Connecting Automated Speech Recognition to Transcription Practices

Blaine Billings
University of Hawai'i at Mānoa
blainetb@hawaii.edu

Bradley McDonnell
University of Hawai'i at Mānoa
mcdonn@hawaii.edu

Johan Safri

Wawan Sahrozi

Abstract

One of the greatest issues facing documentary linguists is the transcription bottleneck. While the large quantity of audio and video data generated as part of a documentary project serves as a long-lasting record of the language, without corresponding text transcriptions, it remains largely inaccessible for revitalization efforts and linguistic analysis. Automated Speech Recognition (ASR) is frequently proposed as the solution to this problem. However, two issues often prevent documentary linguists from making use of ASR models: 1) the thought that the typical documentary project does not have sufficient data to develop an adequate ASR model and 2) that correcting the output of an ASR model would be more time-consuming for transcribers than simply creating a transcription from scratch. In this paper, we tackle both of these issues by developing an ASR model in the larger context of a documentation project for Nasal, a low-resource language of western Indonesia. Fine-tuning a larger pre-trained language model on 25 hours of transcribed Nasal speech, we produce a model that has a 44% word error rate. Despite this relatively high error rate, tests comparing speed of transcribing from scratch and correcting ASR-generated transcripts show that the ASR model can significantly speed up the transcription process.

1 Introduction

The use of Automated Speech Recognition (ASR) in language documentation and revitalization contexts has been met with almost universal enthusiasm due to its promise to loosen the transcription bottleneck (see e.g. Berez-Kroeker et al., 2023). The basic idea behind this approach is that while the limited data of low resource languages is often not able to produce highly accurate models comparable with those of high resource languages, it is more efficient to correct the output of an ASR model than to produce a transcription from scratch (see Foley et al., 2018; Bird, 2021). This approach crucially relies on an ASR model to generate transcriptions accurate enough that making corrections requires less time and effort than creating a transcription anew. More recently, there is a growing number of ASR studies that demonstrate how the accuracy of models with relatively little training data are able to be improved through the use of pre-trained models of high resource languages (Coto-Solano et al., 2022) or supplemental written corpora in the target language (Bartelds et al., 2023; San et al., 2023). Despite improvements to ASR models in language documentation and revitalization projects, it remains difficult to assess the usefulness of such models as there has been very little reported about how transcribers, who are often native speakers and members of the speech community, interact with the outputs of these ASR models.

In this paper, we address this issue through a case study of Nasal, an endangered, under-resourced Austronesian language of Sumatra, Indonesia. The study comprises two parts. The first discusses the development of an ASR model for Nasal through the fine-tuning of a pre-trained high-resource language model using Whisper (Whisper 2024). The second addresses the usefulness of such a model for Nasal transcribers by comparing the process of transcribing in ELAN from scratch against correcting transcriptions generated by the ASR model.

This paper is organized as follows. The remainder of this section provides an introduction to the Nasal speech community (§1.1) and the ongoing Nasal documentation project (§1.2). §2 describes the development of the ASR model for Nasal. §3 discusses the results from the model and the comparison between the two transcription methods. §4 provides some discussion on the viability of ASR models for documentation projects, and §5 gives some summary remarks.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 128–132

1.1 The Nasal speech community

Nasal [glottocode: nasa1239] is a Sumatran language spoken by approximately 3,000 people in southwest Sumatra, Indonesia (Billings and McDonnell, 2024). The Nasal speech community represents a fringe case of small-scale multilingualism (Pakendorf et al., 2021) where, in addition to Nasal, members of the community use two regionally significant varieties of Malay – Kaur [glottocode: kaur1269] and South Barisan Malay [glottocode: cent2053] – in daily life. Nasal was not known to linguists until 2007 (Anderbeck and Aprilani, 2013) and thus little documentation of the language existed until the authors, a team of outsider linguists and members of the Nasal speech community, initiated a documentation project in 2017 that continues to the present (McDonnell 2017; McDonnell et al. ongoing). At the onset of the project, several members of the Nasal community, including the third and fourth authors, were provided training in a simplified system of Discourse Transcription (Du Bois et al., 1993), methods for free translations into Indonesian (Schultze-Berndt, 2006), and ELAN (ELAN 2024) and Fieldworks Language Explorer (FLEx; Fieldworks 2024) software.

