TOWARDS AUTOMATIC SPEECH RECOGNITION WITHOUT PRONUNCIATION DICTIONARY, TRANSCRIBED SPEECH AND TEXT RESOURCES IN THE TARGET LANGUAGE USING CROSS-LINGUAL WORD-TO-PHONEME ALIGNMENT

Felix Stahlberg[⋆] *Tim Schlippe*[⋆] *Stephan Vogel*[†] *Tanja Schultz*[⋆]

[⋆] Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany
[†] Qatar Computing Research Institute, Qatar Foundation, Qatar

ABSTRACT

In this paper we tackle the task of bootstrapping an Automatic Speech Recognition system without an a priori given language model, a pronunciation dictionary, or transcribed speech data for the target language Slovene – only untranscribed speech and translations to other resource-rich source languages of what was said are available. Therefore, our approach is highly relevant for under-resourced and non-written languages. First, we borrow acoustic models from a strongly related language (Croatian) and apply a Croatian phoneme recognizer to the Slovene speech. Second, we segment the recognized phoneme strings into word units using cross-lingual word-to-phoneme alignment. Third, we compensate for phoneme recognition and alignment errors in the segmented phoneme sequences and aggregate the resulting phoneme sequence segments in a pronunciation dictionary for Slovene. Orthographic representations are generated using a Croatian phoneme-to-grapheme model. Finally, we use the resulting dictionary and the Croatian acoustic models to recognize Slovene. Our best recognizer achieves a Character Error Rate of 52% on the BMED corpus.

Index Terms— pronunciation dictionary, non-written languages, word-to-phoneme alignment, language discovery, zero-resource

1. INTRODUCTION

Nowadays the majority of state-of-the-art Automatic Speech Recognition (ASR) systems heavily relies on large amounts of data which is necessary to train such systems. Transcribed speech resources, large amounts of text for language modeling, and pronunciation dictionaries are of great importance to create such systems. Authors in [1] estimate that transcription of 1 hour conversational speech data can take up to 20 hours of effort. Therefore, in recent years ASR research has shifted its focus to low- and under-resourced settings [2] to address less prevalent languages as well, e.g. by exploring new ways to collect data [3, 4, 5, 6], using grapheme-based approaches [7], or sharing information across languages [8]. Zero-resource (ZR) ASR goes one step further and even refrains from assuming the availability of a pronunciation dictionary, transcribed audio data, or a language model (LM) in the target language [9] – only untranscribed audio data are available in the target language. Language discovery for ZR ASR can be subdivided into two steps: 1. *Phonetic discovery* aims to find subword units suitable for acoustic modeling. 2. *Lexical discovery* identifies word-like structures and phrases based on phonetic transcriptions of continuous target language speech. Word segmentation describes the task of segmenting the phonetic target language transcriptions into word-like units and thus is a form of lexical discovery. Among a variety of monolingual approaches to word segmentation [9, 10, 11, 12], recent studies have shown [13, 14, 15, 16], that adding written translations in a resource-rich source language can help the word segmentation process.

In our scenario a human translator produces utterances in the target language (Slovene) from prompts in one or many resource-rich source languages (German, English, Croatian) as illustrated in Fig. 1. Since we assume the source languages to be well-studied, we have access to a phoneme recognizer trained on speech data in a related language (Croatian). Since Croatian is phonetically closely related with Slovene, we simply use this phoneme recognizer for Slovene instead of applying phonetic discovery methods that learn from target language speech. In this paper, we combine recent findings in language discovery research, a novel string clustering method, and resources from other resource-rich languages to build an ASR system for the target language Slovene without an a priori given LM, a pronunciation dictionary, or transcribed speech data in the target language – only untranscribed speech in Slovene and translations to other resource-rich source languages of what was said is available. Our approach is in particular intended for non-written languages and dialects since no written form is required for the target language.

To the best of our knowledge, although isolated steps in this process have been studied extensively within restricted experiment settings, the complete pipeline has never been set up before. In particular, research in word segmentation often assumes error-free phonetic transcriptions [12, 9]. In this work, we compensate for alignment and recognition errors.

2. RELATED WORK

Pronunciation dictionaries are used to train speech processing systems by describing the pronunciation of words in manageable units such as phonemes [17]. The production of pronunciation dictionar-

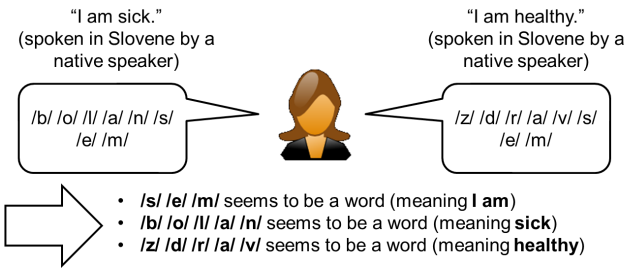


Fig. 1. Scenario.

	Vocabulary Size	Avg. Word Frequency	Avg. Sentence Length	Speakers	Audio
Croatian (hr)	280 words	3.19	4.47 words	8	96 min.
English (en)	163 words	6.90	5.62 words	-	-
German (de)	184 words	6.11	5.62 words	-	-
Slovene (si)	279 words	3.24	4.50 words	5	50 min.

Table 1. Text analysis on BMED.

ies can be time-consuming and expensive if they are manually written by language experts. Therefore several data-driven approaches to automatic dictionary generation, e.g. [18], and to leverage off pronunciations from the World Wide Web have been introduced [3, 4, 19, 5]. However, those approaches do not help for dialects or languages without a written form. Speech processing for non-written languages has been studied in context of speech translation [13, 14, 20], speech syntheses [21, 22] and speech recognition [15, 16]. In this paper, we use a phoneme-to-grapheme model from a related language to generate orthographic representations from extracted pronunciations and thus do not rely on a written form of the target language.

