Automated speech recognition of Indonesian-English language lessons on YouTube using transfer learning

Zara Maxwell-Smith
The Australian National University
Zara.Maxwell-Smith@anu.edu.au

Ben Foley
The University of Queensland
b.foley@uq.edu.au

Abstract

Experiments to fine-tune large multilingual models with limited data from a specific domain or setting has potential to improve automatic speech recognition (ASR) outcomes. This paper reports on the use of the Elpis ASR pipeline to fine-tune two pre-trained base models, Wav2Vec2-XLSR-53 and Wav2Vec2-Large-XLSR-Indonesian, with various mixes of data from 3 YouTube channels teaching Indonesian with English as the language of instruction. We discuss our results inferring new lesson audio (22-46% word error rate) in the context of speeding data collection in diverse and specialised settings. This study is an example of how ASR can be used to accelerate natural language research, expanding ethically sourced data in low-resource settings.

1 Introduction

Accent, speaker-class characteristics, and the use of dialects are among many factors impacting automatic speech recognition (ASR) performance (Jurafsky and Martin, 2023). The dominance of 'high-resource' languages in natural language processing (NLP) and impact of market forces have produced strong outcomes for some applications of ASR when dialects, accented speech or particular speaker populations are excluded (Faisal et al., 2021; Koenecke et al., 2020; Bishop, 2022). However, many human speech scenarios, especially outside monolingual English contexts, require technologies more robust to language mixing and situated language usage — as well as performance measures of these technologies that prioritise the needs of users (Birhane et al., 2022).

This study worked with data from a non-standard context, that is, data from three YouTube channels teaching Indonesian with English as the language of instruction. It records the teaching practice of three teachers who: 1) explore a broad definition of 'Indonesian language', 2) demonstrate various linguistic behaviours associated with teaching (e.g., hyper-articulation and repetition), 3) would typically be described as 'accented' in either one or both languages, and 4) recorded their speech amidst background noise, adding music and sound effects. We hypothesised that repetition and simplifications in 'teacher-talk' intended to create comprehensible input for students (Krashen, 1981), and the use of transfer learning, could balance the many challenging characteristics in the data and allow ASR to create useful transcriptions for editing and analysis.

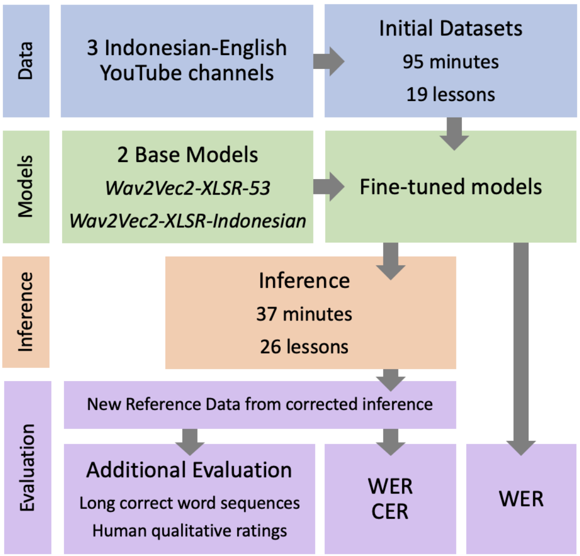
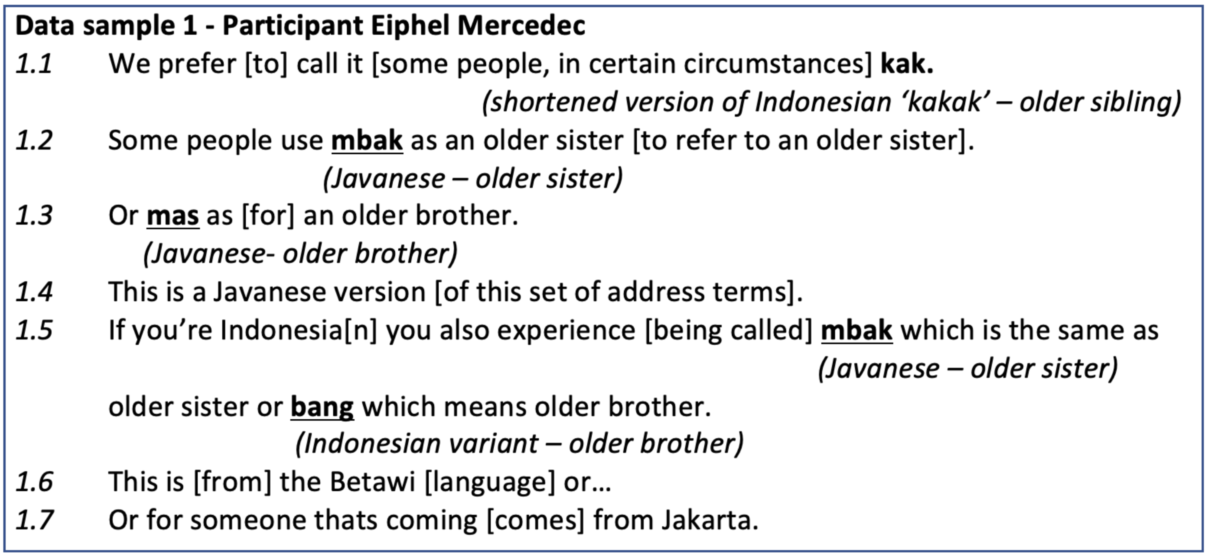


Figure 1: Study Design. See Section 2 for a detailed description of data used to fine-tune and evaluate models.

In recent years, transfer learning approaches have achieved state-of-the-art ASR performance on benchmark tasks with small quantities of data (Church et al., 2021). These approaches fine-tune a base model previously trained on a large dataset. Some pretrained models have been made publicly available, allowing more people to take advantage of their performance, and their advantages to be shared more equitably (Scao et al., 2022).

