ASR pipeline for low-resourced languages: A case study on Pomak

Chara Tsoukala Kosmas Kritsis Ioannis Douros Athanasios Katsamanis
Nikolaos Kokkas Vasileios Arampatzakis Vasileios Sevetlidis
Stella Markantonatou George Pavlidis
Institute for Language and Speech Processing, Athena R.C.
{chara.tsoukala, kosmas.kritsis, ioannis.douros, nkatsam, nikolaos.kokkas,
vasilis.arampatzakis, vasiseve, marks, gpavlid}@athenarc.gr

Abstract

Automatic Speech Recognition (ASR) models can aid field linguists by facilitating the creation of text corpora from oral material. Training ASR systems for low-resource languages can be a challenging task not only due to lack of resources but also due to the work required for the preparation of a training dataset. We present a pipeline for data processing and ASR model training for low-resourced languages, based on the language family. As a case study, we collected recordings of Pomak, an endangered South East Slavic language variety spoken in Greece. Using the proposed pipeline, we trained the first Pomak ASR model.

1 Introduction

Speech technologies have gained popularity in the past decade and several people use voice commands to communicate with their devices or to dictate messages. Furthermore, such technologies can be of use in field and corpus linguistics. Manually transcribing one minute of recorded speech takes on average 40 minutes; Automatic Speech Recognition (ASR) models can facilitate the transcription of spoken corpora by providing the first iteration of the transcription (Foley et al., 2018). If high-quality recordings are available, Text-to-Speech (TTS) models can augment speech corpora by generating audio files from text.

However, training robust models requires several hundred hours of recorded speech, while most languages do not have enough such resources. Therefore, in low-resource settings, one typically bootstraps the process using a model that has been pre-trained in a related language with sufficient resources (e.g., wav2vec2 (Baevski et al., 2020), XLS-R (Conneau et al., 2021), and Whisper (Radford et al., 2022)). The pre-trained model is then fine-tuned on the target language data to obtain the final model (e.g., (Khare et al., 2021; Baevski et al., 2020; Hjortnaes et al., 2020)). To aid linguists, Foley et al. (2018) proposed a pipeline ("Elpis") to help train a Kaldi-based (Povey et al., 2011) ASR model with minimal scripting. The pipeline assumes that the transcription has been done via ELAN [1] which includes timestamps.

However, in case the available transcriptions lack time annotations, creating a dataset for a low-resource language can be a demanding task; one of the reasons is that, typically, ASR systems require short audio segments for the training process. Therefore, to create a dataset, any available recordings must be segmented into smaller parts while retaining the corresponding transcription. Splitting an audio file on its own is a relatively straightforward task in specific conditions. One can use, for instance, a Voice Activity Detection (VAD) algorithm (e.g., using PyAnnote (Bredin and Laurent, 2021) or Praat[2] (Boersma and Van Heuven, 2001)) that segments based on whether speech is present in the signal. However, in the case of missing audio-transcription time alignments, VAD alone cannot split the transcription.

We propose a pipeline (Section 2) for low-resourced languages that i) normalizes the available audio and transcription files, ii) extracts speech-text word-level alignments, iii) segments the audio files into smaller parts to create a dataset, and iv) fine-tunes an ASR model based on the language family. As a use case, we have focused on Pomak, an endangered South East Slavic language variety spoken in Greece (Karahóğa et al., 2022).

Specifically, we have recorded over 14 hours of Pomak read speech (Section 3.1) and used the proposed pipeline to train the first Pomak ASR model (Section 3.2). Even though 14 hours of speech is considered a low-resourced setting in the field of Automatic Speech Recognition, for many endangered languages the available recordings are even fewer. For this reason, we further trained an

[1] https://archive.mpi.nl/tla/elan
[2] https://www.fon.hum.uva.nl/praat/

Proceedings of the Second Workshop on NLP Applications to Field Linguistics, pages 40–45

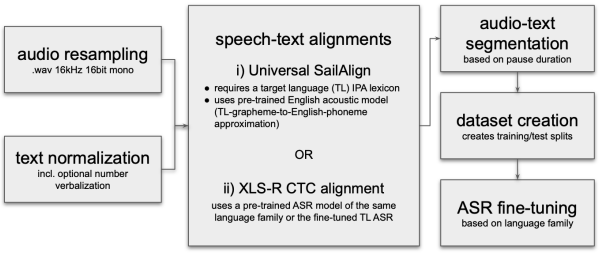

Figure 1: Proposed ASR pipeline

ASR model using only 1 hour of speech to show the applicability of the proposed approach even in the case of a simulated endangered language scenario.

