Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2893–2900
Marseille, 11–16 May 2020

Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu

Alexander Zahrer, Andrej Žgank, Barbara Schuppler
Institute for Linguistics, University of Münster, Germany
Laboratory for Digital Signal Processing, University of Maribor, Slovenia
Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
a.zahrer@uni-muenster.de, andrej.zgank@um.si, b.schuppler@tugraz.at

Abstract

Since at least half of the world's 6000 plus languages will vanish during the 21st century, language documentation has become a rapidly growing field in linguistics. A fundamental challenge for language documentation is the "transcription bottleneck". Speech technology may deliver the decisive breakthrough for overcoming the transcription bottleneck. This paper presents first experiments from the development of *ASR4LD*, a new automatic speech recognition (ASR) based tool for language documentation (LD). The experiments are based on recordings from an ongoing documentation project for the endangered Muyu language in New Guinea. We compare phoneme recognition experiments with American English, Austrian German and Slovenian as source language and Muyu as target language. The Slovenian acoustic models achieve the by far best performance (43.71% PER) in comparison to 57.14% PER with American English, and 89.49% PER with Austrian German. Whereas part of the errors can be explained by phonetic variation, the recording mismatch poses a major problem. On the long term, *ASR4LD* will not only be an integral part of the ongoing documentation project of Muyu, but will be further developed in order to facilitate also the language documentation process of other language groups.
Keywords: Muyu, language documentation, phone recognition, American English, Austrian German, Slovenian

1. Introduction

Language documentation is a rapidly growing field in linguistics due to its urgency. At least half of the world's 6000 plus languages will vanish during the 21st century. Along with the speech communities themselves, linguists are assiduously working to document these languages through recordings, which they translate to global languages, for example English. A fundamental challenge for language documentation is the "transcription bottleneck". Transcribing spoken languages is a rather mechanic but very time-consuming task, with a recording-to-transcription-time-ratio of up to 1:60.

Speech technology, as a fast developing field in itself, may deliver the decisive breakthrough for overcoming the transcription bottleneck. The existing methods, however, are not directly applicable to language documentation, as good performing automatic speech recognition (ASR) systems are based on machine learning techniques requiring large-scale corpora. For the development of a transcription system suitable for language documentation, two main requirements need to be considered: 1) The tool needs to be able to deal with little or no data from the target language, and 2) it needs to integrate well into the work-cycles of a typical language documentation process. This leads to a constantly growing body of primary data and opportunity to increase the algorithm's performance in course of the language documentation process. A working transcription system can also serve as a good base for future projects at the intersection of language documentation and technology.

The aim of this paper is to report the first steps towards building such an ASR based language documentation tool (*ASR4LD*). It shows a comparison of experiments from different methods to build a phone recognizer for Muyu language. *ASR4LD* makes use of existing speech databases from well-resourced languages (i.e., American English, Austrian German, Slovenian). All primary data from Muyu for this project is taken from an ongoing documentation project for the Muyu language in New Guinea. Muyu is an endangered language spoken by around 2.000 people in the rain forest of West New Guinea, Indonesia. *ASR4LD* will be an integral part of the ongoing documentation project of Muyu. On the long term, we plan to make a tool available to the language documentation community that can easily be adapted to the challenges of another language of another language group.

The remainder of the paper is structured as follows. After presenting already existing ASR based tools applied to language documentation (1.1.), we show how *ASR4LD* may integrate into the typical language documentation workflow in 1.2.. Sections 1.3. and 1.4. present basic information on Muyu language and describe the phonological system of Muyu. After presenting the characteristics of the recorded test material (2.1.) and training material (2.2.), we present our experimental methods in 3.. We show the phone classification results for each source language separately, compare the classification with the different phoneme recognition experiments, and discuss some qualitative aspects of our results in 4.. Finally, we give an outlook to the next steps in the development of *ASR4LD* and conclude with some remark on its applicability to language documentation.