Proceedings of the Second Workshop on NLP Applications to Field Linguistics, pages 1–16
May 6, 2023 ©2023 Association for Computational Linguistics



Data sample 1 - Participant Eiphel Mercedec

1.1 We prefer [to] call it [some people, in certain circumstances] **kak.**
(shortened version of Indonesian 'kakak' – older sibling)

1.2 Some people use **mbak** as an older sister [to refer to an older sister].
(Javanese – older sister)

1.3 Or **mas** as [for] an older brother.
(Javanese- older brother)

1.4 This is a Javanese version [of this set of address terms].

1.5 If you're Indonesia[n] you also experience [being called] **mbak** which is the same as
(Javanese – older sister)
older sister or **bang** which means older brother.
(Indonesian variant – older brother)

1.6 This is [from] the Betawi [language] or…

1.7 Or for someone thats coming [comes] from Jakarta.

Figure 2: **Participant Sample 1 - Eiphel Mercedec.** This teacher grew up in Jakarta, speaking Mandarin and Cantonese at home, Mandarin and English in education settings, and Jakartan Indonesian in community settings. This study assessed her Indonesian accent to be Jakartan, and her English as mixing aspects of Hong Kong, Singaporean, American, and Australian accents. Here the participant demonstrated some of the linguistic stance-taking described by Abtahian et al. (2021), as she explained various address terms or substitutions for English 'you' appropriate in a market in Jakarta when buying an iPhone. [ ] – square brackets are additions from the transcriber to clarify meaning. ( ) – are translations and notes on linguistic and audio features. Underlined text is in a language other than standard Indonesian or English

Claims of state-of-the-art performance from fine-tuning a pretrained ASR model with as little as 10 minutes of labelled data (Baevski et al., 2020) often depend on large-vocabulary language models (San et al., 2023). For contexts where matching language models are not readily available, more realistic results are to be expected, such as in Coto-Solano et al. (2022) where median word error rate (WER) ranged from 18-66% for Cook Islands Maori. Even with language models, WERs remained high for low-resource languages: 32.91% for read speech in Bemba language in Sikasote and Anastasopoulos (2022) and 48% for Kurmanji Kurdish in Gupta and Boulianne (2022).

The aim of this study was to achieve a useful level of accuracy in machine transcription, creating drafts for human correction to expand the Online Indonesian Learning Dataset (OIL) (Maxwell-Smith, 2023). The study used the Elpis ASR toolkit to fine-tune models with a small set of human transcribed data. We trialled different base models, parameters, and mixes of fine-tuning data against various evaluation measures to better understand the performance of the tools and achieve this goal.

The rest of the paper is organised as follows: We begin with sociolinguistic and language-teaching commentary on the data, and then provide information about the experiment design, fine-tuning parameters, and standardised results. We discuss how different models performed on audio from new lessons and for different speakers. Finally, we reflect on technologies guided by direct and indirect user need, especially how evaluation measures inform decisions about usability of machine transcription for downstream tasks such as inference editing.

2 Methodology

The experiments used Elpis, a tool to aid linguists to apply sophisticated ASR tools and approaches such as Kaldi (Povey et al., 2011), Wav2Vec2 (Baevski et al., 2020). Elpis enabled us to work with ASR base models that are available on the Hugging Face Hub[1], a repository of public and private datasets and models suitable for machine learning. Models trained in Elpis were uploaded to the Hugging Face Hub, and then used for subsequent analysis.

An initial dataset of manually transcribed audio from YouTube videos totalled 1 hour and 35 min-

[1] See github.com/CoEDL/elpis & huggingface.co

utes (19 lessons). This initial dataset was used to fine-tune ASR base models. Inference texts from an additional seven lessons were used as a draft for human editors to expand the corpus to 2 hours and 13 minutes (26 lessons) of transcribed data.

We used a mixed methods approach to analyse our results, supplementing standardised ASR evaluation with qualitative commentary on transcription workflow and corpus analysis.

Table 1: Fine-Tuned Models: Epochs and WER

Model	Epochs	WER
fb_all	40	36.95
fb_NatInd	40	40.95
fb_JER_e60	60	**30.39**
ind_nlp_all	40	**36.89**
ind_nlp_NatInd	40	41.46
ind_nlp_JER_e60	60	**32.51**

Prefix *'fb_'* used base model Wav2Vec2-XLSR-53 and *'ind_nlp'*, Wav2Vec2-XLSR-Indonesian.

2.1 Data

YouTube channels specifying a purpose to teach Indonesian were identified using keyword searches and recommendations from professional teaching networks. The listed email on these YouTube channels was contacted, progressing from channels with more to less content, until three participants were recruited (see Table 4, Appendix A). These three channel owners confirmed their ownership of materials, and their explicit consent was obtained for their materials to be used for ASR development, language and teaching analysis, as well as sharing as both audio and audio-visual files for future research (see our Ethics Statement).

To ensure our system would be robust to future data from this setting we did not exclude data with characteristics known to be challenging for ASR. These characteristics include background noise, accented speech, task specific intonation/articulation, and frequent language mixing. By using so-called 'noisy' data, our study has provided realistic insight into the performance of ASR for the real-world task of converting teacher speech from YouTube into searchable text.

Anecdotally, we observed a high rate of repetition of sounds, words and phrases in the data. We hypothesised that this would persist throughout the data as teachers aim to present 'learnable' language. That is, the data would be influenced by a common intention among teachers to present 'comprehensible input' to students (Krashen, 1981).

The language background of participants was gained via interviews, with all participants having spoken languages other than Indonesian as children. Participants reported varied language backgrounds and daily use of Indonesian at the time they filmed their videos. In their interviews, the teachers self-described their projected YouTube identity and indicated that they varied their content, tone, and language choice from video to video. Their projected identities varied and were described as 'friend', 'teacher/educator' and 'entertainer'. Participant teaching experience ranged from many years of paid work teaching Indonesian, to no experience as a professional teacher.

Almost all videos contain a mix of languages, with some dominated by Indonesian or English. Some videos explicitly focused on variation in Indonesian or words from other languages which are commonly mixed into Indonesian by speakers. Table 5 (Appendix D) contains notes on the main languages in each file, as well as a subjective comment on whether language mixing tended towards inter-utterance or intra-utterance mixing.

Noise levels were variable according to the channel and each individual video. Some videos were recorded in quiet spaces with minimal reverberation, while others have frequent high volume loudspeaker announcements from local Musholla, added sound effects, muffled voices from other speakers, and road noise. Typical noise associated with each channel is listed in Table 4 (Appendix A).

To illustrate some of the speech phenomena and other characteristics in this data we have produced a excerpt with relevant annotation (Figure 2). Examples from the two other participants are included in Appendices B and C.

2.2 Transcription

The initial transcription of files was completed by Author 1, who is an Indonesian-English bilingual, teacher, and linguist. Reference texts for each audio file were transcribed by the same transcriber using inference texts from the ASR experiments as drafts to speed the process. Reference files were checked by a second expert transcriber (Indonesian-English bilingual and linguist - see Ethics Statement below) to verify the reference transcription quality.

Transcribers erred towards recording words found in both languages with the orthography of Indonesian[2]. Non-standard forms (those not found in KBBI [3]) were transcribed as an approximation of sounds. For example '*lapan*', which is derived from '*delapan*' with first syllable deletion, and '*udah*', a Jakartan variant of '*sudah*'. Where possible, existing literature documenting variants was used to inform spelling (e.g. '*ngapain*' in Sneddon (2006)).