Studies in the fields of psychology and cognitive science investigated infants ability to segment fluent speech into words [24, 25, 26]. Unsupervised word or morphology segmentation in machine learning relies primarily on statistical models [27, 10, 12, 11] or Minimal Description Length analysis [28, 29]. Using translations in a source language for word segmentation and pronunciation extraction is addressed in [13, 14, 15, 16]. This paper investigates how to integrate translations in multiple source languages.

Phonetic language discovery (i.e. Identifying phoneme like subword units for acoustic modeling in an unseen target language) is addressed among others in [30, 31, 32]. Bootstrapping it using ASR systems from other languages and adaptation techniques are presented e.g. in [33]. In this work, we use the phoneme set and acoustic models of Croatian for the Slovene language since they are strongly related.

Document clustering [34] addresses the task of grouping a set of strings into clusters, but usually deals with text documents consisting of a large number of words (e.g. web documents [35]). In contrast to this, we aim to cluster short strings (word pronunciations) based on the Levenshtein distance.

3. BMED CORPUS

We collected our BMED corpus (*Basic Medical Expression Database*) to evaluate our methods on short but realistic sentences. The BMED corpus consists of 200 parallel written sentences in Croatian, English, German and Slovene in the scope of common medical phrases. We recorded 50 minutes Slovene speech from 5 Slovene native speakers. Each Slovene sentence was read by 3-5 Slovene speakers. Tab. 1 shows high average word frequencies (i.e. frequent word recurrences), small vocabularies and low average sentence lengths. Tab. 2 summarizes the word-level IBM-4 perplexities from the BMED languages to Slovene given by GIZA++ [23].

Source Language	IBM-4 Perplexity
Croatian (hr)	3.25
English (en)	9.15
German (de)	7.95

Table 2. IBM Model 4 perplexity according GIZA++ [23] on the BMED corpus (Target language: Slovene).

Since our speakers were geographically widely distributed, we developed the web-based recording tool *CorpusGong*. This tool is designed with major respect to usability and stability so that speakers can record their utterances at home and no supervision is necessary. The user interface (Fig. 2) is available in Croatian, English and Slovene. The open source tool *NanoGong* [37] was integrated to provide voice recording functionality via a Java applet.

Croatian and Slovene are phonetically and linguistically related. Tab. 3 shows the high similarity of both phoneme sets. Both are South Slavic languages and their orthographies are based on *Gaj's Latin alphabet* [38, 39]. This enables us to utilize the Croatian phoneme set, acoustic models, and a phoneme-to-grapheme model for the Slovene language with limited performance degradation.

4. PRONUNCIATION DICTIONARY EXTRACTION

We bootstrap an ASR system for the target language Slovene without a Slovene pronunciation dictionary or transcribed Slovene audio data using resources from Croatian and written translations in Croatian, English, and German. First, we build the pronunciation dictionary for Slovene:

1. **Phoneme Recognition**: Transform the Slovene speech with a Croatian phoneme recognizer to phoneme sequences.
2. **Cross-lingual Word-to-Phoneme Alignment**: Align the phoneme sequences to the written translations in Croatian, English, and German. The alignments induce a segmentation of the phoneme sequences in word-like chunks.
3. **Phoneme Sequence Clustering**: The phoneme sequence segments extracted from the alignments suffer from frequent alignment and phoneme recognition errors. Therefore, group different realizations of the same Slovene word into clusters.

Fig. 2. Recording interface of *CorpusGong* in Slovene.

Related Lang. \ Target Lang.	Croatian (33 phonemes)	English (42 phonemes)	German (39 phonemes)	Slovene (34 phonemes)
Croatian	100%	45.24%	61.54%	**88.24%**
English	75.76%	100%	69.23%	70.58%
German	72.73%	61.90%	100%	82.35%
Slovene	**90.91%**	57.14%	71.79%	100%

Table 3. Phoneme set coverages of the BMED languages according to IPA [36].

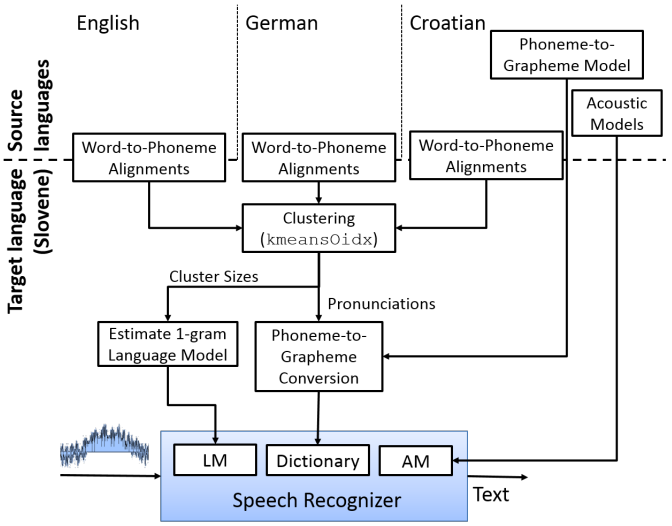


Fig. 3. Overview of the system design.