1.1. Existing ASR based approaches in language documentation

Previous attempts on using ASR technology for language documentation have been conducted with several languages, including Seneca (Jimerson et al., 2018), (Jimerson and Prud'hommeaux, 2018) and Yoloxóchitl Mixtec (Mitra et al., 2016), and Yongning Na (Adams et al., 2018), (Michaud et al., 2018). Though the value of this re-

search must be highly appreciated, these attempts are conducted in the context of long-term research (Yongning Na: 20 months over 12 years) or on languages with abundant data (Yoloxóchitl Mixtec: 125 hours of transcribed audio). Documentation projects on endangered languages typically lack both time and data.

A recent project connecting ASR technology with actual workflows in linguist fieldwork is Elpis (Foley et al., 2018). Elpis (the Endangered Language Pipeline and Inference System) was designed as a pipeline of tools arranged around a Kaldi based speech recognizer. Migration to new languages is facilitated via containerisation with Docker (Boudahmane et al.,). Fieldworkers have been trained to use Elpis in several workshops. To date Elpis has been tested successfully with 16 languages from the Asia-Pacific region using limited training data from 3 hours to less than 1 hour.

As other ASR-systems, Elpis makes use of dictionaries to specify the pronunciation of words. Therefore, it cannot be used in the earliest stages of language documentation when most lexical data is still missing. To overcome this obstacle, Persephone (Adams et al., 2018) was designed for phonemic and tonal transcription. Experiments on the languages Yongning Na and Chatino show that this is a promising approach even with low training data. It must be noted, however, that bare strings of phonemes and tonemes are unlikely to outperform a trained transcriber when a simple orthographic transcription is needed.

Benefits reported from research on ASR-assisted transcription exceed the mere gain in performance. Firstly, the faithfulness to the signal. Transcribers tend to overhear or ignore hesitations, repetitions or repairs. Nonetheless all of this can be important for subsequent analysis (Adams et al., 2018). Secondly, correction seems to level out performance differences. A time-sensitive experiment (Sperber et al., 2017) found that not only correction of their ASR-generated transcripts was 15% faster on average but also that slow transcribers gained up to 42%. Given the fact that a fieldworker often has to work with available assistants rather than most qualified ones, this is an important finding. Similar results have been reported in the area of machine translation (Green et al., 2013).

Some projects introduce new methods for data gathering mostly based on increasing access to the internet with the global spread of smartphones (Bird et al., 2014), (Blachon et al., 2016). Such approaches try to bypass the transcription bottleneck by depending on oral data altogether. Along with the primary data, a smartphone app records phrase-to-phrase translations. However, it is not clear if and how this kind of data will be used in further research since linguists depend heavily on written transcripts. Moreover, the use of smartphone based applications is not always possible in remote areas as internet connections may be sparse.

1.2. Workflows for language documentation

In order to apply ASR in language documentation, acoustic model training has to adapt smoothly into existing documentation workflows. Although linguists aim to use the data for their analysis, in recent years language documentation has gained reputation as a field of practice in its own right (Gippert et al., 2006) especially when dealing with endangered languages. Nonetheless, a language documentation as a "lasting multipurpose record of a language" (Himmelmann, 2006) is usually achieved through linguistic fieldwork (Newman and Ratliff, 2001), (Dixon, 2007), (Bowern, 2015).

Documentation projects often share the following features: low personal resources (often only one PhD student), limited time (1-3 years), several field trips (e.g., 3x4 months). While in the field, the field worker will record communicative events in the object language, transcribe the recordings with the help of native field assistants and translate them into a common lingua franca (e.g., English). Field trips are interrupted by several months of data analysis, reporting and preparation for the next field trip, all of which is usually carried out at the university where the linguist is based with only minor contact to the speakers.

Especially the first field trip is often exploratory and not rewarded with lots of high quality data. A top priority is, however, to investigate the phonological system of the object language in order to develop orthographic conventions for the next visit. So in this period (between first and second field trip) the linguist should by then have a good impression of all the phones he will encounter in the following months. In this phase there is not enough data collected yet to accurately train an acoustic model, thus we propose to use the target language in this phase to test which acoustic models trained on other well-resourced languages perform best. In addition, the small amount of data from this first field trip can already be used to train a phone-level language model (as we also do in this paper). The resulting phone-transcription tool can be used to provide a first transcription of the rest of the recordings, which is subsequently corrected by the linguist. In this stage, the tool does not yet speed up the work of the linguist, the required additional work for tool development might even cost the linguist more time than without the tool. We consider this time a good investment, as the following stages in the language documentation process profit from the use of the tool. This paper reports experiments of up to this stage.