2.3 Experiment

The experiments consisted of fine-tuning multiple pre-trained ASR transformer models using combinations of datasets listed in Appendix D. The data in Appendix D was YouTube data manually transcribed from scratch and used to fine-tune multiple models with different characteristics. Machine transcriptions, or inference, were then used as a draft for human editing. These corrected inference files were checked by another transcriber and then considered 'gold standard' reference files, adding further data to the corpus. The original machine inference was then compared with the 'gold standard' reference files to measure WER and other performance markers. Standard ASR word error rate metrics were calculated, along with other metrics. A qualitative review of inference texts was undertaken as described in Section 3.

Fine-tuning and inference files. To enable us to select a balance of characteristics of audio and speech in our data, Author 1 listened to and coded all files from each YouTube channel for a range of characteristics (see Section 2.1 for commentary). This enabled us to choose files which roughly represented the spread of content (the topics taught and the focus of each lesson on language learning skills such as vocabulary, grammar explanation, or the teacher modelling authentic speech). We also sought to include files containing a spread of background noise and sound effects typical of channels.

When selecting files we considered speech and language behaviours such as: 1) which language dominated in a given lesson, 2) whether code-switching or translanguaging occurred between or within utterances (inter-utterance or intra-utterance), and 3) the frequency and degree of hyperarticulation by each teacher. This coding was carried out on untranscribed audio, and represented a first pass impression of audio characteristics. We sought to balance these characteristics, but note these are highly complex speech behaviours and difficult to assess even with a well-trained team of transcribers. In Maxwell-Smith et al. (2020), we discussed the complexity of measuring these behaviours in similar data at length.

2.3.1 Fine-tuning and Evaluation

Multiple models were fine-tuned to compare the effects of different combinations of data across different base models, using files manually transcribed by Author 1 (see Training Data in Appendix D). Models were evaluated using standard ASR metrics of WER from Elpis-internal train/validation/test splits for each model (see Table 1). Evaluation of inference files (see Inference Data in Appendix D) was enhanced by calculating the number of common word sequences of different lengths, and performing qualitative user rating of inference texts.

Base Models. Elpis was used to fine-tune two pre-trained base models, using combinations of labelled data for fine-tuning. One base model was Facebook's Wav2Vec2-XLSR-53 multilingual model (Conneau et al., 2021) which has been pre-trained on 56K hours of speech from 53 languages. The other base model was an Indonesian ASR model released by Indonesian NLP. Indonesian NLP used a subset of Indonesian-labelled speech from the CommonVoice dataset to fine-tune Wav2Vec2-XLSR-53, releasing it as a general Indonesian language ASR model, Wav2Vec2-XLSR-Indonesian, with 14.29% WER reported.

The base model is indicated in the first section of the model name. Models beginning with *fb_* indicates Facebook's multilingual model, and *ind_nlp_* indicates Indonesian NLP's model.

Parameters. Audio files were prepared in 16kHz, 16bit, mono, WAV format; the internal specifications used by Elpis. Transcription files were created in ELAN format, sharing a common tier name for ease of text selection in Elpis.

[2] Generally these were words loaned from English or other European languages into Indonesian.

[3] *Kamus Besar Bahasa Indonesia*, The Big Indonesian Dictionary, is produced by The Agency for Language Development and Cultivation of the Indonesian Ministry of Education, Culture, Research, and Technology.

Table 2: Inference Results for *ind_nlp*, *fb_all*, *ind_nlp_all*

Model		**ind_nlp**			**fb_all**				**ind_nlp_all**			
File	**Token**	**L:6**	**WER**	**CER**	**R**	**L:6**	**WER**	**CER**	**R**	**L:6**	**WER**	**CER**
EIP_010	35	1	80.00	46.07	e	1	28.57	**6.28**	e	**3**	**22.86**	6.81
EIP_011	598	0	79.93	42.03	r	**15**	**43.65**	**13.78**	e	11	46.82	15.05
EIP_013	629	1	83.47	42.86	e	**20**	44.36	**14.60**	e	**20**	**43.40**	15.97
GUN_004_01	654	1	73.70	38.07	e	**26**	**28.75**	**10.05**	e	**26**	31.65	11.20
GUN_004_10	847	6	83.47	46.74	e	**29**	41.20	13.64	e	27	**35.42**	**13.35**
JER_019	333	1	87.09	51.54	e	**14**	**33.33**	**10.98**	e	7	37.84	15.59
JER_079	992	0	94.05	54.01	e	**47**	**36.29**	**13.33**	r	34	43.45	16.32

L:6 — The number of correct word sequences of length 6 and above.
R — A human transcriber rating for the perceived usefulness of the inference as a basis for editing. Inferences rated 'e' would be edited, 'r' would be used as a reference while transcribing from scratch.
Bold scores — Best or equal best score. Table 6 in Appendix E includes results for all models in Table 1.

Preliminary rounds of fine-tuning with subsets of the data were used to identify suitable learning rates, ideal number of epochs, and batch size. Reductions in WER and loss for training conducted over 40 epochs were negligible. A trade-off was made to limit the number of epochs to reduce training time, possibly at the expense of very minor improvements in WER. A range of learning rates were used in preliminary rounds, with $1e{-4}$ determined to be the most suitable for the final models.

Verification. After being trained, the fine-tuned Elpis models were uploaded to the Hugging Face Hub and used in Google Colab[4] to obtain inferences for untranscribed audio. Using Colab was a workaround for restrictions on the length of inference audio which Elpis would process at the time of the experiment. A custom Python script was used in Colab to load Elpis-trained models from Hugging Face and run inferencing with Hugging Face pipeline tools. Colab was later used to run evaluation scripts to calculate word and character error rates, and to find the longest correct word sequences for these inferences.

Evaluation. WER values up to 30% were reported by Gaur et al. (2016) as being useful as a 'canvas' or starting point for correction. However, due to the intricacies of manually editing transcription files in ELAN, an inference with WER within this threshold might still be cumbersome to correct, while inference outside this threshold might actually have extended sections of correct transcription. From Author 1's personal experience, editing text with frequently alternating correct/incorrect sequences was known to be more labour-intensive than editing text with long sequences of correct words, indicating that the standard WER metric of performance does not necessarily correlate with user experience.

Before reference transcriptions had been created, Author 1 made a qualitative review of inference from models that had low WER. Inference output was reviewed and rated according to the estimated frequency of extended sequences of correct words, as well as the position of necessary edits and the number of keystrokes required to correct the text in ELAN. Based on this assessment of the anticipated manipulation process, a rating for each inference text was made from a five-point scale (useless, glance, refer, edit, wow).

3 Results

The speaker specific models *fb_JER_e60* and *nlp_all_JER_e60* achieved the lowest WER from elpis-internal train/test splits (30.39% and 32.51% respectively). Train/test evaluation is compared in Table 1. The initial results from *fb_JER_e60* and *nlp_all_JER_e60* may have been due to the greater number of epochs. However, the performance of the models when trained was not reflected in their application to new lessons from the same speaker (WERs of 38.44% and 44.36% from *fb_JER_e60*).

