EXTENDING THE REACH OF ALLOSAURUS: A UNIVERSAL PHONE RECOGNITION SYSTEM

Matthew R. Lee

Dallas International University &
SIL International
Matthew_Lee@diu.edu

ABSTRACT

The mighty Allosaurus of old is not known for its long reach, as its arms were only somewhat longer than the much-ridiculed Tyrannosaurus Rex, but in the world of Automated Speech Recognition, Allosaurus (**allo**phone **s**ystem of **au**tomatic **r**ecognition for **u**niversal **s**peech), developed at Carnegie Mellon University, seeks to use high-resource languages' data to extend the reach of phone-level automated speech recognition to low-resource languages. This paper is a trial of Allosaurus across recordings from two low-resource languages, Chamikuro of Peru and Ewondo of Cameroon, with analysis of the strengths and shortcomings of such a system.

Index Terms— multilingual speech recognition, universal phone recognition, phonology

Fig. 1: Allosaurus from Wikimedia Commons by Nekar. https://commons.wikimedia.org/wiki/File:Allosaurus_silhouette_01.jpg

1. ALLOSAURUS

1.1. Beginnings

Allosaurus, or **allo**phone **s**ystem of **au**tomatic **r**ecognition for **u**niversal **s**peech) [1], is the phone recognition system developed primarily by Xinjian Li at Carnegie Mellon University running behind www.dictate.app [1]. It is a machine learning system designed to transform audio recordings into written phonetic transcriptions. The learning model was trained on eight resource-rich languages with a wide phonetic inventory and uses what was learned to identify phones in uploaded audio regardless of language.

1.2. Discussion

When developed for high-resource languages, most ASR (automated speech recognition) systems are trained with large amounts of data in the target language (lexicons, running and tagged texts, and grammars) to identify segments at the word level, only outputting known or approved sequences. Every bit of training makes the system more accurate for that language, but if the granularity is too low, it has the side effect of locking it into that language and limiting transfer of learning, and the "creativity" of the system.

Rather than words, Allosaurus focuses on transcription into "phones" and is intentionally agnostic as to whether these are phonemes or their allo-

[1] https://www.dictate.app/phone

phones. An automated speech recognition system only has access to the fully-realized surface form of the utterance. This means that, whatever the desired output, each level of further abstraction beyond the surface form must be deciphered. This would be true of extension to a phonemic or orthographic representation. Even though the number and complexity of possible segments are significantly higher for phonetic segmentation, a phonetic transcription is the most direct goal for this type of system.

1.3. Customization

I was introduced to Allosaurus's early version during a workshop in Pittsburgh on Language Documentation for Low Resource Languages at Carnegie Mellon University. At the time of the workshop, when provided with an audio clip, Allosaurus would output a space-separated list of phones using its full repertoire of circa 180 phones that it had learned to recognize. While this wide vocal range would seem to be an advantage, in practice, relatively rare phones that were prevalent in the training languages were often unexpectedly common in the output. This meant that transcriptions frequently included phones like voiceless vowels and a voiceless b (both of which would normally be quite restricted in their environments). With all of the unexpected phonemes, human editing of the output was likely to be less efficient than typing the transcription from scratch.

The speech-processing workgroup found that by removing some of the more unexpected phones from the inventory, it was easy to improve the output, and removing some of the worst offenders quickly provided a considerable improvement in the output. Thus, limiting the inventory was shown to be an effective method of cleanup.

The user could provide the list of phones, and the system could adapt its output, only delivering phones from the list. An interface resembling an IPA chart[2] was developed to facilitate this workflow.

[2] https://www.dictate.app/chart

1.4. Language Templates

While the user could theoretically define each phone that they expected, I'm a big fan of useful templates that can be customized. When we found the wealth of phonemic (and sometimes allophonic) data on Phoible [2] for over 2300 languages, it was obvious that this could be used to bootstrap the system by providing presets for each language.

Link to PhoneInventory [3]

The Phoible dataset, after some cleanup and filtering, proved to be a valuable resource. This brings us up to the current project.

2. A CORD OF THREE STRANDS

For any machine learning system, there are three major elements that affect the output, the training data, the testing data, and the learning algorithm.

2.1. The Training Data

As shown in Table 1, Allosaurus's model was trained on the high-resource languages of English, Japanese, Mandarin, Tagalog, Turkish, Vietnamese, Kazakh, German, Spanish, Amharic, Italian, and Russian. The largest contributions to the corpus were from English, Japanese, and Mandarin.