The data from the first field trip will be used to automatically create a (small) pronunciation lexicon (with variants) from the annotated data and subsequently a word level recognizer can be developed. This transcription tool is then already available for the second field trip. Given the small amount of different word types available for the ASR system, high out of vocabulary (OOV) rates can be expected. In order to make sure that the transcription system actually facilitates the work of the linguist, the tool will provide the transcription for words with high system-internal confidence, but coloured questionmarks in places of low confidence. We expect that having this tool at hand during the second trip, a good amount of data can be recorded and annotated. After the second trip, the word-level recognizer can be adapted with the new material available: acoustic models can be adapted with the material from the target language or even be trained on it, the lexicon can be increased, Forced Alignment tasks will be used to develop a variant lexicon, and overall, the transcription system can be improved such that for the third field trip it will actually speed

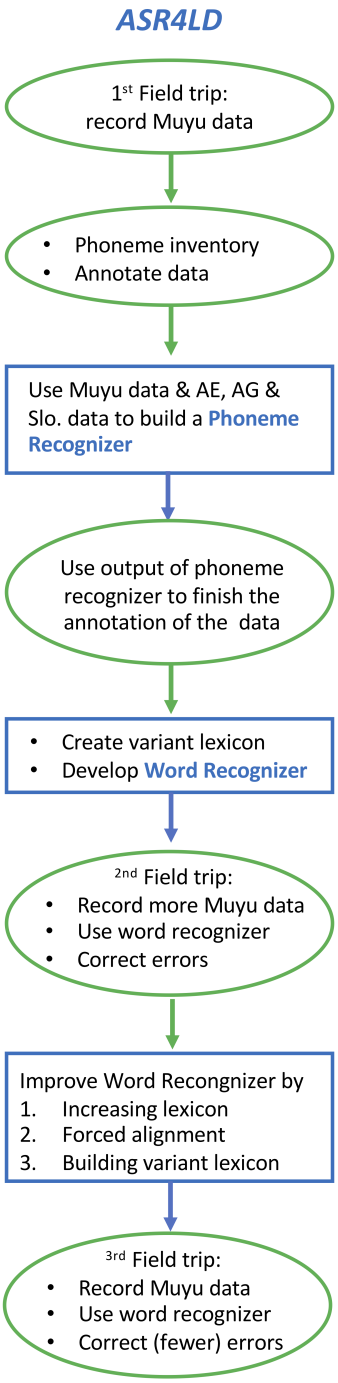

Figure 1: **How *ASR4LD* integrates in the typical language documentation workflow**. The oval, green boxes show working steps of the linguist, the squared, blue boxes those of the speech technologist.

up the transcription process. In this proposed workflow, the linguist is not only supported by the tool in creating transcriptions, but also secondary language resources (lexica, variant lexica) are developed at the same time. Figure 1 shows how the envisioned tool *ASR4LD* integrates in the currently ongoing Muyu language documentation project.

Phoneme	Allophones	Examples
/b/	[b], [p]	AG **[b]**uch *'book'*, AG de**[p]** *'fool'* - no aspiration
/t/	[t], [d]	AG **[t]**elefon - no aspiration, AG o**[d]**er *'or'*
/k/	[k], [g]	AG **[k]**ann *'can'* - no aspiration, AE **[g]**ame
/m/	[m]	AE **[m]**ake
/n/	[n]	AE **[n]**oise
/ŋ/	[ŋ]	AG Leitu**[ŋ]** *'wire'*
/l/	[l], [r], [ɾ]	AG a**[l]**e *'all'*, SL v**[r]**at *'neck'* (short trill with tip of the tongue), AE le**[ɾ]**er (single flap)
/ʋ/	[ʋ], [w]	AG **[ʋ]**enden *'to turn'* (lower lip to upper teeth), AE **[w]**indow (round lips)
/j/	[j]	AE **[j]**oung
/a/	[a]	SL kr**[a]**j *'region'*
/e/	[ɛ]	SL *ž***[ɛ]***na'wife'*
/i/	[i]	SL d**[i]**m *'smoke'*
/o/	[ɔ]	SL m**[ɔ]***č'power'*
/u/	[u]	AG **[u]**nter *'under'*