[4] https://colab.research.google.com

The next best training evaluation scores were from models fine-tuned with all our data: *ind_nlp_all* and *fb_all*[5]. Experimental models fine-tuned with a subset of data[6] from teachers who were long-term residents of Indonesia (models with suffix *_NatInd*) had higher WER.

While Indonesian NLP reports WER of 14.29% for the base model *ind_nlp*, it did not score well on inference of our multilingual audio (Table 2 sets out evaluation metrics for inference of new lesson audio). No WER below 70% was achieved using *ind_nlp* and very few long correct sequences were produced. Using Indonesian NLP's model, which is fine-tuned with monolingual Indonesian data, was not suitable for our data.

Our own fine-tuned models were a dramatic improvement on these results. The best WERs on inference files ranged from 22.86% to 43.65%. No single model consistently achieved the best WER, CER or correct sequence of six or more on inference of new audio (Table 2). The *fb_all* model achieved a greater number of better scores, and on closer inspection of correct sequence counts would appear to have produced inference which is more easily editable (Table 3). Merged word errors, such as *'ribuseratus'* rather than *'ribu seratus'* (thousand one hundred), were prevalent in inference from all models.

Table 3: Correct word sequences in inference from *fb_all* and *ind_nlp_all* models

	fb_all				**ind_nlp_all**			
File	**L:4**	**L:5**	**L:6**	**L:7**	**L:4**	**L:5**	**L:6**	**L:7**
EIP_010	2	1	1	1	3	3	3	2
EIP_011	31	24	15	12	28	19	11	10
EIP_013	39	30	20	17	31	28	20	17
GUN_004_01	42	34	26	22	39	32	26	16
GUN_004_10	51	41	29	26	52	39	27	19
JER_019	18	17	14	10	18	14	7	6
JER_079	85	68	47	36	77	52	34	26

L:x — The number of correct word sequences of length x is marked with L:x.

Two files from participant Jeremy Snyder (JER), received best scores with the Facebook base model fine-tuned on all data (*fb_all*). Data from other participants achieved better scores across both the *fb_all* and *ind_nlp_all* models. Jeremy was the only participant with English as a first language. This weighting towards the Facebook base model may be related to Jeremy's spoken English more closely matching English in the Facebook base model data.

Qualitative human rating of inference indicated inferences from *fb_all* and *ind_nlp_all* were suitable for editing (see R in Table 2). Verifying this finding with timed transcription experiments to ascertain the degree of acceleration was beyond the scope of this project. However, the suitability of inference files for editing was confirmed by Author 1 as she used them to expand the dataset. The process of editing inference also led to interesting reflections on the data itself, as discussed below.

4 Discussion

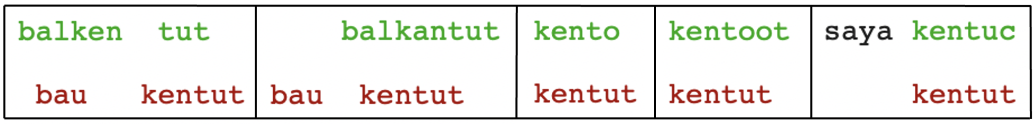
Principal findings. This paper makes a unique contribution in demonstrating the viability of using ASR for an explicit and executed purpose. Machine transcription was successfully edited to increase the size of a noisy, mixed-language, Indonesian-English, YouTube language teaching dataset with three speakers. The expanded dataset will improve analysis of teacher speech by a teacher-researcher. It also provides ethically sourced and openly released materials to engineer and enhance bespoke NLP solutions in a setting that is currently low-resource.

While machine transcription accelerated the transcription process, the process of fine-tuning base models and preparing data required an upfront investment which was not compensated for by this acceleration over seven inference files. We hope that our upfront investment can be useful to others via our models and data on Hugging Face.

The process of editing machine transcriptions revealed workflow and evaluation needs. It also impacted human transcriber interpretation of the data, provoking discussion of how multilingual, accented, language teaching plays out. Meanwhile, so-called 'errors' in inference were less concerning than they would be in other fields where accuracy is of paramount importance (such as in medical applications of ASR (Joseph et al., 2020; Miner et al., 2020).

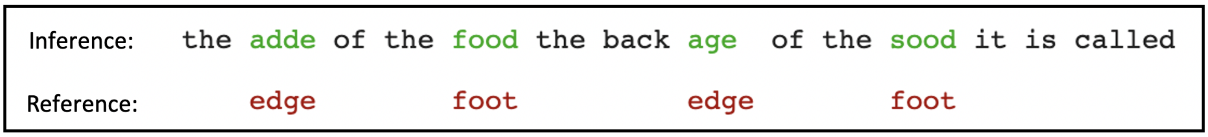
[5] (Maxwell-Smith and Foley, 2023b) & (Maxwell-Smith and Foley, 2023a)

[6] See data marked with [a] in Appendix D



balken tut	balkantut	kento	kentoot	saya kentuc
bau kentut	bau kentut	kentut	kentut	kentut

Figure 3: Incorrect inference (Green) and reference (Red) of a lesson using fart humour to teach grammar. Reference transcription is: *bau kentut* (fart stench), *bau kentut* (fart stench), *kentut* (fart), *kentut* (fart), *saya kentut* (I farted)



Inference:	the	adde	of	the	food	the	back	age	of	the	sood	it	is	called
Reference:		edge			foot			edge			foot			

Figure 4: Insights into accented speech via error correction of inference.

Correct sequences: length and location. The placement of correct sequences of inference influenced the usability of an inference as an editable draft. Specifically, longer correct sequences and those that were left-aligned reduced the time spent editing, an impact not measured by WER. Similarly, word final spelling errors were less disruptive to the editing process as they required less keystrokes to correct. As an example, for reference *'satu ribu'*, the inference *'a satu ribu'* is more disruptive than *'satu ribua a'*. This is despite having lower WER and CER.

'Out-of-domain' lexicon. In a lesson using humorous discussion of farts to teach grammar, Jeremy Snyder produces the words *'bau'* (stench) and *'kentut'* (fart) repeatedly. These are consistently inferred incorrectly (see Figure 3), despite minimal hyperarticulation and background noise, and fairly clear articulation. This is likely due to their absence from training data — they belong to language rarely used in public settings though they are not uncommon in everyday life[7]. Reflecting on this limitation of machine transcription highlights the domain of use for certain language and how students may encounter, or not encounter, certain words in their learning journey.

Accented speech. The reflection of speech behaviours in machine transcription also stimulated reflection on teacher pronunciation. The use of context and language knowledge in understanding and interpreting teacher speech is highlighted in the following examples.