2 ASR pipeline

Typically, popular ASR models use sample rates of 8kHz (8000 samples/sec) or 16kHz. The first step of the pipeline (Figure 1) is to convert all recordings to the latter sample rate (16kHz 16-bit mono channel wav files) because it provides more accurate high-frequency information and it matches the sample rate of the pre-trained models we use (see Section 2.3). Additionally, we normalize the text and convert dates and numbers to their literal equivalent.

To be able to verbalize Pomak dates and numbers, we have extended the num2words package[3]. This part is language-specific and will need to be customized for a new under-resourced language. If that is not possible, the conversion step needs to be done manually and the user will need to convert the numbers into their lexical equivalents.

The most challenging part of the data preprocessing task is the segmentation of audio files while retaining the correct transcription of words. To do so, we first need to obtain speech-text alignments in order to get the exact onset and offset times of each word.

2.1 Speech-text alignments

Speech-text alignments, also known as forced alignments, require an acoustic model (AM) of the language to successfully match audio with a transcript. However, training the AM requires lots of data, which a low-resourced language does not have. To bypass this issue, we have extended SailAlign (Katsamanis et al., 2011) to be able to align new languages using an English pre-trained model.

[3]https://github.com/savoirfairelinux/num2words

Pomak	IPA	English phone
hálove	h a l o v e	hh aa l ow v eh
hadaíčko	h a d a i ʧ k o	hh aa d aa iy ch k ow
haklýje	h a k l i j e	hh aa k l iy y eh

Table 1: Examples of Pomak words, their phonetic representation, and their transformation to an English phone representation that allows SailAlign to use the pre-trained English acoustic model

2.1.1 Universal SailAlign

SailAlign is a toolkit for robust speech-text alignment of long audio files, that implements an adaptive, iterative speech recognition and text alignment scheme. It currently supports English, Spanish, and Greek.

To obtain the alignments in a new language (Pomak in this case), we provided the toolkit with a Pomak IPA dictionary and a Pomak grapheme to English phoneme approximation (see Table 1). This allowed us to utilize the pre-trained English acoustic model, without training a Pomak ASR.

Since Pomak is a Slavic language there is no perfect match between Pomak and English phonemes. However, even this approximation results in good alignments that can be used to segment the original recordings (see Section 2.1.3). The big advantage of this method is that no AM training is needed. The only input needed is the audio-transcription pairs and an IPA (pronunciation) lexicon.

Typically, the IPA lexicon is difficult to obtain because it requires that a phonetician provides the phonetic representation of several words. In case there is no IPA dictionary available in the target language, we have created a helper script that generates an approximation based on a language that has a similar phonology. More specifically, the script is based on Phonemizer (Bernard and Titeux, 2021) that employs eSpeak NG[4] TTS which supports over 100 languages. To test this method, we generated an IPA dictionary in Pomak based on the phonology of another Slavic language.

SailAlign does not handle out-of-vocabulary (OOV) words; the IPA dictionary should contain all words in the transcription files. To facilitate this process, if the user has an incomplete existing IPA dictionary (i.e., if the IPA dictionary lacks some of the words in the transcription files), the Universal SailAlign script can use Phonetisaurus (Novak et al., 2016) to generate the missing

[4]https://github.com/espeak-ng/espeak-ng

items. Phonetisaurus is an open-source grapheme-to-phoneme tool based on Weighted Finite States Transducers (WFSTs). Universal SailAlign is available at https://gitlab.com/ilsp-spmd-all/filotis/universal_sail_align.

2.1.2 Wav2vec2 XLS-R alignments

An alternative method of obtaining alignments is using the CTC-segmentation algorithm proposed by Kürzinger et al. (2020). This method uses a Connectionist Temporal Classification (CTC)-based end-to-end network; in this case, we are using a wav2vec2 (Baevski et al., 2020) ASR model. The model can be a pre-trained model of the same language family (e.g., Slavic), but, ideally, it should be the fine-tuned model of the target language (the process is described in Section 2.3). The advantage of this alignment method is that it is readily available once an initial ASR model is obtained. However, the process heavily depends on the model used; especially when using a generic pre-trained model, alignment success is not guaranteed, making it a less reliable alignment method than Universal SailAlign for low-resourced languages.

2.1.3 Manual evaluation of alignments

To evaluate the performance of the alignments, we manually corrected a few Pomak alignment files and compared the performance of Universal SailAlign and wav2vec2 XLS-R alignments. Specifically, we sampled four audio files of a total of 20 minutes. Using the corresponding Universal SailAlign alignment files as a baseline, we manually corrected the generated alignments using Audacity[5]. As displayed in Figure 2, the percentage of correctly aligned words is at peak for tolerance durations larger than 0.2 seconds, i.e., when the automatically aligned boundaries are considered correct even if they differ up to 200 milliseconds from the manually corrected ones. For smaller time differences (i.e., a tolerance alignment of 0.1 seconds and below), Universal SailAlign clearly outperforms the XLS-R alignments. [6]

2.2 Audio segments

As mentioned above, ASR systems require short audio segments as training input. Typically, audio segments of up to 30 seconds are used to train or fine-tune a model.