Table 1: Phonomes and allophones of Muyu including similar sounds in Austrian German (AG), American English (AE) or Slovenian (SL).

1.3. Muyu language

Muyu (kti, kts) is an under-documented Papuan language (non-Austronesian) spoken by probably 2.000 speakers in the central lowlands of West New Guinea, Indonesia. It belongs to the lowland Ok branch of the Trans New Guinea (TNG) language family. Due to the lack of transmission to a younger generation of speakers, Muyu is considered severely endangered. At least nine mutually intelligible dialects have been reported which differ in lexical material and phonetic realisation of a shared phonemic inventory. The data for this paper focuses on the Kawip dialect of Muyu.

In several respects, Muyu exhibits typical features of TNG languages: SOV as most frequent word order, clear grammatical distinction between nouns and verbs, little or no morphology on the noun, complex verb morphology. Unlike other well studied Ok languages like Mian (Fedden, 2011) and Telefol (Healey, 1964) lexical tone is not a feature of Muyu.

1.4. Muyu phonetics and phonology

Table 1 illustrates phonemes and their allophones. In the right column, similar sounds in American English, Austrian German or Slovene are given.

The vowel system comprises of five phonemes differentiated in vowel quality: /a, e, i, o, u/. Mid-vowels are realized rather open ([ɛ] and [ɔ]). There are no oppositions rounded/spread lips. Likewise an opposition in vowel quantity cannot be stated safely yet, minimal pairs tend to be rare and restricted to /a/ only. Word stress seems to be an intervening factor with longer vowels in stressed syllables.

However, the analysis of vowel duration is not completed yet.

The consonant system mostly relies on a three-way distinction in plosives /b, t, k/ and nasals /m, n, ŋ/. Additionally, there is a liquid /l/, realized as one of three allophones [l, ɾ, ɾ], and two approximants /ʋ/ (allophonic labio-dental [ʋ] before [i], [ɛ] or rounded [w] before [u], [o], [a]) and /j/. Most noticeable in Muyu phonology is the total lack of fricatives both as single consonants and in affricates. The obstruents /t/ and /k/ appear as voiced [d] and [g] when preceded by a vowel or a homorganic nasal (i.e. /n/ or /ŋ/ respectively) in the same prosodic word. Contrary to the alveolar and velar plosives, /b/ is voiced as a default and appears as devoiced [p] when realized as syllable coda. Allophones of the liquid /l/ appear exclusively intervocalic. Allophonic variation seems to be of dialectal origin though further investigation is needed.

Muyu phonology strongly disfavours complex syllable structures. Syllable-medial consonant clusters are excluded. Most frequent syllables are CV, VC, and CVC. Consonantic onsets are not obligatory. Also bare V-syllables occur word initially. Although isolated lexical items prefer closed syllables, connected speech tends to resyllabification when following words start with an initial vowels (/nup ambip/ 'our house' = [nu.wam.bip], incl. lenition). Muyu lexemes are typically di- or trisyllabic. However, longer word forms appear quite frequently due to complex verb morphology.

2. Materials

2.1. Muyu recordings

Recordings were made during a 4-month field trip in 2019. For this paper we used a subset of the recordings which comprises of 123 min of monologues with a total of 7073 transcribed tokens. 61 min and 6121 tokens are narratives from 6 male speakers, 62 min and 952 tokens are part of a reading task with 4 male speakers where speakers had to read aloud single words presented on a screen. Speakers of the two tasks were partially overlapping which leads to a total of 8 male speakers in total. Our recordings do not contain female speakers yet. All materials are fully transcribed and translated to English. The working orthography for the transcription was based on the Latin script as used in Indonesian since this is the language of education in the area. Figure 2 shows the spectrogram of a typical single-word-utterance of the reading task.