In a lesson from Gunawan Tambunsaribu (GUN) the word final 't' in 'foot' was converted to a 'd' in the machine inference (see Figure 4). Human analysis found the production of 'oo' (in 'foot') by the participant matched with the common grapheme to phoneme pair in words like 'too' and 'roo'. However, 'foot', confusingly given it's spelling, is pronounced like 'put'. The inference highlighted the transfer of Indonesian vowel production and possibly a speech error resulting from irregularities in orthographic conventions in English.

Similar to a language model (LM), a human transcriber editing the inference in Figure 4 would step through each word, finding 'adde' to be a non-word. Presuming correction to 'edge' was substituted, the sentence 'The edge of the food' would be judged improbable and corrected despite the vowel production described in the previous paragraph. Further, a human and a LM would preference 'foot' over 'food' as the preceding data indicates body parts are likely, being the topic of the lesson.

In another inference, the transcription of 'tv' as 'tipi' matched the participant's production of the word. The inference reflected a common characteristic displayed by Indonesian speakers in which fricatives and plosives[8] are not always differentiated (Nurhayati, 2020).

'Non-words'. The machine transcription of 'non-words', or words invented by the teacher to illustrate a point, also spurred discussion and reflection. For example, data, again from Gunawan Tambunsaribu, in which he purposefully produced 'yuk' incorrectly with the glottal stop aspirated was inferred as 'youk'. This estimated orthography for a non-existent word was found stimulating for transcribers rather than harmful.

[7] This is not a comment on the authors' own level of flatulence, though it is relevant to the topic of domain shift in computational linguistics (Paraskevopoulos et al., 2023).

[8] In this example (/v/) and (/p/) respectively.

Editing inference 'errors' highlighted patterns in teachers' speech and illustrated incidental learning encountered by students. All participants demonstrated non-standard pronunciation of Indonesian and English. The examples above offer evidence for the role of intermediary targets of pronunciation in language teaching and techniques in pronunciation instruction; a lively research area in second language acquisition (Lee et al., 2014).

Language models. Often the addition of a LM will be used to improve ASR and other NLP. However, in this application of ASR, introducing a LM would be unlikely to assist as code-switching behaviours, non-standard grammar and accents, as well as situated language from the language learning setting has largely been excluded from language technologies (Scao et al., 2022). In other words, LMs built from data similar to ours are not yet available.

In our study, human transcribers took on the role of LM correction. However, this placed signficant demands on transcribers to be multilingual and knowledgeable in the language learning setting. These demands make transcription and error correction of this data a true bottleneck. Optimistically perhaps, we see this work as potentially enriching for teachers and their reflective teaching practice. It can bring attention to interlanguage and movement between native speaker modelling and intermediary productions of sounds and language structures.

Future work. ASR systems fine-tuned with very small quantities of data often rely on LMs trained with large amounts of text data (San et al., 2023). These systems typically use a multilingual base model that has been fine-tuned to a monolingual language, with a monolingual LM[9]. In this setting, further work to develop a complex multilingual LM could improve results with a pre-trained multilingual model fine-tuned with multilingual data.

A major challenge in the development of multilingual LMs for contexts such as this is the varying inter-utterance and intra-utterance code-switching that occurs in teacher speech (Maxwell-Smith et al., 2020). These switches are likely to disrupt potential identification of language for an n-gram LM. An n-gram sequence identified as English may in fact erroneously negate a correctly identified Indonesian word in the sequence.

Further work to investigate initial diarisation/language identification may be a fruitful approach to handling this language complexity. Such an approach was taken in Szalay et al. (2022) to assist with mixed data from adult and child speakers. In this setting, multiple mono-lingual LMs used on identified languages and then compiled could be helpful (Shen, 2022). However, with the degree of hyperarticulation and accent evident in this study's audio, reliable language identification itself is likely to be difficult.

The prevalence of merged word errors identified in inference texts (e.g. 'reduplicationand' rather than 'reduplication and'), would be resolved in a monolingual system through the use of a LM. Given a LM may not work well on data like this, future work for complex language systems could investigate the benefits of a rudimentary splitting step based on matches with combined bigrams or trigrams from a multilingual vocabulary list.

Audio content analysis indicates that errors concentrated at the beginning and ends of files were associated with background music. Given the consistent poor inference text in these sections, better performance would be likely by excluding these sections of the files.

5 Conclusion

Our findings offer a reality check of ASR performance with 'difficult' data, including newer techniques of transfer learning. Our results clearly indicated that publicly available models for Indonesian are not suitable for processing holistic language teaching data. Inference from a model fine-tuned on a small dataset of complex language was much more useful. The WER remained high, however, rather than discarding results based on the industry-standard/internal expectations, we persisted and edited inference text to expand our dataset. The resulting insights into user workflows encourage investigation of task-specific evaluation measures. Meanwhile, insights into data characteristics that were highlighted by editing the inference texts go some way to counterbalancing the time spent in interactions with ASR output by language-teaching professionals. Our ethically sourced dataset[10] and best models[11] are available on Hugging Face.

[9] https://discuss.huggingface.co/t/how-to-create-wav2vec2-with-language-model/12703

[10] Online Indonesian Learning (OIL) Dataset

[11] OIL ASR models

Limitations

This study represented complex human language with simple orthography, including language mixing, hyperarticulation and variation. Further linguistic annotation would enrich the dataset and enable deeper insights into language teaching behaviours. For example, phonetic transcription would help to differentiate words that occur in both languages and allow for exploration and comparison of accented speech between participants.

The potential benefits of using a multilingual LM to improve ASR results were not studied due to the language complexities of the dataset. Further work is required to: 1) develop complex multilingual LMs matching the language and, 2) conduct subsequent studies on the efficacy of a complex LM in the ASR system.

Ethics Statement

The audio (and visual) data from the three YouTube channels was transferred by participants after discussing the project and possible impacts of sharing their data (Ethics Approval No. 2017/889 of the *Australian National University* Human Research Committee, *Speech Recognition; Building Datasets from Indonesian Language Classrooms and Resources* protocol). Files were screened for intelligible speech from people other than the participant and those containing such data were removed from the dataset. The non-author transcriber referred to in Section 2 completed the transcription as part of an exchange of editing and proof reading. Our appreciation for his contribution to the project is expressed in our Acknowledgements.

With a view to advancing the language technologies available for Indonesian, and especially Indonesian and English bilingual data, and to support research into Indonesian language teaching, the dataset has been made available for other researchers to further develop these tools and complete their own analysis. Our study documented one approach to developing NLP in understudied language situations, contributing to realistic expectations of NLP in settings outside monolingual English settings most supported by the investment of business interests.

The study and release of data does embody some risks for participants as data stored in an open repository could be downloaded to create other derivative works not aligned with this research (Kale 2019). As videos contain the professional teaching practice of some participants, and the 'YouTuber' persona of others, there is a risk of reputational damage. This risk and that of derivative works was made clear in the participant information sheet and storage in an open repository was subject to explicit consent on the consent form. To further reduce risk, videos with individuals not explicitly involved in the making of the video (bystanders) were excluded from the dataset. We believe the risk of misappropriation of content from YouTube was already significant for participants as their work could be copied relatively easily from YouTube; their involvement in this project increased the risk of misappropriation only slightly.