1.2 Documentation on Nasal

The documentation of Nasal consists of audiovisual recordings of everyday conversations, culturally important events, active elicitation sessions to elicit and discuss word meanings, and structured tasks to elicit aspects of Nasal phonology and grammar. In the vast majority of recordings, speakers were recorded on separate channels using lapel or headset microphones. Recording in this way better facilitates the ability to train, test, and use ASR models on conversational data by targeting individual speaker audio and reducing signal bleeding.

The largest portion of the documentation consists of everyday conversations, followed by active elicitation sessions. The majority of recordings that fall into the prior category have been transcribed using Discourse Transcription and translated into Indonesian using ELAN and later glossed in FLEx. However, the majority of the recordings that fall into the latter category have yet to be transcribed or translated. The aim of these active elicitation sessions is to discuss a large number of lexical items with their associated meanings and uses by facilitating conversations of various semantic domains. This documentation forms the basis of a Nasal dictionary project.

At the outset of this documentation project, project leaders (which includes the second author) hosted a series of meetings to discuss issues such as project outcomes and orthography development as well as training workshops in Discourse Transcription, ELAN, and FLEx at the Atma Jaya Catholic University of Indonesia. The third and fourth authors participated in the workshop. During the workshop, they began producing transcriptions, and with the help of project team members with linguistics training, they began transcribing recordings of conversations. Over subsequent years, Nasal transcribers have honed transcriptions and translations as well as their methods for transcribing. Currently, transcriptions and corresponding translations are produced by the third and fourth authors in ELAN with the following workflow:

1. **Segmentation**: Segment recording into Intonation Units in *Segmentation Mode* in ELAN
2. **Transcription**: Fill empty annotations with Nasal transcriptions
3. **Discourse & translation**: Input corresponding discourse transcriptions and Indonesian translations necessary for analysis
4. **Context**: Create additional annotations for various types of sporadic notes (speech context, code-switching, etc.)

Of these four steps, the one to be addressed by the ASR model is *transcription*. Depending on the granularity of transcription, it can take upwards of forty minutes to transcribe one minute of audio (Seifart et al., 2018), often requiring listening to each individual annotation up to five or ten times. Given its drastically greater time requirement over the other steps, we decided to work on implementing ASR for transcription first with plans to tackle the remaining three in the future.

2 Methods

2.1 Data preparation

The data used for training and testing the ASR model consist of transcribed audio from twenty-five recordings of various genres: everyday conversation (13), brief map game and role-play tasks originally recorded for prosody elicitation (10), and semantic domain active elicitation sessions recorded for dictionary development (2). Sessions lasted anywhere from fifteen minutes to three hours for approximately 25 total hours of recording time.

Each session included two to four speakers with 49 unique speakers in total (seven appeared in two recordings, one appeared in three recordings). While nearly all of the dictionary elicitation sessions feature both the third and fourth author as facilitators, one or two additional speakers, differing by recording, are also present, and thus training the model on a diverse speaker population accords with the intended use case. The transcriptions for these recordings were produced by the third and fourth authors over a period of four years. The text for these transcriptions constitute the corresponding text input for the language model. Audio segments corresponding to the timestamps for the annotations were extracted from each individual speaker audio file. This training data for the ASR model totaled more than 160,000 words in 66,500 annotations accounting for 17.5 hours of speech time (that is, excluding all silence, i.e. non-annotated segments, from each speaker audio file).

2.2 ASR training

The data was split into two sets, training data and testing data, with a simple 80/20 division of the annotations, respectively. The ASR model was built by fine-tuning the small model from Whisper (Whisper 2024), using the small model's pretrained tokenizer and feature extractor from Indonesian, a related language. Fine-tuning ran over 5,000 steps with evaluation according to word error rate (WER) taken at every 500-step checkpoint. The best of these checkpoints was used in generating the transcriptions for the transcription task.