The narratives were recorded on video using a Canon XA40 on a tripod and a Audio-Technica AT831b lavalier microphone. The reading task was recorded on audio only with a Tascam DR-40 audio recorder and a head-mounted AKG C520 microphone.

The recordings were made in the village Upyetetko. All but one narratives were recorded indoors at different times of the day depending on the schedules of the speakers. Recording quality varies according to background noise of playing children, crowing roosters, passing motor-cycles etc. Another issue in recording free narratives is to foresee volume peaks. Speakers tend do get emotionally involved and speak up at several parts of their stories. Audio recording levels have to be adjusted accordingly to avoid overdrive. Therefore, the volume of the recordings differ from speaker to speaker.

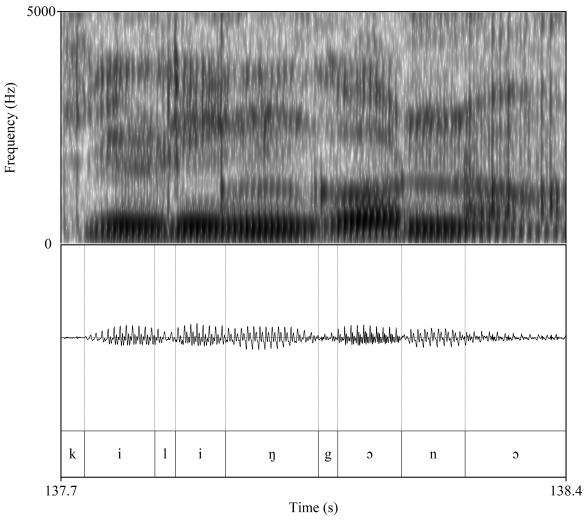

Figure 2: Spectrogram of the word *kilinggono* 'before'.

2.2. Training materials

The monolingual source acoustic models were built for three languages: American English, Austrian German and Slovenian. The American English speech recognition system was trained on the freely available LibriSpeech database (Panayotov et al., 2015). The LibriSpeech database comprises audiobook recordings, where books were read by 2087 speakers. Out of in total 982 hours of recordings in the database, we used the 100 hours clean data training subset. The reason for this choice was to have all source language resources of a comparable size. The audio material is of high quality, recorded in a quiet indoor environment, with 251 speakers.

The Slovenian BNSI Broadcast News speech database (Žgank et al., 2005) consists of the daily evening and late-night TV news shows recorded from the broadcaster's archive. There are 42 news shows in a total duration of 36 hours available. Altogether, the BNSI speech database has 1,565 speakers (1,069 male, 477 female, 19 unknown), which is significantly more than the other two speech databases used in this work. The BNSI speech database recording conditions and speech style are also different. The speech in news shows, which can be read, prepared or spontaneous, is often accompanied by other sources, such as audio background from the video footage. This makes the BNSI speech database conditions more diverse and challenging than for the other two languages.

The GRASS corpus (Schuppler et al., 2017) was used for Austrian German speech recognition modeling. It comprises read and spontaneous speech, which are placed into database subsets. The total duration of the recordings is 32 hours, the speech was uttered by 38 speakers (19 female, 19 male). The recording took place in a professional recording studio, which results in high acoustic quality. The orthographic transcriptions for Austrian GRASS and Slovenian BNSI were prepared manually, but the phonetic segmenta-

tion was generated with forced re-aligning.
All three speech databases were recorded in controlled environments, which is in contrast to the field environment, where usually endangered languages such as Muyu are being recorded. This difference between language resources could lead to additional accuracy reduction along with the one produced by the acoustic-phonetic differences between the languages involved.