Acknowledgements

We would like to acknowledge the support and guidance of our supervisors Janet Wiles, Professor Hanna Suominen, Dr Michelle Kohler, and Assoc Professor Danielle Barth, and colleagues Nay San and Daan van Esch. Our special thanks to our research participants, Gunawan Tambunsaribu, Jeremy Snyder, and Eiphel Mercedec, and to Yustinus Ghanggo Ate for contributing his expertise and time in our editing exchange.

References

Maya Ravindranath Abtahian, Abigail C. Cohn, Dwi Noverini Djenar, and Rachel C. Vogel. 2021. Jakarta indonesian first-person singular pronouns: Form, function and variation. *Asia-Pacific Language Variation*, 7(2):185–214.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Judith Bishop. 2022. *Linguistic Diversity in AI: A Provocation*. ARC Centre of Excellence for the Dynamics of Language, Panel: New connections for language and technology CoEDL End-of-Centre Event, Friday, 30 September 2022.

Kenneth Ward Church, Zeyu Chen, and Yanjun Ma. 2021. Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, 27(6):763–778.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*. ISCA.

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.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, W4A '16, New York, NY, USA. Association for Computing Machinery.

Vishwa Gupta and Gilles Boulianne. 2022. Progress in multilingual speech recognition for low resource languages Kurmanji Kurdish, Cree and inuktut. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6420–6428, Marseille, France. European Language Resources Association.

Joseph Joseph, Zena EH Moore, Declan Patton, Tom O'Connor, and Linda Elizabeth Nugent. 2020. The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: A systematic review. *Journal of Clinical Nursing*, 29(13-14):2125–2137.

Daniel Jurafsky and James H Martin. 2023. Automatic Speech Recognition and Text-to-Speech. In *Speech and Language Processing*.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Stephen Krashen. 1981. *Second Language Acquisition and Second Language Learning*. Pergamon Press Inc.

Junkyu Lee, Juhyun Jang, and Luke Plonsky. 2014. The Effectiveness of Second Language Pronunciation Instruction: A Meta-Analysis. *Applied Linguistics*, 36(3):345–366.

Zara Maxwell-Smith. 2023. Online Indonesian Learning dataset (OIL) (revision b2a39e5).

Zara Maxwell-Smith and Ben Foley. 2023a. ZMaxwell-Smith/OIL_YT_fb_all Automatic Speech Recognition (ASR) model (revision 1fb3a19).

Zara Maxwell-Smith and Ben Foley. 2023b. ZMaxwell-Smith/OIL_YT_ind_nlp_all Automatic Speech Recognition (ASR) model (revision 1a14ec0).

Zara Maxwell-Smith, Simón González Ochoa, Ben Foley, and Hanna Suominen. 2020. Applications of natural language processing in bilingual language teaching: an indonesian-english case study. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–134.

Adam S Miner, Albert Haque, Jason A Fries, Scott L Fleming, Denise E Wilfley, G Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A Arnow, W Stewart Agras, et al. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digital Medicine*, 3(1):1–8.

Dwi Nurhayati. 2020. Plosive and fricative sounds produced by efl students using online media: A perspective on learning english phonology. In *Proceedings of the 1st International Conference on Folklore, Language, Education and Exhibition (ICOFLEX 2019)*.

Georgios Paraskevopoulos, Theodoros Kouzelis, Georgios Rouvalis, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. 2023. Sample-efficient unsupervised domain adaptation of speech recognition systems a case study for Modern Greek.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

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 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*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Gaofei Shen. 2022. Does where words come from matter? Leveraging self-supervised models for multilingual ASR and LID. Master's thesis, Center for Information Technology of the University of Groningen, Campus Fryslan, August.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

James N. Sneddon. 2006. *Colloquial Jakartan Indonesian*, volume 581. Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, Canberra.

Tünde Szalay, Louise Ratko, Mostafa Shahin, Tharmakulasingam Sirojan, Kirrie Ballard, Felicity Cox, and Beena Ahmed. 2022. A semi-automatic workflow for orthographic transcription of a novel speech corpus: A case study of AusKidTalk. In *Proceedings of the Eighteenth Australasian International Conference on Speech Science and Technology*. Australasian Speech Science and Technology Association.

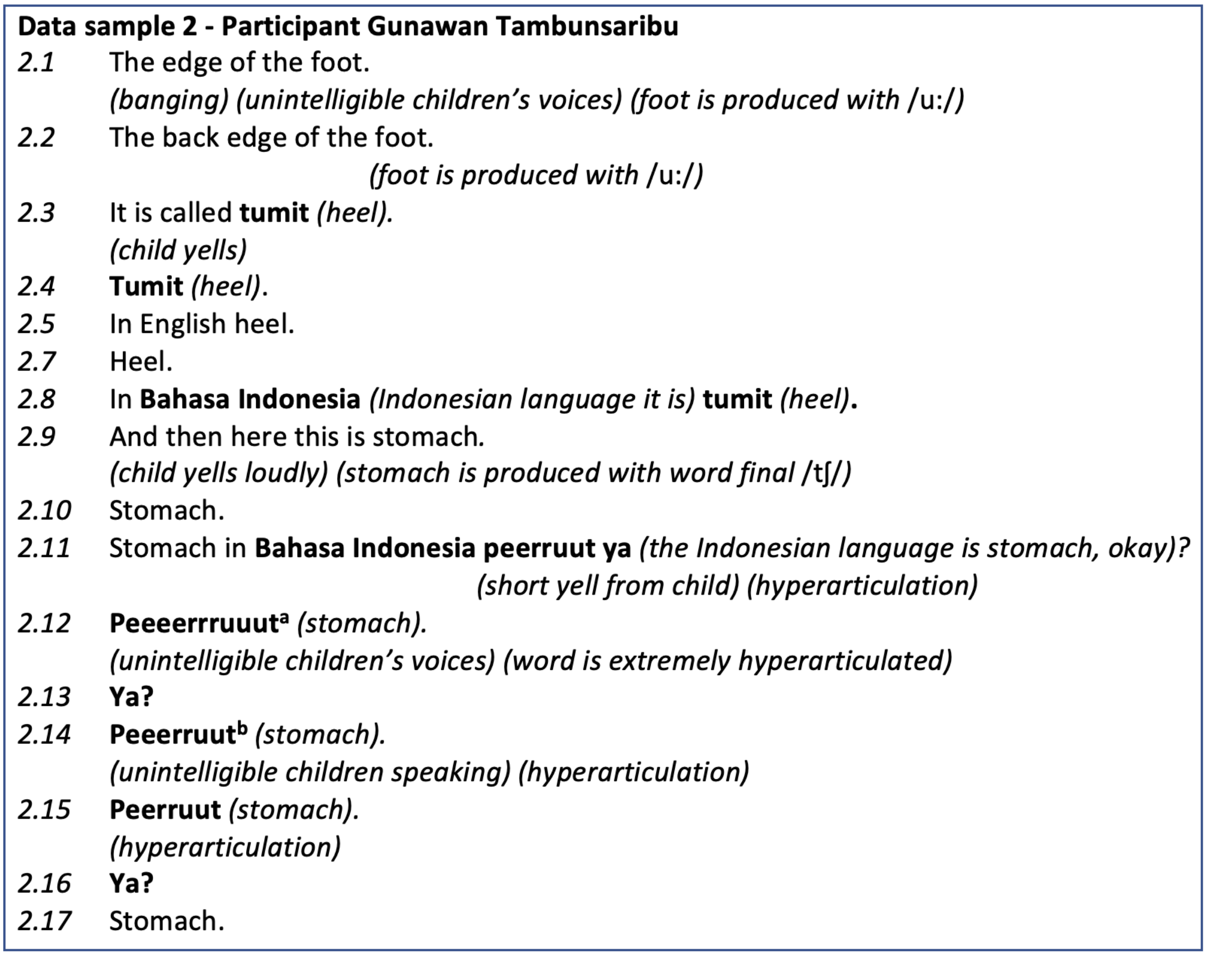
Appendix A Speaker/channel characteristics