These large corpora consist of aligned audio and orthographic text. The text was converted programmatically to phonemes using the profiles of the grapheme-to-phoneme mapping tool Epitran [4]. This was an ingenious solution to the problem, as it would quickly produce consistent and regular transcriptions from the orthographic data. Such a rule-based system is unlikely to reproduce the wealth of phonological variety that would be found in human transcriptions. This means that rather than teaching the system to transcribe from human transcriptions, the system was taught to transcribe as a human taught a system to transcribe, an unfortunate, but ultimately understandable, abstraction.

Table 1: Training corpora and size in utterances for each language. Reproduced with permission from ([Li, 2019 Forthcoming –ML])

Language	Corpora	Utt.
English	voxforge, Tedlium [5], Switchboard [6]	1148k
Japanese	Japanese CSJ [7]	440k
Mandarin	Hkust [8], openSLR [9, 10]	377k
Tagalog	IARPA-babel106b-v0.2g	93k
Turkish	IARPA-babel105b-v0.4	82k
Vietnamese	IARPA-babel107b-v0.7	79k
Kazakh	IARPA-babel302b-v1.0a	48k
German	voxforge	40k
Spanish	LDC2002S25	32k
Amharic	openSLR25 [11]	10k
Italian	voxforge	10k
Russian	voxforge	8k
Inukitut	private	1k
Tusom	private	1k

2.2. In search of ideal test data

Test data with the highest transcription quality would be a high-quality recording in a language included in the training data, but what fun is that? The goal of machine learning is to apply the model to novel data.

The second tier of ideal data is a high-quality recording in a language that only includes phones that are attested in the training data (Language *x* in Fig. 2).

The third tier of ideal data is a high-quality recording in a language with significant overlap of Allosaurus's training data (Language *y* in Fig. 2). You might have noticed that there are some constants. The first is a high-quality recording. This will be discussed in more detail in Section 3. If you were paying attention, you might have also noticed that the defining mark of each tier is the amount of overlap with the training data. Thus, more training data with more varied phones (and more exemplars of rare phones) would improve the recognition of Tier 3.

2.3. The Algorithm

2.3.1. Finding the Sweet Spot

One challenge for Allosaurus will be when phones exist in the training set but they do not exist in environments that are allowed in the language. Large enough datasets could be designed to include a wide variety of positions for each phone. Nevertheless, this is for the sort of thing that a machine Learning System is quite good at discerning.

Consider the Venn Diagram in Fig. 2. The largest circle contains all of the phones that the system has learned to recognize, about 180 phones with the eight-language dataset. The smaller circles contains an list of phones that are found in language *x* and language *y*.

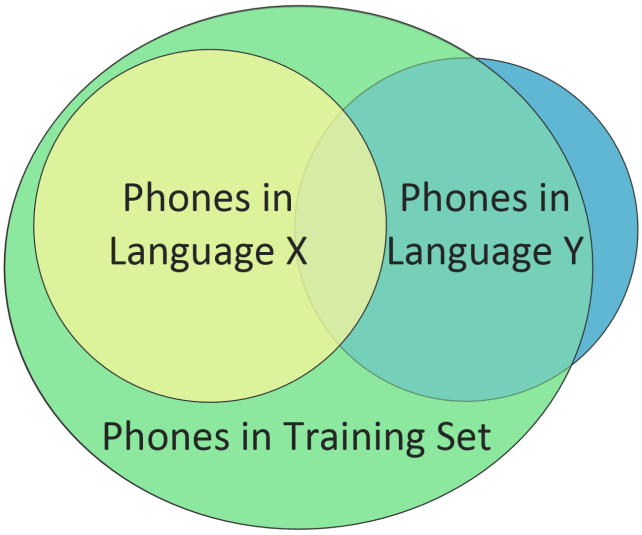

Fig. 2: A diagram of Phone Overlap Between Languages and the Training Set

The filters discussed in section 1.3 successfully limit the output options, effectively limiting transcription of Language *x* to the inventory of *x*. For languages with a relatively straightforward phone inventory, recognition is greatly improved, and it is easier to edit the few mistakes manually.

When we consider Language *y*, the phones that fall outside of the training set are not magically recognized, as the system has no profiles to recognize them. Any phone not in the training set will not appear in the output, and the system will be forced to fall back to the most similar sound existing in both inventories.