3. Methods

3.1. Phone set

Languages from three different language groups are covered in this work. The number of phones applied to speech recognition can differ from the number defined for a particular language by linguists. The main cause is phone frequency, which results in the inability to collect enough training material for in-frequent phones. The number of phones used for the speech recognition experiments for each of the languages is:

- American English (AE): 40
- Austrian German (AG): 33
- Muyu (M): 18
- Slovenian (Slo): 40

The small number of Muyu phones is a benefit for cross-lingual speech recognition, as most of them have an equivalent matching pair in the source languages. This is important for the zero resource speech recognition task, where no target language spoken resources are used. The mapping from the Muyu phoneme set to the phoneme sets of the source languages was based on expert knowledge.

3.2. Training of acoustic models

The automatic speech recognition systems applied were trained using the Kaldi toolkit (Povey et al., 2011). Separate monolingual speech recognition systems were trained for each of the three source languages. The training set recordings were first converted into Mel Frequency Cepstrum Coefficients (MFCC) together with first- and second-order derivatives. A final feature vector with 39 values was calculated per each 25 ms frame, which was shifted by 10 ms.
The 3-states left-right GMM-HMM acoustic models were first trained as monophones and then the context was expanded to triphones. The training set transcriptions were improved using the forced re-aligning procedure. The resulting triphone system was applied to train the DNN-HMM speech recognition system using the p-norm feed-forward deep neural networks. The DNN training procedure with 30 epochs was frame-level based, where the architecture had two hidden layers.

3.3. Free phone recognition task

The free phone recognition task is the most difficult case of cross-lingual speech recognition, as no Muyu spoken language resources is available. A simple phone-loop grammar, with equal probabilities for all phones, was built. This resulted in an effect where only the acoustic model performed the cross-lingual speech recognition role and full focus was given to the acoustic-phonetic similarities between the phone pairs, as defined by an expert. The complexity of the phone-loop grammar is entirely proportional to the number of target phones, where Muyu has the advantage of a small number of phones.

3.4. Phone recognition with language models

The available Muyu recordings and their transcriptions were split into training and test sets. The training set was used only for training the phone-based language model from the transcriptions. After removing the examples of code-switching, the training set transcriptions contained 572 utterances with 23k phones. The phone-based bigram and trigram language models were generated from this small corpus of Muyu transcriptions and later used for cross-lingual speech recognition.

4. Results and discussion

The speech recognition evaluation was carried out for different configurations. Word Error Rate (WER) and Phone Error Rate (PER) were used as a metric. First, the three source speech recognition systems were tested, to assess their monolingual performance. The WER results are given in Table 2.

Monolingual system	WER(%)
American English	9.21
Austrian German	11.25
Slovenian	31.39

Table 2: Monolingual American English, Austrian German and Slovenian speech recognition results

The American English and Austrian German speech recognizers achieved comparable results, while the speech recognition performance for the Slovenian system was lower by approx. 20% absolute, mainly because of more degraded acoustic conditions and highly inflected, free word order language characteristics. The achieved monolingual speech recognition results for source languages are comparable with similar other ASR systems for these languages.
A dedicated test set from 476 field recordings was defined for Muyu. Two speakers uttered a set of 238 isolated words each. This test scenario was the most convenient for the proposed cross-lingual speech recognition experiments with zero speech resources.

4.1. Results for free phone recognition

The zero resources Muyu cross-lingual phone recognition used PER as a metric. It was first carried out with a phone-loop grammar, with equal probabilities for all phones. The results are given in Table 3.
The best Muyu cross-lingual free phone speech recognition result was achieved with the Slovenian acoustic models. The phone error rate (PER) was 48.26%. The PER degraded to 67.88% with the American English data. The worst result was achieved with Austrian German, where the PER was as high as 91.15%.

Cross-lingual source AM	PER(%)
American English	67.88
Austrian German	91.15
Slovenian	48.26

Table 3: Muyu cross-lingual speech recognition results with the free phone recognition system

4.2. Results for phone recognition with language model

The second part of Muyu cross-lingual speech recognition included the phone-based n-gram language models, combined with 3 different monolingual source acoustic models. The results in the form of PER are given in Table 4.