Table 4: Speaker/channel characteristics

Eiphel Mercedec	**Gunawan Tambunsaribu**	**Jeremy Snyder**
5-Minute Indonesian	Indonesian Language for Beginners ENG-INA	Dua Budaya
Languages used at home		
Mandarin, Cantonese, Indonesian, English	Batak Simalungun, other varieties of Batak	English, Indonesian
Language of formal education		
Mandarin, English	Indonesian, English	English, Indonesian
Use of Indonesian		
Community interactions	Family, work, community	Teaching, family interactions
Residency		
Indonesia, Jakarta	Indonesia, Jakarta	Australia, Perth
Typical ‘noise’ in audio		
Clear, music, sound effects	Background noise (call to prayer, other speakers, street noise)	Clear, some music
Duration		
34 minutes	7 hours 51 minutes	2 hours 53 minutes
Number of files		
13	22	63

Characteristics: This table is characteristics drawn from participant descriptions of their lives at the time of video/channel creation.

Appendix B Speaker sample 2



Data sample 2 - Participant Gunawan Tambunsaribu

2.1 The edge of the foot.
(banging) (unintelligible children's voices) (foot is produced with /u:/*)*

2.2 The back edge of the foot.
(foot is produced with /u:/*)*

2.3 It is called **tumit** *(heel).*
(child yells)

2.4 **Tumit** *(heel)*.

2.5 In English heel.

2.7 Heel.

2.8 In **Bahasa Indonesia** *(Indonesian language it is)* **tumit** *(heel)***.**

2.9 And then here this is stomach*.*
(child yells loudly) (stomach is produced with word final /tʃ/*)*

2.10 Stomach.

2.11 Stomach in **Bahasa Indonesia peerruut ya** *(the Indonesian language is stomach, okay)?*
(short yell from child) (hyperarticulation)

2.12 **Peeeerrruuut**[a] *(stomach).*
(unintelligible children's voices) (word is extremely hyperarticulated)

2.13 **Ya?**

2.14 **Peeerruut**[b] *(stomach).*
(unintelligible children speaking) (hyperarticulation)

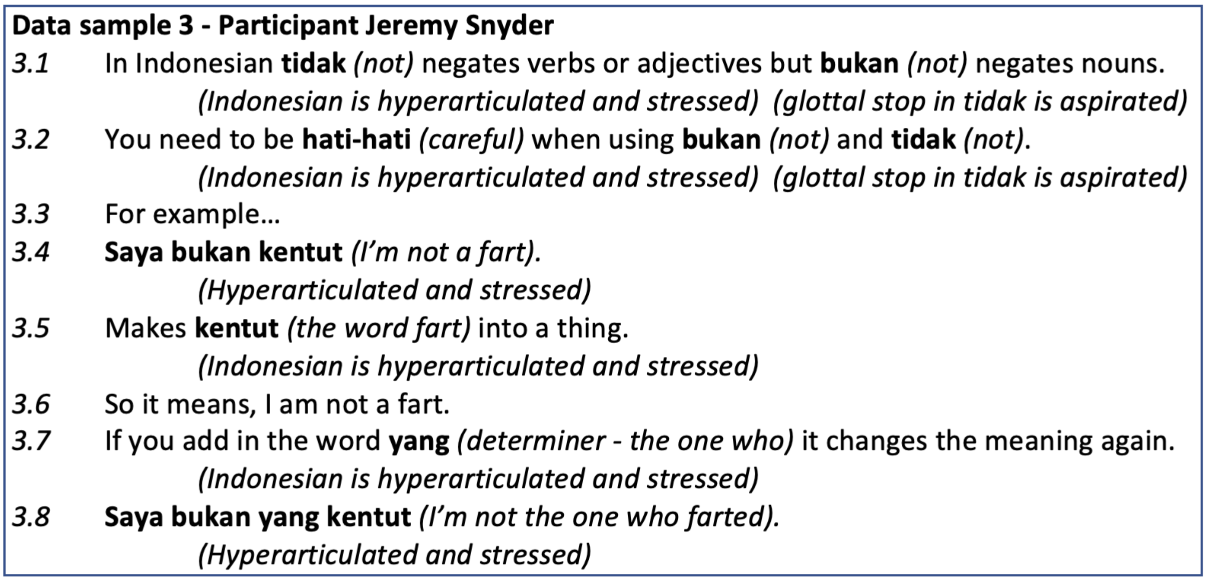
2.15 **Peerruut** *(stomach).*
(hyperarticulation)

2.16 **Ya?**

2.17 Stomach.

Figure 5: **Participant Sample 2 - Gunawan Tambunsaribu.** This teacher grew up speaking Batak Simalungun, completed his education in Indonesian and English and has lived in Jakarta for more than 15 years, speaking Indonesian and Betawi. His Indonesian accent is Jakartan, while his English could be described as having an international and Indonesian accent. Here he produces hyperarticulated speech to highlight the sounds of new vocabulary. The duration of the most hyperarticulated instance of 'perut" is perut[a], (1.4 seconds). While still hyperarticulated, perut[b] is much shorter (0.84 seconds). The token 'stomach' is transcribed orthographically here but varies, with the first instance produced with a 'tch' sound, as in 'latch' which is then corrected by the participant. The audio includes background noise from children playing and unintelligible childrens' speech. ( ) – are translations and notes on linguistic and audio features. / / - provide phonetic information.