The downside of this approach is that any phone that exists in the language inventory but not the training set cannot be output. These are effectively blind spots of the system. In the current

version of Allosaurus (as of December 2019), there is no provision for recognizing phones that are not in the training set, even if they are phonetically similar to attested phones in the training set.

Future work will allow the use of phonetic distance to assist the system in finding a nearby phone from the language. For example, if a language that has retroflexed consonants instead of dental consonants. If the system incorrectly guessed dental consonants, the output could be adjusted using phonetic distance to realize that retroflex alveolar consonants are phonetically similar and choose those for output. This would allow the system to sometimes recognize sounds that are outside of its phone inventory.

3. DISCUSSION OF RECORDING QUALITY

Recording quality affects both ends of Allosaurus's performance. Many factors influence the quality of the test data. Some of these factors are discussed here. What follows is a discussion of these effects.

3.1. Effects of Excessive Microphone Distance

The first challenge found in the Chamikuro data was the distance from the microphone. As 80% of the training data for Allosaurus consisted of telephone recordings, the microphone would have been mere inches from the speaker's mouth. Though I don't know the origin of the other 20%, it was likely recorded in a near-studio situation. This means that the training set doesn't include significant data recorded from a large distance (far-field), and thus it does not know how to handle it.

What changes happen when an audio source is too far from the microphone?

3.1.1. Phonetic Detail

The most significant loss at a distance is fine phonetic detail. While phones that are resonant [12, p. 167] such as prototypical vowels will often be perceptible at a distance, voiceless stops and aspiration will be hard to identify and distinguish. This was the largest challenge with the Chamikuro data. Bilabial stops /p/, while common in the data, were very infrequently transcribed by Allosaurus. This was the ultimate reason that I abandoned pursuit of the Chamikuro transcription.

3.1.2. Low Volume and High Noise

Recording from a distance is quieter. Due to the inverse square law [**?**], as the Sound Source gets further from the microphone, the volume decreases exponentially. A voice six inches from the microphone will seem four times louder than the same voice twelve inches from the microphone.

If a speaker at a large distance is too quiet, the usual solution is to increase the gain in the microphone or recorder. While this does improve the volume of the speaker, it also amplifies everything else in the immediate surroundings. This increases noise in the recording. Noise can be loosely defined as anything in your recording you don't want.

With increased amplification, nearby fans, breathing, and outside noise will now also feature more prominently in the recording. The relationship between the desired data and the extraneous data is called the signal-to-noise ratio (SNR). A lower SNR usually means that the important data is less likely to be recoverable. While some minimally-intrusive noise can be reduced programmatically after the fact, some of the target recording data will be irreparably destroyed in the process. If strong noise in an important frequency range obscures the target data, the data may be unrecoverable.

As noise is to be expected in recordings, a reasonable amount of noise is often added to some training data which can improve the quality of transcription /cite??.

3.1.3. Echo

The second challenge introduced by a distant speaker is echo. Every non-soft surface within range of the microphone will produce an echo that will be audible in the recording. Distance and amplification increase the number of reflective surfaces that can negatively influence the quality of the recording. Thus you will start to find echo in the recording delayed from the original speech, obscuring later utterances.