Cross-lingual source AM	2g PER(%)	3g PER(%)
American English	57.14	57.14
Austrian German	91.30	89.49
Slovenian	44.19	43.71

Table 4: Muyu cross-lingual speech recognition results with phone-based bigram and trigram language models

The usage of language models instead of a phone-loop grammar significantly improved the cross-lingual speech recognition performance for most configurations, as it was expected. The best overall phone recognition result was again achieved with the Slovenian acoustic models. The PER reduced from 48.26% for the experiment with the phone-loop grammar to 44.19% with bigrams and further to 43.71% with trigrams, respectively. Even a larger reduction of PER occurred for American English. The 67.88% PER for the recognition experiment with the phone-loop grammar was reduced with language models to 57.14%. The American English produced the same PER with bigram and trigram language models. With trigrams, the number of substitutions was higher and the number of deletions and insertions was lower. In the case of Austrian German, the cross-lingual results were again the worst, with almost no improvement by the language models. It is also important to notice that in general, the phone-based trigram language model did not improve significantly the cross-lingual speech recognition results over the bigram language model. The probable cause is the limited size of the Muyu language model training corpora.

4.3. Qualitative evaluation and discussion

Unsurprisingly, PER depended upon the underlying phonological system of the language. In the Slovenian language model, for example, difficulties arose with recognizing the approximant [w] before vowels which is not attested in the Standard Slovenian phonology (Šuštaršič et al., 1995). In contrast, [j] before vowels (attested in Slovenian) was recognized correctly most of the times. The same holds true for nasals of which the velar position [ŋ] is not used in Slovenian whereas [m] and [n] are recognized quite satisfyingly.

Another issue was with voiceless plosives in final position. At the end of Muyu lexemes [p, t, k] are quite frequent. However, in most cases these voiceless offsets do not remain intact. Single word utterances tend to delete the final burst and therefore end with a mere closure [p̚, t̚, k̚]. These closures are rarely recognized as segments by the language models in our experiment.

As already mentioned earlier, the recording mismatch of the data poses a big problem. This is indirectly also reflected in our results, as cross-lingual recognition with acoustic models trained on the best recording quality (i.e., the Austrian German *GRASS* database) has the worst performance, and this despite the high overlap in the typical realization of phonemes. Since in Austrian German, plosives are not aspirated, as in German German or English, and since the Muyu vowels all exist also in Austrian German (except for a open mid back vowel), we actually had hoped for a high performance with Austrian German acoustic models. Our results thus show the importance of matching recording conditions in addition to phonetic similarity.

4.4. Future directions

In order to account for the earlier mentioned issues with recording quality, two directions will be taken in our future work. First, we will have to implement more sophisticated speech enhancement methods (Vincent et al., 2017) than the ones implemented automatically by the speech recognition toolkit Kaldi. One possibility will be, to use multi-conditioned training (Soni et al., 2019), the other one to use robust processing (Suh et al., 2007). Second, since one of the research areas of the SPSC laboratory is on room acoustics, there is the plan to design a portable recording booth. The envisioned booth should be easy to build up (e.g., similar to a pop up tent) and be made of a light material of high acoustic absorbance. The idea is to test a prototype during the upcoming field trips to Indonesia.

The results presented here are the first speech recognition experiments in the development of the *ASR4LD* toolkit presented in 1.2.. The recognition is of high enough performance to go to the next steps: annotating data, building a better word recognizer, increasing the lexicon, and using forced alignment to extract pronunciation variants , to improve the transcription system.

5. Conclusions

The transcription bottleneck in language documentation is an obstacle that can be tackled from early stages on. Endangered languages are particularly challenging. Time and resources for fieldwork are often very limited. Therefore, it makes sense to start on a phonemic level when lexical data are still sparse.

Training language models from other languages can help to approximate phoneme recognition in the target language. Our best results with 43.71% for Slovenian as a source language are not good enough yet to compete with human transcribers. Correcting almost half of the phonemes would actually slow down the transcription process compared to transcribing manually from scratch. Nevertheless, the time spend by the transcriber in this phase (after first field trip) does not only create transcriptions (as primary resource),

but also contributes to the creation of pronunciation dictionaries (incl. pronunciation variants). In addition, from experiments with annotating well resourced languages, we know that an ASR based reference yields to higher inter and intra- labeller agreement. The latest by the time of the second field trip, the tool can be expected to also facilitate the transcription process.