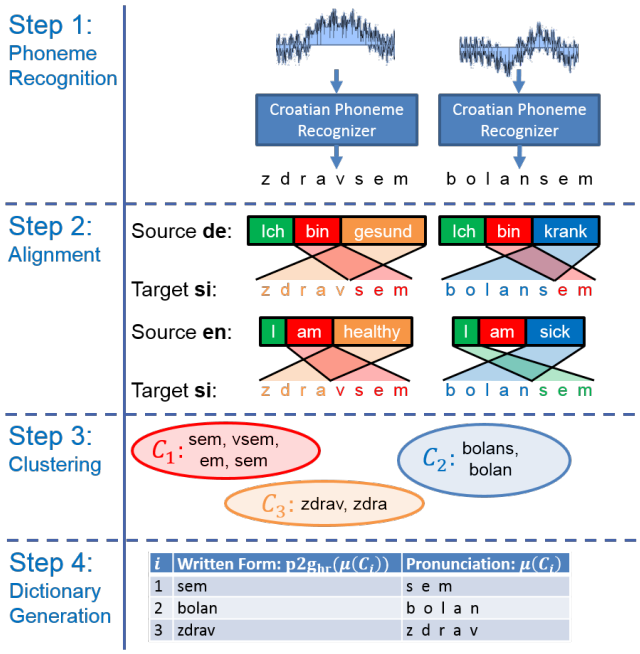


Fig. 4. Steps for the pronunciation dictionary extraction.

4. **Building the Pronunciation Dictionary**: For each cluster, find a representative phoneme sequence and write it as word pronunciation to the dictionary. Orthographic representations are generated with a phoneme-to-grapheme model from Croatian.

The next sections discuss these steps in depth. Fig. 3 shows the complete system design and Fig. 4 illustrates the steps with the help of a small example.

4.1. Phoneme Recognition

The phoneme recognition in Slovene is done with a Croatian context-independent phoneme recognizer (50 phonemes) trained on 20 hours Croatian speech data from the GlobalPhone project [40] using the Janus Recognition Toolkit [41]. The GlobalPhone corpus is a collection of read speech in 20 widespread languages in the world. The recognizer uses a uniformly distributed phoneme level LM (0-gram) because we have no knowledge about n-gram phoneme frequencies in the target language. The preprocessing consists of feature extraction applying a Hamming window of 16ms length with a window shift of 10ms. Each feature vector has 143 dimensions by stacking 11 adjacent frames of 13 Melscale Frequency Ceptral Coefficients (MFCC) frames. A Linear Discriminant Analysis (LDA) transformation is computed to reduce the feature vector size to 42 dimensions. The AM uses a fully-continuous 3-state left-to-right HMM with emission probabilities modeled by Gaussian Mixtures with diagonal covariances (64 Gaussians per state). The recognizer achieves a Phoneme Recognition Error Rate of 33.0% on the Croatian GlobalPhone test set, and 55.2% Phoneme Recognition Error Rate on Slovene speech data from the BMED corpus (Tab. 4). The error rate for the Slovene speech was calculated using an IPA based mapping from the Slovene phoneme set.

	Phoneme Error Rate
Croatian GlobalPhone test set	33.0%
Croatian BMED corpus	43.4%
Slovene BMED corpus	**55.2%**

Table 4. Performance of the Croatian phoneme recognizer.

4.2. Cross-lingual Word-to-Phoneme Alignment

Cross-lingual word-to-phoneme alignments introduced in [20, 13, 14] and tackled by [15] with the alignment model *Model 3P* are the basis for our pronunciation extraction algorithm. The word segmentation problem describes the task of segmenting phoneme sequences into word units. [15] and [16] show that unsupervised learning of word segmentation is more accurate when information of another language is used. *Model 3P* (implemented in the PISA Alignment Tool[1]) for cross-lingual word-to-phoneme alignment extends the generative process of IBM Model 3 by a word length step and additional dependencies for the lexical translation probabilities. Alignments are used for the segmentation task as illustrated in Fig. 5. For

[1]available at `http://pisa.googlecode.com/`

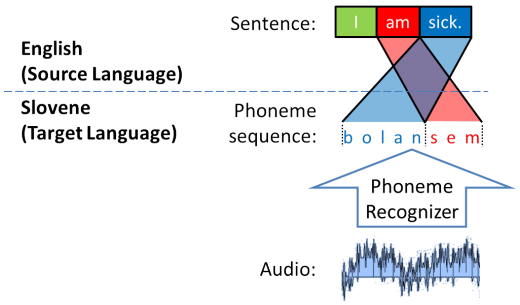


Fig. 5. Word segmentation through word-to-phoneme alignment.

a more detailed description of *Model 3P* used in this paper we refer to [15].

4.3. Phoneme Sequence Clustering

Let $PhonemeSet_{hr}$ be the Croatian phoneme set and $\Sigma \subset {PhonemeSet_{hr}}^+$ be the set of phoneme sequence segments which we extract from the word-to-phoneme alignments (see Sec. 4.2). The function $c : \Sigma \to \mathbb{N}_+$ indicates how often a phoneme sequence segment was found in the word-to-phoneme alignments in the previous step. Elements in Σ correspond to Slovene words, but are often corrupted by alignment and phoneme recognition errors. In this step, we therefore group different realizations of the same Slovene word into clusters

$$C = \{C_1, C_2, \ldots, C_n\} \subset \mathcal{P}(\Sigma) \text{ with } \Sigma = \biguplus_{i \in [1,n]} C_i. \quad (1)$$