[5]Audacity is an open-source audio and label editor. www.audacityteam.org/

[6]The results are available at https://osf.io/dkbnu

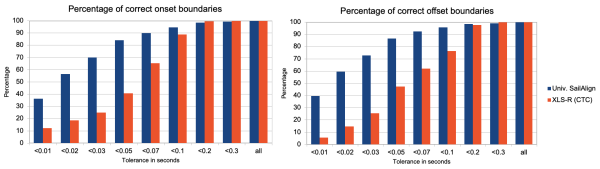

Figure 2: Percentage of correct alignments for Universal SailAlign (Section 2.1.1) and XLS-R (Section 2.1.2)

Speaker	Gender	Total recording duration
NK9dIF	F	4h 44m 45s
xoVY9q	M	4h 36m 12s
9G75fk	F	1h 44m 03s
n5WzHj	M	3h 44m 04s

Table 2: Total recording duration per speaker for the original (i.e., pre-segmented) recordings

To be able to split the audio files while retaining their transcription, we use the alignment files from Section 2.1. We split the files based on i) a silence duration threshold between two words ("pause duration"), which we set to 0.3 and ii) a minimum number of words per segment, which we set to two [7]. Campione and Véronis (2002) studied silent pause durations based on the analysis of 5 ½ hours of speech in five Indo-European languages, and categorized silences in brief (< 0.2s), medium (0.2 - 1s) and long (> 1s) pauses. Therefore, we suggest using a pause duration threshold between 0.2 and 1 second to segment the audio.

The final step of the data processing pipeline splits the audio segments-transcription pairs into a training, validation, and test dataset (80-10-10 respectively).

2.3 ASR fine-tuning

As mentioned in the introduction, in low-resource settings one typically fine-tunes a model that has been pre-trained on several hundred hours of a related language. In our pipeline, we are using a language-family-specific version of the wav2vec2 XLS-R model (Babu et al., 2022) that has been exposed to 56k hours in 53 languages[8]. The script that allows one to create a dataset and fine-tune a HuggingFace XLS-R model based on the language family is available at https://gitlab.com/ilsp-spmd-all/filotis/speech_to_text.

[7]The segmentor is available at https://gitlab.com/ilsp-spmd-all/filotis/silent-pause-segmentation

[8]huggingface.co/facebook/wav2vec2-large-xlsr-53

Model		WER	CER
Slavic model			
	Fine-tuned	9.06	3.12
	Baseline	87.31	31.47
Multilingual model			
	Fine-tuned	12.43	3.90
	Baseline	97.27	49.77

Table 3: ASR error rates for pre-trained ('Baseline') and fine-tuned Slavic and multilingual models (11h Pomak segments)

Slavic model	WER	CER
Fine-tuned (11 hours)	8.57	2.31
Fine-tuned (1 hour)	18.15	4.59
Baseline	87.14	30.13

Table 4: ASR model results on the 1-hour dataset split for the full fine-tuned model (11 hours), mini fine-tuned model (1 hour) and pre-trained Slavic model (baseline).

3 The case of Pomak

3.1 Recordings

Pomak has a rather weak online presence, which typically involves folk singing, so we could not simply crawl the web to create a dataset; only a few texts, and even fewer recordings, are available online. For instance, Salakidis et al. (2016) collected a few songs, recipes, and lullabies from the area of Thrace, including the Pomak community[9].

To build a Pomak corpus, we collected texts from different authors and sources (e.g., blogs and books) and included various genres (e.g. news items, folk tales, essays, biographical texts, short stories). We asked 4 native Pomak speakers to read the texts at the ILSP audio-visual studio in Xanthi, Greece. Pomak does not have an official script; the few existing texts are written in various alphabets: Cyrillic, Greek, IPA, and variations of the Latin alphabet (Karahóğa et al., 2022). For uniformity, all texts were converted to the alphabet presented in Karahóğa et al.

The duration of each recording ranges from 20 to 846 seconds, resulting in a total of over 14 hours. The total recording duration per speaker is displayed in Table 2. We also recorded a short free dialogue (4m 33s) between the two male speakers which we transcribed and added to the dataset.

3.2 ASR experiments: Low-resourced and endangered scenario

Using the proposed pipeline, we created a Pomak dataset to train our ASR model. Note that smaller segments also mean fewer pauses in the dataset. This results in a reduction of the total audio duration: The final duration of the audio files is 11 hours and 8 minutes in total.