2.3 Transcription task

In order to determine the viability of using ASR-generated transcriptions over transcribing recordings from scratch, we designed a short transcription task. In this task, the first author selected two excerpts, one from a conversation and the other from an active elicitation session and neither of which was used in training and testing the model. The 2 minute and 30 second excerpts were segmented in ELAN, leaving empty annotations. The third and fourth authors then each produced four transcriptions on the two excerpts. The third author transcribed the elicitation session first and the conversation second, while the fourth author transcribed the conversation first and the elicitation session second. Both started their first file by correcting the ASR-generated transcript and then transcribed from scratch, whereas with the second file, they first transcribed from scratch and then corrected the ASR-generated transcript. The task was designed in this way so as to balance any confounding influence from the order of session or transcription method. The third and fourth authors screen-recorded the process of transcribing each of the four files and later compared their experiences in each.

3 Results

3.1 ASR results

Over the 5,000 steps of fine-tuning the ASR model, the lowest error rate attained in the training checkpoints was 43.9%, a significant improvement over the 67.2% of the previous model trained on Nasal data (San et al., 2023). When tested against two segments of audio not included in the testing set (one from an everyday conversation, one from an active elicitation session), WER was higher at 60.1% (conversation) and 54.1% (active elicitation). Character error rate (CER) was similarly calculated for both to gain a better understanding of the kinds of errors made by the model. These came out to be 21.4% (conversation) and 20.4% (active elicitation), corresponding as expected with the WER above given the distribution of word length in Nasal. On further inspection of the ASR output, these rates were found to be inflated by errors with interjections (e.g. transcribing a single syllable *m* rather than a two syllable *mm*) and orthographic variations introduced by the training data (see *Limitations* at the end of this paper).

3.2 Transcription comparison results

After completing the transcription task, the third and fourth authors found that correcting the ASR-generated transcriptions was able to significantly speed up transcription time, with all four tests showing improvements. It is unsurprising that the gains were higher (23.49% compared to 11.30%, 32.29% compared to 21.92%) when the ASR-assisted method occurred second — as the authors had already been exposed to the media once — but the improvement is nonetheless apparent. Although many corrections needed to be made to the generated transcripts, changes were most often minor, single-letter or single-word edits and rarely required reannotating an entire IU. Furthermore, revising the automatic transcriptions was preferred over transcribing from scratch, since having a baseline of transcribed text meant the audio needed to

be listened to fewer times in order for the recorded speech to be accurately determined.

4 Discussion

Documentary projects typically result in the production of a large body of audio and video recordings of various genres, from narratives to conversation and elicitation. Whether to assist in the production of language materials or in linguistic analysis, creating transcriptions of these recordings is often a normal part of a documentary linguists' workflow. As has been demonstrated here for Nasal, once a small body of transcribed audio data has been produced, these transcriptions can be leveraged to fine-tune an ASR model to speed up the transcription of remaining or future documentary data. The authors, who themselves work directly with transcribing the Nasal data, have found that correcting the transcriptions generated by a model trained on such data significantly aids in the transcription process.

One of the primary motivations for developing this ASR model for Nasal is the ongoing compilation of the Nasal dictionary. A large corpus of active elicitation sessions, now totaling 70 hours of data, remains to be transcribed. Transcriptions of these sessions – many of which contain lexical items absent from the corpus of everyday conversation – would make the data more usable and more easily linked with dictionary outputs. Although recordings of active elicitation sessions contain a greater frequency of new lexical items, the ASR model developed here did not show significant differences in accuracy in transcribing the active elicitation recording and the everyday conversation recording, proving equally useful for the dictionary compilation process.

As discussed above (§1.2), the transcription workflow in the Nasal documentation project consists of four steps. Since transcription time is the most significant problem in this workflow, we decided to tackle it first. ASR-generated transcriptions for Nasal speech have already proven to significantly speed up the transcription process. Addressing the remaining three steps could further contribute to faster transcription of documentary data. For example, further AI models targeting prosody and intonation could be implemented to speed up the IU-based segmentation used in our transcriptions. Discourse transcription will likely need to remain manual, but machine translation has also shown promise in low resource contexts (see van Esch et al., 2019) and Whisper AI may even prove to be useful in this regard (see the description at Whisper 2024). Finally, the creation of additional contextual notes, while important for linguistic analysis, results in less than two percent of the total number of annotations and thus needs not be immediately addressed with computer-assisted methods.