In order to build C automatically, we first run k-means clustering [42] based on the Levenshtein distance for 8 iterations. The means are initialized with the k most frequent elements in Σ (indicated by c). To find a mean $\mu(C_i) \in {PhonemeSet_{hr}}^+$ for a cluster C_i, we use the `nbest-lattice` tool [43]. k is an initial guess for the target language vocabulary size, that can be derived from the vocabulary size of Croatian. Therefore we set $k = 280$. However, k-means fails to separate phonetically similar Slovene words reliably: For example, different inflections (like *bo* ("it will") and *bom* ("I will")) or completely different words (like *da* ("yes") and *dan* ("day")) are often placed in the same cluster. Instead, we want elements in C_i to be realizations of a single Slovene word $\mu(C_i)$. Therefore we introduce the *outlier index* $\texttt{oidx} : C \to \mathbb{Q}_+$:

$$\widetilde{C_i} := \{p \in C_i | p \neq \mu(C_i)\} \quad (2)$$

$$\texttt{oidx} : C_i \mapsto \begin{cases} 1 & \text{if } \widetilde{C_i} = \emptyset \\ \frac{\max_{p \in \widetilde{C_i}} c(p)}{\text{Median}(\{c(p) | p \in \widetilde{C_i}\})} & \text{otherwise} \end{cases} \quad (3)$$

$\widetilde{C_i}$ denotes the set of all elements in C_i that differ from the mean $\mu(C_i)$. If the elements in $\widetilde{C_i}$ are corrupted realizations of $\mu(C_i)$, we assume that they are approximately uniformly distributed ($\text{Var}(c(\widetilde{C_i}))$ is small). This is only an approximation, because errors made by the phoneme recognizer or the alignment model usually depend on the context and the actually spoken phoneme. However, an element $o \in \widetilde{C_i}$ that occurs significantly more often than other elements in $\widetilde{C_i}$ is likely to be a different Slovene word rather than a corrupted version of $\mu(C_i)$: Correct pronunciations are assumed to occur more often than incorrect ones and thus result in a high `oidx`. Fig. 6 illustrates the `oidx` for seven clusters – i.e. seven possible distributions in $\widetilde{C_i}$. Peaks in the distributions indicate Slovene words that are incorrectly grouped to cluster C_i. In each iteration, the maximal element in each $\widetilde{C_i}$ with `oidx` higher than a threshold ϵ_{oidx} is moved in a separate cluster and k is incremented (i.e. k is increased to 12 in Fig. 6). We refrained from using other outlier detection methods: Some (e.g. Grubbs' test or Chauvenet's criterion [44]) take the standard derivation into account and thus score examples 3) to 5) differently. Some (e.g. Grubbs' test) only work reliably on larger sample sizes. Dixon's Q test [45] would penalize example 3) containing two outliers because of the small gap between both. Tests based on the (edit) distance between the cluster mean and the putative outlier are misleading since a small distance does not necessary indicate that both elements belong to the same cluster (e.g. *mine* and *fine* in English). On the contrary, the proposed `oidx` is a simple criterion that has proved to be effective in our case. After the initial 8 k-means iterations, we therefore run a modified version of k-means that searches for clusters with high *outlier indices* in each iteration, and would propagate o as a new mean by putting it in a separate cluster. Thus, the final number of clusters may differ from k since it is incremented whenever an element is moved to a separate cluster because of the `oidx` criterion. The complete algorithm description is listed in Alg. 1.

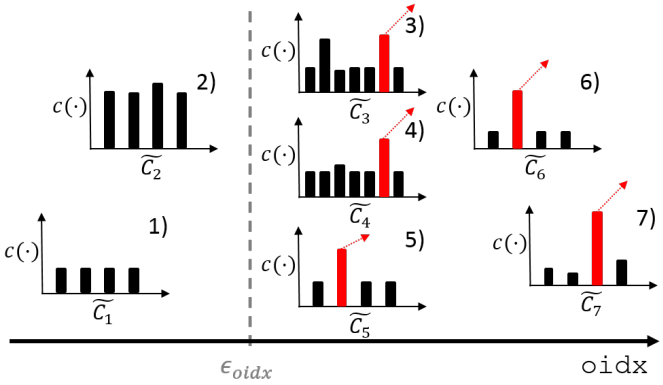


Fig. 6. The outlier indices for different distributions in $\widetilde{C_i}$.

Algorithm 1 `kmeansOidx`$(\Sigma, k \in \mathbb{N}_+, \epsilon_{oidx} \in \mathbb{Q}_+)$

Require: $\Sigma \neq \emptyset$
Require: $M \subset {PhonemeSet_{hr}}^+ \times \mathcal{P}(\Sigma)$
$M \leftarrow \text{initializeMeans}(\Sigma, k)$
for $i \leftarrow 1$ **to** 8 **do**
  assignmentStep(M)
  updateStep(M)
end for

for $i \leftarrow 1$ **to** 8 **do**
  assignmentStep(M)
  updateStep(M)
  for all $\{(\mu, C) \in M | \texttt{oidx}(C) \geq \epsilon_{oidx}\}$ **do**
    $o \leftarrow \arg\max_{\mu \neq p \in C} c(p)$
    $M \leftarrow M \cup \{(o, \{o\})\}$
  end for
end for

4.4. Building the Pronunciation Dictionary

The previous step results in a set of clusters $C = \{C_1, C_2, \ldots, C_n\}$, where each cluster C_i stands for a single Slovene word. We use its mean $\mu(C_i)$ as pronunciation in the pronunciation dictionary for Slovene. Consequently, the number of clusters $|C|$ is reflected by the size of the extracted dictionary in Tab. 5 and 7. To find an orthographic representation for $\mu(C_i)$, we transform the phoneme sequence to a written form with a phoneme-to-grapheme model from a related language. Croatian is closely related to Slovene and is even written with the same script (Sec. 3). Both languages have a good phoneme-to-grapheme relation. We trained a phoneme-to-grapheme model $\mathrm{p2g_{hr}}$ on the Croatian GlobalPhone pronunciation dictionary using Sequitur G2P [18]. This model is used for generating the written form from a pronunciation $\mu(C_i)$. Applying this model to the correct pronunciations in the Slovene reference dictionary and comparing the generated written forms with the correct Slovene written words results in 5.4% character error rate.