To obtain a Pomak ASR model, we fine-tuned existing XLS-R models for 35 epochs[10] using the Pomak segments. Specifically, we fine-tuned i) an XLS-R model that had been exposed to Slavic languages (Ljubešić et al., 2022) ("Slavic model"[11]) and ii) an XLS-R model that had been exposed to 56 languages of the Common Voice dataset ("multilingual model"[12]). The results can be seen in Table 3. The fine-tuned Slavic model (i.e., the Slavic model that was further fine-tuned on the Pomak training set) has the lowest Word Error Rate (WER) and Character Error Rate (CER) on the test set. The multilingual model has a higher error rate, although it can also be useful if there is no language-family-specific pre-trained model available for the target language. The test set error rates of the pre-trained models are also given as a baseline and the best Pomak model (i.e., "Slavic fine-tuned") is available at https://huggingface.co/ilsp/wav2vec2-xls-r-slavic-pomak.

As for endangered languages, available recordings may consist of a few minutes or hours in total. Thus, we repeated the training using only one hour of speech. Specifically, we split the test set from Table 3 into three parts: training, validation, and test. In this new sub-dataset, the total recordings per speaker ranged from 13 to 20 minutes. We repeated the fine-tuning process of the baseline Slavic model for 35 epochs. While the error rates are higher than those reported in the full 11-hour model, the results are promising even with one hour of recorded speech (Table 4). Note that the 1h-dataset split is different than the 11h-dataset split reported in Table 3, therefore the baseline and fine-tuned error rates are also slightly different.

[9]Their recordings are available at http://ct-audiolink.eee.uniwa.gr/

[10]We initially fine-tuned for 100 epochs; the best checkpoints, based on the validation WER, were between the 30th and 40th epoch.

[11]The pre-trained (Baseline) Slavic model we selected is available at: https://huggingface.co/classla/wav2vec2-xls-r-parlaspeech-hr

[12]Pre-trained multilingual model: https://huggingface.co/voidful/wav2vec2-xlsr-multilingual-56

4 Conclusion

We presented a pipeline that facilitates data processing and enables ASR model training for low-resourced languages. Using this pipeline, we created the first transcription model in Pomak. We used the same dataset to train a TTS model, which we plan on using to augment the Pomak corpus.

Limitations

While we are confident that this pipeline can work for most low-resource languages, we have only tested it with Pomak, which belongs to the Slavic language family. Hugging face does not currently have pre-trained models for all language families (e.g., for indic). Therefore, for some low-resourced languages, a more generic (e.g., multilingual) pre-trained model will be selected, which will likely result in a higher error rate as shown in Table 3. Furthermore, as mentioned in Section 2.1, the wav2vec2 XLS-R-based alignments are heavily dependent on the ASR model used, while the Universal SailAlign-based alignments require an IPA dictionary of the target language. We have proposed a solution using a phonetic dictionary approximation, but this approach may also lack accuracy and it requires some manual verification. Last, the audio samples we used were of high quality as they were recorded in a studio. Noisy recordings are likely to result in less accurate i) alignments, ii) segmentations, and therefore iii) higher error rates.

Ethics Statement

All four participants have signed an informed consent form with Athena R.C. for their contribution to the narration of the voice samples.

Acknowledgements

We acknowledge support of this work by the project "PHILOTIS: State-of-the-art technologies for the recording, analysis and documentation of living languages" (MIS 5047429), which is implemented under the "Action for the Support of Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech*.

Estelle Campione and Jean Véronis. 2002. A large-scale multilingual study of silent pause duration. In *Speech prosody 2002, international conference*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *In S. S. Agrawal (Ed.), The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 205–209.

Nils Hjortnaes, Niko Partanen, Michael Rießler, and Francis M Tyers. 2020. Towards a speech recognizer for komi, an endangered and low-resource uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37.

Ritván Jusúf Karahóǧa, Panagiotis G Krimpas, Vivian Stamou, Vasileios Arampatzakis, Dimitrios Karamatskos, Vasileios Sevetlidis, Nikolaos Constantinides, Nikolaos Kokkas, George Pavlidis, and Stella Markantonatou. 2022. Morphologically annotated corpora of pomak. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 179–186.

Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and Shrikanth Narayanan. 2011. Sailalign: Robust long speech-text alignment. In *Proc. of workshop on new tools and methods for very-large scale phonetics research*.

Shreya Khare, Ashish R Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529–1533.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings 22*, pages 267–278. Springer.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. Parlaspeech-hr - a freely available asr dataset for croatian bootstrapped from the parlamint corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116.

Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose. 2016. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework. *Natural Language Engineering*, 22(6):907–938.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Georgios Salakidis, Evagelia Thomadaki, Christina Markou, Theodoros Kontogiorgis, Gavriil Kamaris, and John Mourjopoulos. 2016. A database of narrations and songs recordings with cultural interest from the area of thrace. In *8th conference 'Ακουστικη'*, pages 149–157.