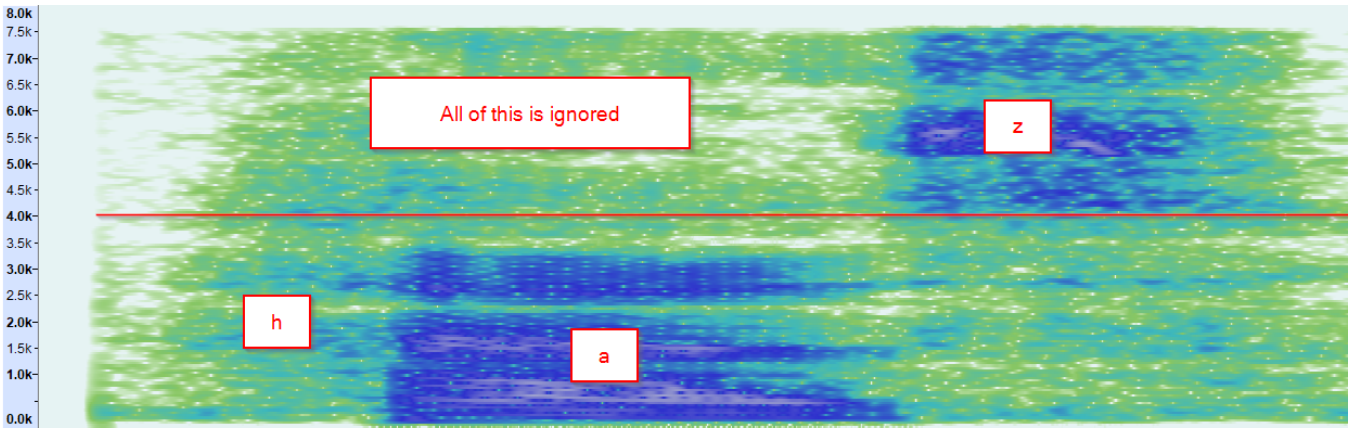

Fig. 3: A spectrograph of /haz/. All data above the red line will be lost when downsampling to 8kHz, including most of the amplitude of the voiced sibilant /z/

3.2. Frequency Response

Human hearing typically reaches from just above 20 Hz to just below twenty thousand hertz. Adults usually can only hear to 15 or 16 thousand hertz. As predicted by the Nyquist theorem, to faithfully sample digitally everything that a human can hear, we need to record at twice the highest frequency that can be heard. This is why 48,000 and 41,000 samples per second are quite common for recordings. Both sampling rates give just over twenty thousand Hertz that can be accurately sampled and reproduced.

Human voice typically falls between that 20 Hz and about 8000 Hz. This means that to faithfully sample the full spectrum of a human voice, one needs to record in excess of 16 thousand cycles per second. While 16,000 samples per second is the ideal minimum for speech, telecommunications technicians have realized that speech may be understandable at a lower sample rate. Telephone conversations are typically sampled at 8,000 cycles per second, which gives a maximum reproducible frequency of 4,000 Hz. Early Communication in space pushed the sampling down to about 5,000 samples per second reproducing only 2,500 Hz. This is why "One Small Step for Man" doesn't sound like a very good recording. The recordings in Chamikuro and Ewondo were collected at a standard 48,000 samples per second, and while this doesn't meet the current archival recommendations of 96,0000 samples per second, test data sampling in itself not a problem for our application.

The challenge is that the data that was used to train Allosaurus is largely telephone data. 80% of the corpus data was received at 8,000 samples per second, and the rest was down-sampled to match. This means that only 4,000 Hz and lower frequencies are retained, and while 4,000 Hz is perfectly acceptable for identifying the fundamental formants of vowels, the higher frequency fricatives, especially sibilants, are not fully captured. As shown in Figure 3, while it may still be possible to recognize sibilants via the lower frequencies, this is the acoustic equivalent of recognizing a close friend by their trousers rather than their face.

Silberer [13, p. 4] sums it up as follows:

> According to ANSI S3.5-1997, as much as 95% of the information necessary for speech recognition is provided when the available frequency bandwidth is 200 Hz to approximately 5-6 kHz. Still, some investigators suggest extending the bandwidth up to 9 kHz for both children and adults (Hornsby & Ricketts, 2006; Stelmachowicz, Pittman, Hoover, & Lewis, 2001). Stelmachowicz and colleagues (2001) have shown that when listening to children and female voices, some children will detect the phoneme /s/ more accurately with a bandwidth of 9 kHz compared to a bandwidth of

> 4 or 5 kHz. In contrast, other investigations have shown that extending bandwidth beyond 3-4 kHz does not always improve speech recognition performance and in some instances it may be detrimental to performance (Ching et al., 1998; Hogan & Turner, 1998; Turner & Cummings, 1999).

The rest of the data came at a higher sampling rate and was down sample to 8 kHz to match. This means that there is no high frequency data in the training set. Sadly, resampling the telephone audio data will not reveal data where there is none. The solution here seems to be include higher quality data, which will not be easy to find in such a quantity.

Due to the logarithmic nature of the Mel scale, doubling the recording rates of the audio will take up considerably more space, but it only adds about 6 new data points to the MFCC that will be used by the voice recognition system.

[Discuss Mel Scale –ML]

3.2.1. Selective Attenuation

The next challenge is attenuation of higher frequencies through air. According to my research, the higher frequencies excite the air molecules and produce heat. When the energy is transformed into heat, it ceases to propagate through the air as sound. The result is that higher frequencies are attenuated more over distance as they travel through the air. Low frequencies are the most efficient to transfer and are less susceptible to attenuation.

