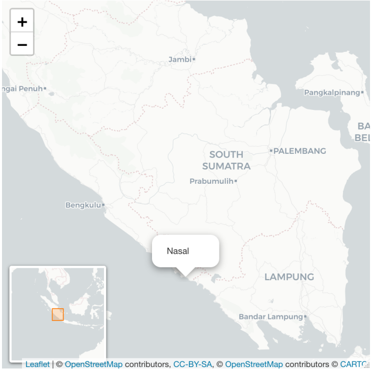
Connecting automated speech recognition to transcription practices

Blaine Billings[1], Bradley McDonnell[1], Johan Safri, Wawan Sahrozi

[1]University of Hawai‘i at Mānoa

March 4, 2025
ComputEL-8
Honolulu, HI, USA

Nasal



- Austronesian > Malayo-Polynesian > Sumatran language
- Spoken by ∼3,000 people in coastal southwestern Bengkulu province, Indonesia
- LEI: Endangered (Lee & van Way 2018)
- Not known to linguists until 2007 (Anderbeck & Aprilani 2013)
- Sustained intensive contact with neighboring Lampungic and Malayic languages

Nasal documentation

Documentation project began in 2017:

- Little existing prior documentation
  1. 3 written sources with limited lexical information
  2. No audio/video recordings with transcriptions
- Large corpus of conversation, elicitation, and narratives (McDonnell 2017, McDonnell et al. ongoing)
  1. ≈360hrs recording time
  2. ≈50hrs transcribed (mostly conversation)

Transcription workflow

Four primary steps in our transcription workflow:

1. Segmentation
2. Transcription
3. Discourse & Translation
4. Context

Time to transcribe audio is often upwards of forty times its length (Seifart et al. 2018)

The most time consuming step in transcription is #2.

ASR motivations

Sufficient input:

- Relatively large corpus of transcriptions (for training)
- Large corpus of untranscribed recordings (for testing)
- Availability of well-developed pre-trained models

Valuable output:

- Slow progression of manual transcription
- Reliance on transcripts for linguistic analysis
- Ongoing development of Nasal dictionary

Overview

Project goals:

1. Assess feasibility of implementing ASR in a typical documentary context
2. Determine if adequate results can be obtained
3. Knowing the limitations, determine if ASR could help speed up the transcription workflow

ASR

Data for training ASR model

Data used for model development:

- Transcriptions of 25 recordings
- Genres:
  - Everyday conversation (13)
  - Prosody elicitation (10)
  - Semantic domain elicitation (2)
- 49 unique speakers
  - 7 represented twice, 1 represented three times

Diversity in speakers and genres reflects intended use case of the ASR model

Data preparation

Data preparation:

- 25hrs recording time
- Timecodes in ELAN's XML used for clipping speech segments
- Final data
  1. 17.5hrs actual speech
  2. 160,000 words
  3. 66,500 annotations

80/20 split of training data and testing data

ASR model

Built by fine-tuning Whisper's small ASR model (Whisper (version 20240930) [Computer software] 2024)

Utilized pre-trained tokenizer and feature extractor from Indonesian (related language)

Run over 5,000 steps, evaluation of WER at every 500-step checkpoint

Best checkpoint used in generating test transcriptions

ASR results

```
wer : 65.40656872836156
wer : 57.97698137924582
wer : 55.721905118368106
wer : 53.242256947693456
wer : 48.40460372415084
wer : 49.10015283366083
wer : 48.058388696547205
wer : 46.28052774398802
wer : 44.71164342971211
wer : 43.938117962633726
```

Final checkpoint: 43.9% WER

- Significant improvement over previous model's 67.2% WER (San et al. 2023)

Tested against two segments not included for model development:

- Everyday conversation: 60.1% WER, 21.4% CER
- Dictionary: 54.1% WER, 20.4% CER
- WER and CER correspond exactly as expected with distribution of word length in Nasal

Transcription task

Comparison of manual vs. ASR-assisted transcription:

- Four 2.5min segments
  - Everyday conversation, dictionary recording
  - Empty annotations, ASR-generated annotations
- Block design transcription
- Screen-recorded for later analysis

Intended as impressionistic evaluation of ASR's effectiveness.