5. LANGUAGE MODEL EXTRACTION

Language modeling for the target language is especially hard because we do not assume the availability of text data in the target language. In initial experiments, we apply a uniformly distributed LM (0-gram). However, Sec. 6.3 shows that our best results are achieved with a unigram LM. The unigram word probabilities are estimated using the sum of occurrences of elements in a cluster:

$$\widehat{P}(\mathrm{p2g_{hr}}(\mu(C_i)) = \frac{\sum_{p \in C_i} c(p)}{\sum_{p \in \Sigma} c(p)}, i \in [1, n]. \quad (4)$$

6. EXPERIMENTS

6.1. Evaluation Measures

To evaluate the extracted pronunciation dictionaries, we apply the evaluation measures for the pronunciation extraction from phoneme sequences introduced in [16]: The mapping m maps each entry in the extracted dictionary to the most similar pronunciation in the reference dictionary containing the correct pronunciations as shown in Fig. 7. The **Phoneme Error Rate (PER)** is the average edit distance between pairs mapped by m. The **Out-Of-Vocabulary rate (OOV)** is calculated using the set of all reference dictionary entries mapped by m. The **Hypo/Ref ratio** indicates how many hypothesis entries in the extracted dictionary are mapped to a single reference dictionary entry on average. The higher the Hypo/Ref ratio, the more pronunciations are extracted unnecessarily.

We calculate the **Character Error Rate (CER)** to evaluate the final word recognizer for the target language Slovene because it is more robust against minor spelling or segmentation errors than the Word Error Rate (WER). Blanks are treated as separate characters. In all experiments, the word recognizer uses context-dependent acoustic models trained on 20 hours Croatian speech from the GlobalPhone project [40] to recognize Slovene. The preprocessing is described in Sec. 4.1. The AM uses a fully-continuous 3-state left-to-right HMM with emission probabilities modeled by Gaussian Mixtures with diagonal covariances. For our context-dependent AMs with different context sizes, we stopped the decision tree splitting process at 2,000 triphones. After context clustering, a merge-and-split training was applied, which selects the number of Gaussians according to the amount of data (19 on average, 38k in total). We did not apply adaptation techniques to improve the acoustic models since Slovene transcriptions are not given. This recognizer achieves a CER of 13.6% on the Slovene portion of the BMED corpus with the correct Slovene pronunciation dictionary and an 1-gram LM trained on the 200 BMED sentences (Tab. 6).

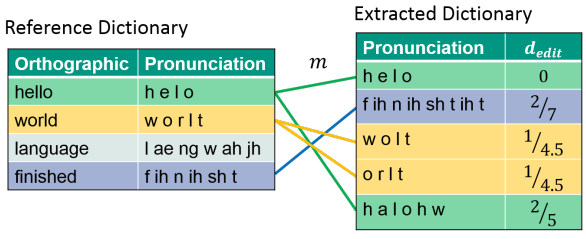


Fig. 7. Mapping m between extracted pronunciations and written words for evaluation (Target language in this example: English)

6.2. Experiments with Error-Free Phonetic Transcriptions

First we replace *Step 1* in our approach in Sec. 4 and simulate a perfect phoneme recognizer with 0% Phoneme Recognition Error Rate by replacing the words in the Slovene text with their canonical pronunciation and removing word boundary markers. Thereby we initially refrain from dealing with pronunciation variants and phoneme recognition errors.

Tab. 5 summarizes the results when the pronunciation dictionaries are extracted from perfect phonetic transcriptions. As reference we also report the performance of the method presented in [16] that uses a different clustering algorithm and combines elements only if they are aligned to the same source language word. In contrast to this, we cluster solely on the basis of phonetic similarities. Setting $\epsilon_{oidx} = \infty$ constricts `kmeansOidx` to the standard k-means algorithm. k is fixed to 280 in all experiments, which is an initial guess for the target language vocabulary size derived from the Croatian vocabulary size. This initialization is not required to be exact since in general the final size of the extracted dictionary differs from k (as shown in the third column of Tab. 5 and 7): On the one hand, as described in Sec. 4.3, the dictionary size is increased for each new cluster identified by the `oidx` criterion. On the other hand, `nbest-lattice` occasionally calculates the same mean for two separate clusters causing them to merge in k-means' assignment step.

We observe that on the designated task, `kmeansOidx` generally results in lower CERs than the method from [16], although the Hypo/Ref ratio is higher. Limiting the maximum `oidx` with ϵ_{oidx} effectively reduces both the OOV rate and the CER and tends to produce larger dictionaries. The best recognizer has a CER of 44.2%. When blanks are ignored in the evaluation so that segmentation errors do not affect the error rate, this system achieves a CER of 35.3%.