Appendix C Speaker sample 3



Data sample 3 - Participant Jeremy Snyder

3.1 In Indonesian **tidak** *(not)* negates verbs or adjectives but **bukan** *(not)* negates nouns.
(Indonesian is hyperarticulated and stressed) (glottal stop in tidak is aspirated)

3.2 You need to be **hati-hati** *(careful)* when using **bukan** *(not)* and **tidak** *(not)*.
(Indonesian is hyperarticulated and stressed) (glottal stop in tidak is aspirated)

3.3 For example…

3.4 **Saya bukan kentut** *(I'm not a fart).*
(Hyperarticulated and stressed)

3.5 Makes **kentut** *(the word fart)* into a thing.
(Indonesian is hyperarticulated and stressed)

3.6 So it means, I am not a fart.

3.7 If you add in the word **yang** *(determiner - the one who)* it changes the meaning again.
(Indonesian is hyperarticulated and stressed)

3.8 **Saya bukan yang kentut** *(I'm not the one who farted).*
(Hyperarticulated and stressed)

Figure 6: **Participant Sample 3 - Jeremy Snyder.**This teacher grew up speaking English, completed his education in English and Indonesian, and has lived in Australia and Indonesia, speaking English and Indonesian. His Indonesian has an Australian accent, as does his English. In this example he produces hyperarticulated speech to highlight the sounds of target language for learners and for emphasis/comedic effect.( ) – are translations and notes on linguistic and audio features.

Appendix D Datasets

Table 5: Datasets

File name	Duration	Language	Codeswitch	Music	Audio quality	Hyper
TRAINING DATA						
Eiphel						
EIP_002 [a]	0:01:46	Mix	Intra	X	Moderate	High
EIP_003 [a]	0:02:09	Mix	Intra	X	Moderate	High
EIP_006 [a]	0:02:45	Mix	Intra	X	Moderate	Med
EIP_007 [a]	0:01:40	Mix	Inter	X	Moderate	Med
EIP_008 [a]	0:00:26	Mix	Inter	X	Moderate	Med
Subtotal:	*0:08:45*					
Gunawan						
GUN_001 [a]	0:04:27	Mix	Inter		Poor	High
GUN_002 [a]	0:05:42	Mix	Inter		Poor	Very High
GUN_005 [a]	0:05:04	Mix	Inter		Very Poor	Very High
GUN_008 [a]	0:05:37	Mix	Inter		Moderate	Med
GUN_011 [a]	0:33:28	Mix	Inter		Very Poor	Very High
GUN_022 [a]	0:03:44	Mix	Inter		Poor	Very High
Subtotal:	*0:58:03*					
Jeremy						
JER_004 [b]	0:01:38	Mix	Inter	X	Good	Min
JER_013 [b]	0:02:02	Eng	Inter	X	Good	Med
JER_017 [b]	0:01:25	Mix	Inter	X	Good	Min
JER_020 [b]	0:01:25	Mix	Intra	X	Good	Med
JER_049 [b]	0:05:18	Eng	Intra	X	Moderate	High
JER_050 [b]	0:06:06	Eng	Inter	X	Moderate	Med
JER_051 [b]	0:07:13	Eng	Inter	X	Moderate	Med
JER_109 [b]	0:03:29	Ind	na	X	Poor	Med
Subtotal:	*0:28:38*					
Total training:	**1:35:26**					
INFERENCE DATA						
EIP_010	0:00:26	Mix	Inter	X	Moderate	Low
EIP_011	0:00:26	Mix	Inter	X	Moderate	Low
EIP_013	0:04:48	Mix	Inter	X	Moderate	Med
GUN_004_01	0:08:00	Mix	Inter		Moderate	High
GUN_004_10	0:08:00	Mix	Inter		Moderate	High
JER_019	0:03:07	Mix	Intra	X	Moderate	Med
JER_079	0:08:59	Mix	Intra	X	Good	Low
Total inference:	**0:37:45**					

[a] Subset of files used to fine-tune the *fb_NatInd* and *ind_nlp_NatInd* models.

[b] Subset of files used to fine-tune the *fb_JER_e60* and *ind_nlp_JER_e60* models.

Files are identified using part of their filename: E.g. EIP_002 refers to ZMS_EIP_002_L1-Alpha.wav.

Codes: *Language* - the dominant language, *Codeswitch* - whether inter- or intra-utterance switches appeared more common, *Audio quality* - a subjective judgement of ‘noise’ (call to prayer, unintelligible voices from other speakers, chickens, etc.), *Hyper* - the prevalence and degree of hyper-articulation.

Appendix E Extended Results

Table 6: Extended Inference Results

File	EIP_010	EIP_011	EIP_013	GUN_004_01	GUN_004_10	JER_019	JER_079
Words	35	598	629	654	847	333	992
Time	0:26	4:48	4:25	8:00	8:00	3:07	8:59
ind_nlp							
L:6	1	0	1	1	6	1	0
WER	80.00	79.93	83.47	73.70	83.47	87.09	94.05
CER	46.07	42.03	42.86	38.07	46.74	51.54	54.01
fb_all							
R	e	r	e	e	e	e	e
L:6	1	15	20	26	29	14	47
WER	28.57	43.65	44.36	28.75	41.20	33.33	36.29
CER	6.28	13.78	14.60	10.05	13.64	10.98	13.33
ind_nlp_all							
R	e	e	e	e	e	e	r
L:6	3	11	20	26	27	7	34
WER	22.86	46.82	43.40	31.65	35.42	37.84	43.45
CER	6.81	15.05	15.97	11.20	13.35	15.59	16.32
fb_nat_ind							
L:6	1	14	13	22	26	4	17
WER	31.43	52.51	46.42	33.18	43.09	65.47	51.82
CER	8.90	17.84	15.91	10.74	15.50	24.09	18.52
ind_nlp_nat_ind							
L:6	0	4	13	25	24	4	8
WER	42.86	52.01	47.38	32.42	41.20	62.76	60.69
CER	10.47	18.25	18.25	11.72	15.32	28.49	26.39
fb_JER_e60							
L:6	-	-	-	-	-	7	22
WER	-	-	-	-	-	38.44	44.36
CER	-	-	-	-	-	13.28	15.41
ind_nlp_JER_e60							
R	-	-	-	-	-	e	r
L:6	-	-	-	-	-	13	27
WER	-	-	-	-	-	40.24	45.67
CER	-	-	-	-	-	18.94	18.56

Colour — Coloured cells indicate best or equal best scores.
R — A rating given by a human transcriber for the perceived usefulness of the inference as a basis for editing. Inferences rated ‘e’ would be edited, and ‘r’ used as a reference while transcribing from scratch.
L:6 — The number of correct word sequences of length 6 and above.