Transcription task

Results:

- All four transcription times faster correcting the ASR transcripts
  - Time improvement: 11.30%, 21.92%, 23.49%, 32.29%
  - As expected, correcting ASR-generated annotations was faster when done second
- Most often changes were single-letter or single-word edits

Feedback:

- Revising the automated transcription was preferred over transcribing from scratch
- Audio needed to be listened to fewer times in order for speech to be accurately determined

Discussion

Discussion

First thoughts on the model:

- Clearly, 43.9% WER does not seem strong
- Typical documentary data is indeed sufficient for developing an ASR model with usable WER/CER
- Time, cost, and software are not prohibitive for using ASR

So what is holding it back? What are the next steps?

Limitations

Data:

1. Spelling variation: *gawuh* vs. *gauh*
2. Shortenings: *jenu* vs. *nu*
3. Discourse marking: *m*, *uu*, *oo*
4. Signal bleeding and audio clarity

Model:

1. Model size: Started with Whisper small
2. Availability of computational resources
3. Limited training time

Future prospects

- Further study on viability of utilizing ASR to assist in transcription
- Additional development of ASR model: increasing size, normalizing transcripts, adding training data
- Determining the point of diminished gains in training

Addition: *Possibly attempt adding artificial data*

Using Whisper's medium model, (partially) standardized spelling, and seven additional hours of recording: 37% WER, 14.4% CER.

Summary

In response to our goals:

1. For time, cost, and data, ASR is feasible for the documentary context
2. Is 37% WER, 14.4% CER *adequate*? Yes!
3. ASR-generated transcript assisted in transcription

We would love to hear suggestions, feedback, or interest in collaboration

Acknowledgements

We would like to thank the following organizations and people.

1. The Nasal community from Tanjung Betuah, Gedung Menung, and Tanjung Baru, especially Johan Safri, Wawan Sahrozi, and Anton Supriyadi.
2. National Research and Innovation Agency (BRIN) for supporting this research in Indonesia
3. Center for Language and Culture Studies, Atma Jaya Catholic University of Indonesia, especially the center's director Yanti.
4. Google Cloud for the research credits needed to develop the ASR model

This material is based upon work supported by the National Science Foundation under grant BCS-1911641 to the University of Hawaiʻi at Mānoa. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Anderbeck, Karl & Herdian Aprilani. 2013. *The improbable language: survey report on the nasal language of bengkulu, sumatra*. Dallas: SIL International.

Lee, Nala H. & John R. van Way. 2018. The language endangerment index. In Lyle Campbell & Anna Belew (eds.), *Cataloguing the world's endangered languages*, 66–78. London: Routledge.

McDonnell, Bradley. 2017. *Documentation of Nasal: an overlooked Malayo-Polynesian isolate of southwest Sumatra*. Endangered Languages Archive. http://hdl.handle.net/2196/00-0000-0000-0010-798B-E.

McDonnell, Bradley, Blaine Billings, Jacob Hakim, Johan Safri & Wawan Sahrozi. ongoing. *The languages of the Nasal speech community*. Collection BJM02 at catalog.paradisec.org.au [Open Access]. https://dx.doi.org/10.26278/5f46870d43f29.

San, Nay, Martijn Bartelds, Blaine Billings, Ella de Falco, Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley, Bradley McDonnell & Dan Jurafsky. 2023. Leveraging supplementary text data to kick-start 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*, 1–6. Remote: Association for Computational Linguistics. (31 August, 2023).

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

Whisper (version 20240930) [Computer software]. 2024. *OpenAI, San Francisco, CA*. https://github.com/openai/whisper.