6.3. Experiments with Recognized Phoneme Sequences

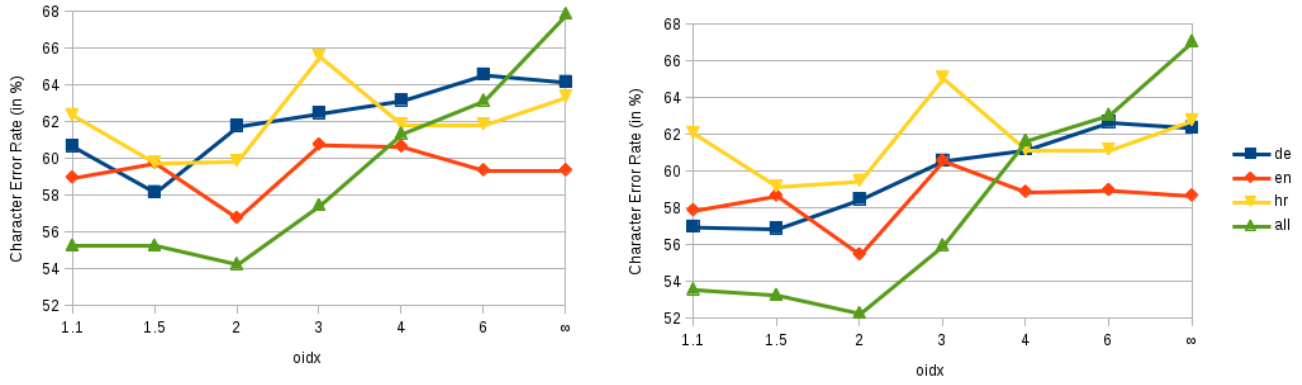
In our scenario we use a Croatian phoneme recognizer to recognize the Slovene target language speech in order to build the pronunci-

Language Model	WER	CER
0-gram	36.2%	15.7%
1-gram	32.0%	13.6%

Table 6. Recognition performance on Slovene with the correct reference dictionary (gold standard).

Method	Src. Lang.	Dict. Size.	PER	OOV		Hypo/ Ref	CER	
				unique	running		0-gram LM	1-gram LM
Method from [16]	de	80	57.8	76.4	59.6	1.13	**66.0**	–
Method from [16]	en	74	49.8	76.4	60.3	1.04	**62.4**	–
Method from [16]	hr	164	56.8	53.9	33.1	1.22	**51.3**	–
`kmeansOidx`, $\epsilon_{oidx} = \infty$	de	280	55.6	39.3	26.2	1.62	**52.1**	**50.7**
`kmeansOidx`, $\epsilon_{oidx} = \infty$	en	279	51.4	33.9	22.7	1.48	**51.4**	**47.9**
`kmeansOidx`, $\epsilon_{oidx} = \infty$	hr	278	54.3	36.1	20.9	1.53	**49.9**	**48.4**
`kmeansOidx`, $\epsilon_{oidx} = \infty$	All	275	47.3	42.1	25.4	1.67	**47.3**	**46.1**
`kmeansOidx`, $\epsilon_{oidx} = 3$	All	282	46.3	42.5	25.2	1.72	**46.6**	**46.2**
`kmeansOidx`, $\epsilon_{oidx} = 2$	All	318	47.5	36.1	20.5	1.75	**45.9**	**45.1**
`kmeansOidx`, $\epsilon_{oidx} = 1.5$	All	324	49.0	34.6	19.7	1.74	**45.8**	**44.4**
`kmeansOidx`, $\epsilon_{oidx} = 1.1$	All	322	48.3	35.0	20.0	1.74	**45.2**	**44.2**

Table 5. Results on error-free phonetic transcriptions (oracle experiments).



(a) 0-gram language model.

(b) 1-gram language model.

Fig. 8. Performance of the final word recognizer using different dictionaries.

ation dictionary (Sec. 4.1). Therefore, we operate on phoneme sequences with a Phoneme Recognition Error Rate of 55.2% rather than on error-free phonetic transcriptions. Tab. 7 shows our evaluation measures for dictionaries extracted from recognized phoneme sequences. Reducing ϵ_{oidx} leads to larger dictionaries and higher Hypo/Ref ratios, but again significantly reduces the OOV rates. To put it another way, a low ϵ_{oidx} produces noisier dictionaries that on the other hand cover more Slovene words.

Ultimately the dictionaries need to prove their usefulness when integrated in a word recognizer for the target language. Fig. 8 plots the CER of the final recognizer over ϵ_{oidx} for the source languages German, English, Croatian, and the combination of all of them. All curves pass through their minimum at $\epsilon_{oidx} = 2$ or $\epsilon_{oidx} = 1.5$. The best CER (55.5%) with only one source language is achieved with English, $\epsilon_{oidx} = 2$ and a unigram LM. Using English as source language consistently performs better than Croatian. We believe that this is due to the poor morphology of English: A small source language vocabulary size reduces the number of parameters in the alignment model *Model 3P* so that they can be estimated more reliably. We report a CER of 52.3% for our best system with a unigram LM and a dictionary extracted using all source languages ($\epsilon_{oidx} = 2$). When blanks are ignored in the evaluation, this system achieves a CER of 44.9%. Recognition examples are listed in Tab. 8 to get an impression of the errors that still remain.

6.4. Human Evaluation

In addition to the automatic evaluation measures we perform a human evaluation for our best systems based on error-free and recognized phonetic transcriptions (44.2% and 52.3% CER): For 100 randomly selected recognition hypotheses from each system we ask a Slovene native speaker to select one sentence from a alphabetically sorted list of all 200 BMED sentences that she thinks is the correct output. If the answer (including reading, understanding, and finding the sentence in the list) takes longer than *Maximum Answering Time* (MAT), we consider it as wrong. A correct answer with a MAT of 15 seconds usually implies instant understanding. Tab. 9 shows the percentage of correct answers for both systems with different MATs. Surprisingly, the system using recognized phonetic transcriptions outperforms the system using error-free transcriptions when MAT$\geq$ 30, indicating that a low OOV is more important for understanding than a low PER or even CER. The speaker is able to identify the correct output for 87.9% of the hypotheses from the system based on recognized transcriptions within 1 minute.