Building this first version of *ASR4LD* did also affect our understanding of the target language. Comparing sounds of Muyu to three other languages added to our phonetic analysis of differences and similarities of the Muyu phonemes with those of well-known and well-studied languages. With the help of the ASR system, linguistic hypotheses about the phoneme system and phonetic variation in the language under investigation can be tested by evaluating the recognition performance on particular phones (e.g., on Muyu [w] using a Slovenian language model). We therefore conclude, that phoneme recognition is not only a first step towards building an automatic transcription system for language documentation, it also contributes to new insights on the target language and improves the quality of phonetic analysis.

6. Acknowledgements

The research of Alexander Zahrer was supported by ELDP (Endangered Languages Documentation Programme) Individual Graduate Scholarship 0367. The research of Andrej Žgank was partially funded by the Slovenian Research Agency under Contract number P2-0069 "Advanced methods of interaction in telecommunication".

7. Bibliographical References

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018*, pages 3356–3365.

Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5.

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, 81:61–66.

Boudahmane, K., Manta, M., Antoine, F., Galliano, S., and Barras, C.). Docker container software. `http://transag.sourceforge.net`.

Bowern, C. (2015). *Linguistic fieldwork: A practical guide*. Springer.

Dixon, R. M. (2007). Field linguistics: a minor manual. *STUF–Sprachtypologie und Universalienforschung*, 60(1):12–31.

Fedden, S. (2011). *A grammar of Mian*, volume 55. Walter de Gruyter.

Foley, B., Arnold, J. T., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., Heath, S., Kratochvil, F., Maxwell-Smith, Z., Nash, D., et al. (2018). Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *SLTU*, pages 205–209.

Gippert, J., Himmelmann, N. P., and Mosel, U. (2006). *Essentials of language documentation*, volume 178. Walter de gruyter.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448. ACM.

Healey, A. (1964). *The Ok Language Family in New Guinea*. The Australian National University.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for. In Jost Gippert, et al., editors, *Essentials of language documentation*, volume 178, pages 1–30. Mouton de Gruyter Berlin.

Jimerson, R. and Prud'hommeaux, E. (2018). Asr for documenting acutely under-resourced indigenous languages. In *LREC 2018*, pages 4161–4166.

Jimerson, R., Simha, K., Ptucha, R. W., and Prudhommeaux, E. (2018). Improving asr output for endangered language documentation. In *SLTU*, pages 187–191.

Michaud, A., Adams, O., Cohn, T. A., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit. *Language Documentation & Conservation*, 12:393–429.

Mitra, V., Kathol, A., Amith, J. D., and García, R. C. (2016). Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *INTERSPEECH 2016*, pages 3076–3080.

Newman, P. and Ratliff, M. (2001). *Linguistic fieldwork*. Cambridge University Press.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Schuppler, B., Hagmüller, M., and Zahrer, A. (2017). A corpus of read and conversational Austrian German. *Speech Communication*, 94C:62–74.

Soni, M., Joshi, S., and Panda, A. (2019). Generative Noise Modeling and Channel Simulation for Robust Speech Recognition in Unseen Conditions. In *INTERSPEECH 2019*, pages 441–445.

Sperber, M., Neubig, G., Niehues, J., Nakamura, S., and Waibel, A. (2017). Transcribing against time. *Speech Communication*, 93:20–30.

Suh, Y., Sungtak, K., and Hoirin, K. (2007). Compensating acoustic mismatch using class-based histogram equalization for robust speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2007, 01.

Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J., and Marxer, R. (2017). An analysis of environment, micro-

phone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535 – 557.

Šuštaršič, R., Komar, S., and Petek, B. (1995). Slovene. *Journal of the International Phonetic Association*, 25(2):86–90.

Žgank, A., Verdonik, D., Markuš, A. Z., and Kačič, Z. (2005). Bnsi slovenian broadcast news database-speech and text corpus. In *Ninth European Conference on Speech Communication and Technology*.