5 Conclusion

We fine-tuned a pre-trained ASR model with 25 hours of data typical to a documentation project. Through comparing the processes of correcting transcripts generated from this model and transcribing from scratch, it was demonstrated that such a model proves effective for improving the transcription workflow and reducing the amount of time necessary for transcribing documentary data. We believe that such models are more accessible to documentary linguists than typically thought and can greatly assist in the transcription process.

Limitations and future prospects

In reviewing the ASR-generated transcripts, it was clear to the authors that a major contributor to the increased WER and CER was the lack of standardization in the Nasal orthography. For example, many words contain two-vowel sequences and can be written with or without a predictably inserted glide (e.g., *gauh*, *gawuh* 'just' are both valid written forms). In other cases, a shortened form of a word used in rapid speech is variably reflected in the transcripts either by the longer or the shorter form (e.g., either *jenu*, *nu* 'before, earlier' may be transcribed even if *nu* is uttered). For these and similar issues training ASR models in the documentary linguistics context, see Meelen et al. (2024).

In an effort to determine if better results could be easily attained, the authors addressed the first issue by standardizing spelling of vowel sequences throughout the transcriptions, included seven additional hours of recording, and used Whisper's medium baseline to generate a new model. Results from this model show an improved WER of 37.0% and CER of 14.4%.

Acknowledgments

We would like to thank people from the Nasal community who took part in this documentation project. We are also grateful to the National Research and

Innovation Agency (BRIN) for supporting this research in Indonesia as well as the Center for Culture and Language Studies at Atma Jaya Catholic University of Indonesia for sponsoring this research, especially the center's director Yanti. This material is based upon work supported by the National Science Foundation under Grant No. (1911641). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Karl Anderbeck and Herdian Aprilani. 2013. *The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra*. SIL International.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Andrea L. Berez-Kroeker, Shirley Gabber, and Aliya Slayton. 2023. Recent Advances in Technologies for Resource Creation and Mobilization in Language Documentation. *Annual Review of Linguistics*, 9(1):–330342568.

Blaine Billings and Bradley McDonnell. 2024. Sumatran. *Oceanic Linguistics*, 63(1):112–174.

Steven Bird. 2021. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

John W. Du Bois, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In Jane Anne Edwards and Martin D. Lampert, editors, *Talking Data: Transcription and Coding in Discourse Research*, pages 45–89. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

ELAN (version 6.8) [Computer software]. 2024. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen.

FieldWorks (version 9.1.25) [Computer software]. 2024. SIL Global, Dallas, TX.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan Van Esch, Scott Heath, František Kratochvíl, 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 *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 205–209. ISCA.

Bradley McDonnell. 2017. *Documentation of Nasal: An overlooked Malayo-Polynesian isolate of southwest Sumatra*. Endangered Languages Archive.

Bradley McDonnell, Blaine Billings, Jacob Hakim, Johan Safri, and Wawan Sahrozi. ongoing. *The languages of the Nasal speech community*. Collection BJM02 at catalog.paradisec.org.au [Open Access].

Marieke Meelen, Alexander O'neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of Nepal. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 83–93, St. Julians, Malta. Association for Computational Linguistics.

Brigitte Pakendorf, Nina Dobrushina, and Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism*, 25(4):835–859.

Nay San, Martijn Bartelds, Blaine Billings, Ella de Falco, Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley, Bradley McDonnell, and Dan Jurafsky. 2023. Leveraging supplementary text data to kickstart automatic speech recognition system development with limited transcriptions. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–6, Remote. Association for Computational Linguistics.

Eva Schultze-Berndt. 2006. Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Trends in Linguistics. Studies and Monographs [TiLSM]*, pages 213–252. Mouton de Gruyter, Berlin, New York.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14–22, Honolulu. Association for Computational Linguistics.

Whisper (version 20240930) [Computer software]. 2024. OpenAI, San Francisco, CA.