7. CONCLUSION AND FUTURE WORK

We have tackled the task of bootstrapping an Automatic Speech Recognition (ASR) system without an a priori given language model (LM), a pronunciation dictionary, or transcribed speech data for the target language Slovene – only untranscribed speech and translations to other resource-rich source languages of what was said were

available. In our scenario a human translator produced utterances in the target language (Slovene) from prompts in resource-rich source languages (Croatian, English, German). First, we cross-lingually aligned the target language speech to the written translations and obtained phoneme sequence segments corresponding to Slovene words, but corrupted by recognition and alignment errors. Second, we introduced a new clustering method `kmeansOidx` and grouped the segments into clusters. The means of these clusters were written as word pronunciations to the pronunciation dictionary. A Croatian phoneme-to-grapheme model provided orthographic representations. The sizes of the clusters were used to estimate unigram LM probabilities. Both the dictionary and the LM together with Croatian acoustic models were then used to recognize Slovene.

We collected a small corpus (BMED) in four languages consisting of 200 parallel sentences and 50 minutes Slovene speech to evaluate our methods. Our best system achieved a CER of 52% on this corpus, using a unigram LM and all three translations to extract Slovene word pronunciations. For 87.9% of the recognizer hypotheses a Slovene native speaker was able to spot the correct output in a list of 200 sentences within 1 minute.

In the future, we plan to focus on acoustic modeling and apply phonetic discovery methods as in [30, 31, 32] on the target language speech rather than a phoneme recognizer of a related language. Acoustic models could be further improved by iteratively recognizing the speech to provide target language transcriptions, and then using the transcriptions to adapt the models. When it comes to evaluating speech recognizers with partially misspelled words, a automatic measure that is more meaningful than the CER, but less erratic than the Word Error Rate is to be found. Although our results on a limited domain and a small vocabulary are encouraging, the evidence for the applicability of our method on a larger vocabulary and a truly under-resourced or non-written language is still pending.

Src. Lang.	ϵ_{oidx}	Dict. Size.	PER	OOV unique	OOV running	Hypo/ Ref
de	∞	257	58.0	43.6	29.2	1.60
	6	258	57.2	44.3	29.8	1.62
	4	257	57.4	42.9	29.1	1.58
	3	262	57.3	38.9	27.1	1.51
	2	334	60.5	32.9	23.0	1.75
	1.5	333	58.9	31.1	21.5	1.70
	1.1	328	58.8	32.5	22.3	1.71
en	∞	251	55.3	42.5	30.1	1.53
	6	252	55.4	42.1	29.9	1.53
	4	255	56.3	43.6	29.2	1.58
	3	266	57.5	41.4	29.8	1.59
	2	306	56.5	37.9	24.7	1.73
	1.5	308	56.6	39.3	26.3	1.78
	1.1	306	56.5	35.4	22.8	1.66
hr	∞	253	59.6	39.6	25.2	1.47
	6	258	58.8	37.5	26.0	1.45
	4	258	58.8	37.5	26.0	1.45
	3	268	59.8	37.9	25.1	1.51
	2	316	58.3	32.1	21.3	1.64
	1.5	320	59.3	32.5	21.6	1.67
	1.1	319	59.3	33.2	20.9	1.68
all	∞	212	54.7	57.1	40.6	1.72
	6	236	55.5	53.6	36.6	1.77
	4	309	54.8	43.6	29.6	1.92
	3	448	56.1	26.4	16.3	2.14
	2	1145	64.8	5.0	3.0	4.26
	1.5	1142	64.4	4.6	2.2	4.23
	1.1	1171	64.2	5.0	2.3	4.35

Table 7. Quality of pronunciation dictionaries extracted from recognized phoneme sequences with the `kmeansOidx` algorithm.

CER	Alignment
18%	REF: n i m a m _ č a * s a HYP: n i m a m _ š a _ s a
25%	REF: p o š k o d o v * a n _ s * e m HYP: p o š k o d o v _ a n e s _ a m
31%	REF: * * p o m e m b n o _ j e HYP: p _ p o m i m b n a _ j e
40%	REF: p o š k o d o v * a n _ s e m HYP: * o ž g o d o v _ a n _ z a m
44%	REF: l e ž i * t e _ u d * * o b n o HYP: * i ž i _ t e _ u d u _ p _ n o
50%	REF: z d r a * v _ s e m HYP: z g r a l b _ z a m

Table 8. Example hypotheses and their CER (on recognized phoneme sequences, unigram LM, $\epsilon_{oidx} = 2$). Blanks are represented by the underscore character (_).

System / MAT	Error-Free Transcript., $\epsilon_{oidx} = 1.1$ (44.2% CER)	Recognized Transcript., $\epsilon_{oidx} = 2$ (52.3% CER)
15 Seconds	37.0%	27.3%
30 Seconds	65.0%	73.7%
45 Seconds	73.0%	83.8%
1 Minute	77.0%	87.9%

Table 9. Percentage of correct answers in the human evaluation (all source languages, unigram LM).

8. REFERENCES

[1] T. Schultz and K. Kirchhoff, Eds., *Multilingual Speech Processing*, Academic Press, Amsterdam, 2006.

[2] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic Speech Recognition for Under-Resourced Languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[3] A. Ghoshal, M. Jansche, S. Khudanpurv, M. Riley, and M. Ulinski, "Web-derived Pronunciations," in *ICASSP*, 2009.

[4] D. Can, E. Cooper, A. Ghoshal, M. Jansche, S. Khudanpur, B. Ramabhadran, M. Riley, M. Saraclar, A. Sethy, M. Ulinski, and C. White, "Web Derived Pronunciations for Spoken Term Detection," in *ACM SIGIR*, 2009.

[5] T. Schlippe, S. Ochs, and T. Schultz, "Web-based Tools and Methods for Rapid Pronunciation Dictionary Creation," *Speech Communication*, vol. 56, pp. 101–118, 2014.

[6] H. Gelas, S. T. Abate, L. Besacier, and F. Pellegrino, "Quality Assessment of Crowdsourcing Transcriptions for African Languages.," in *Interspeech*, 2011.

[7] S. Kanthak and H. Ney, "Context-dependent Acoustic Modeling Using Graphemes for Large Vocabulary Speech Recognition," in *ICASSP*, 2002.

[8] T. Schultz and A. Waibel, "Language-independent and Language-adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–51, 2001.

[9] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, et al., "A Summary of the 2012 JHU CLSP Workshop on Zero Resource Speech Technologies and Models of Early Language Acquisition," in *ICASSP*, 2013.

[10] M. Johnson, "Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure," in *ACL-HLT*, 2008.

[11] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Unsupervised Word Segmentation From Noisy Input," in *ASRU*, 2013.

[12] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling," in *ACL and AFNLP*, 2009.

[13] S. Stüker and A. Waibel, "Towards Human Translations Guided Language Discovery for ASR Systems," in *SLTU*, 2008.

[14] S. Stüker, L. Besacier, and A. Waibel, "Human Translations Guided Language Discovery for ASR Systems," in *Interspeech*, 2009.

[15] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word Segmentation through Cross-Lingual Word-to-Phoneme Alignment," in *SLT*, 2012.

[16] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Pronunciation Extraction from Phoneme Sequences through Cross-Lingual Word-to-Phoneme Alignment," in *SLSP*, 2013.

[17] O. Martirosian and M. Davel, "Error Analysis of a Public Domain Pronunciation Dictionary," in *PRASA*, 2007.

[18] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[19] T. Schlippe, S. Ochs, and T. Schultz, "Grapheme-to-Phoneme Model Generation for Indo-European Languages," in *ICASSP*, 2012.

[20] L. Besacier, B. Zhou, and Y. Gao, "Towards Speech Translation of Non Written Languages," in *SLT*, 2006.

[21] S. Sitaram, G. K. Anumanchipalli, J. Chiu, A. Parlikar, and A. W. Black, "Text to Speech in New Languages without a Standardized Orthography," in *Speech Synthesis Workshop*, 2013.

[22] S. Sitaram, S. Palkar, Y. Chen, A. Parlikar, and A. W. Black, "Bootstrapping Text-to-Speech for Speech Processing in Languages Without an Orthography," in *ICASSP*, 2013.

[23] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[24] P. Cairns, R. Shillcock, N. Chater, and J. Levy, "Bootstrapping Word Boundaries: A Bottom-up Corpus-Based Approach to Speech Segmentation," *Cognitive Psychology*, vol. 33, no. 2, pp. 111–153, 1997.

[25] E. K. Johnson and P. W. Jusczyk, "Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics," *Journal of Memory and Language*, vol. 44, no. 4, pp. 548–567, 2001.

[26] M. H. Christiansen, J. Allen, and M. S. Seidenberg, "Learning to Segment Speech Using Multiple Cues: A Connectionist Model," *Language and Cognitive Processes*, vol. 13, no. 2-3, pp. 221–268, 1998.

[27] J. A. Goldsmith, "Segmentation and Morphology," *The Handbook of Computational Linguistics and Natural Language Processing*, vol. 57, pp. 364, 2010.

[28] C. Kit, "Unsupervised Lexical Learning as Inductive Inference," Tech. Rep., CityU of HK Press, 2000.

[29] J. Goldsmith, "An Algorithm for the Unsupervised Learning of Morphology," *Natural Language Engineering*, vol. 12, no. 4, pp. 353–371, 2006.

[30] C. Lee and J. Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery," in *ACL-HLT*, 2012.

[31] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised Learning of Acoustic Sub-Word Units," in *ACL-HLT*, 2008.

[32] S. Chaudhuri, M. Harvilla, and B. Raj, "Unsupervised Learning of Acoustic Unit Descriptors for Audio Content Representation and Classification," in *Interspeech*, 2011.

[33] N. T. Vu, F. Kraus, and T. Schultz, "Rapid Building of an ASR System for Under-Resourced Languages Based on Multilingual Unsupervised Training," in *Interspeech*, 2011.

[34] N. O. Andrews and E. A. Fox, "Recent Developments in Document Clustering," *Computer Science, Virginia Tech, Blacksburg, VA, Technical Report TR-07-35*, 2007.

[35] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, "A Survey of Web Clustering Engines," *CSUR*, vol. 41, no. 3, pp. 17, 2009.

[36] IPA, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, 1999.

[37] G. Lam and D. Rossiter, "Gong - a Voice for the Web World," 2007, http://nanogong.ust.hk/.

[38] A. Bajec, Ed., *Slovar Slovenskega Knjižnega Jezika*, DZS, 1995.

[39] L. Badurina, I. Marković, and K. Mićanović, Eds., *Hrvatski Pravopis*, Matica Hrvatska, 2007.

[40] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A Multilingual Text & Speech Database in 20 Languages," in *ICASSP*, 2013.

[41] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe-Verbmobil Speech Recognition Engine," in *ICASSP*, 1997.

[42] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[43] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit Word Error Minimization in N-best List Rescoring," in *Eurospeech*, 1997.

[44] V. Barnett and T. Lewis, *Outliers in Statistical Data*, vol. 3, Wiley New York, 1994.

[45] R. B. Dean and W. J. Dixon, "Simplified Statistics for Small Numbers of Observations," *Analytical Chemistry*, vol. 23, no. 4, pp. 636–638, 1951.