Thus, recorded voice from a distance will lack much of the higher frequency data that would have been captured with a nearer microphone.

3.3. Other Effects Influencing Recognition

3.3.1. Crosstalk

Crosstalk is another contributor to poor transcription. While the human mind is extremely talented at distinguishing and following multiple overlapping speakers, seamlessly performing segmentation and recognition. Contextual hints such as directionality and difference of voice are mixed into the same waveform with considerable overlap. Without the simulation of both voices and a lexicon of possibilities needed to untangle them [**?**], sections of crosstalk are outside the scope of Allosaurus and are left to be transcribed by hand.

3.3.2. Prosody

The best recognition will be performed on test data that matches the training data. For clear recognition, it is important that exemplars exist of each segment that needs to be recognized in similar prosodic contexts. While cross-speaker differences such as voice pitch should be normalized through the filtering of F0 that is normally done in ASR, prosodies as extreme as whispering and shouting are unlikely to be recognized correctly and unfortunately adding such extreme training data is likely to reduce the quality of ”normal” transcription, especially with voiceless vowels.

3.4. Repairing the Audio

In most situations, one would just go back and re-record or annotate the audio. This is probably an example of a case where the BOLD methodology would be perfect. In this case, the speaker is in Peru and there are very few native speakers remaining. Thus, within the scope of this project this is not possible.

4. TRIAL 1: CHAMIKURO TRANSCRIPTION

Chamikuro is...

Upon learning that there were some recordings of Chamikuro, a language of Peru, I intended to use Allosaurus to do transcriptions and analyze the output. I had the advantage that my professor has published significant information on the phonology of Chamikuro, giving me a starting point for configuring the language’s phone inventory.

4.1. Data Preparation

The first task was to identify and segment the data, the first of which was a Swadesh wordlist in Spanish and Chamikuro. The recording was an interchange between a Spanish speaker who would

prompt the native speaker, and two repetitions of the word in the language. In about 10 minutes of audio, about three minutes of speech containing the language was found.

Upon listening to the data, it was obvious that the recording was not ideal. It seems that the Swadesh recording was done with the microphone off to the side of the speaker and prompter on a table. Each speaker seems to be several feet from the microphone, and the prompter was considerably louder than the speaker. The target speaker was elderly and soft-spoken, and the prompter had an unfortunate tendency to speak over the target speaker's repetitions. Nevertheless, to my ear, the words seemed quite intelligible, so I didn't worry too much.

The second recording was an autobiography of the same speaker in the language. The speaker was seemingly more distant from the microphone and spoke more softly than during the recording of the Swadesh list.

4.2. Phonetic and Phonemic Inventory

For Chamikuro, the data from Phoible's Chamikuro profile was used.

One list included only the phonemes as defined by Parker [14]:
/a/, /aː/, /ç/, /e/, /eː/, /h/, /i/, /iː/, /j/, /k/, /l/, /ʎ/, /m/, /n/, /ɲ/, /o/, /oː/, /p/, /r/, /s/, /ʂ/, /ʃ/, /t/, /ts/, /t͡ʃ/, /u/, /uː/, /w/, /ʔ/

A second list contained all phonemes and allophones as defined by Parker [14]:
/a/, /a̟/, /aː/, /æ/, /ç/, /ɔ/, /çʰ/, /dl/, /e/, /e̟/, /ə/, /eː/, /ɛ/, /h/, /i/, /ɪ/, /i̟/, /iː/, /j/, /j̟/, /k/, /kʰ/, /l/, /ɬ/, /m/, /n/, /ɲ/, /ŋ/, /o/, /oː/, /p/, /pʰ/, /r/, /s/, /ʂ/, /ʃ/, /t/, /tʰ/, /ts/, /tʃ/, /tʃʰ/, /u/, /uː/, /w/, /x/, /ʎ/, /ʔ/

Note that the phones /a̟/, /çʰ/, /dl/, /e̟/, /i/, /j̟/, /ɬ/, and /tʃʰ/ are not included in the training set, so they will never be output in the current version of Allosaurus. This is similar to language *y* in Fig. 2.

My expectation was that transcription with the phonetic inventory would output a narrow transcription, and the phonemic listing would "fail over" to the phonemic forms that are nearest their allophonic counterpart. I did not expect any representation of tone in either transcription (see section **??**, as this content is ignored.

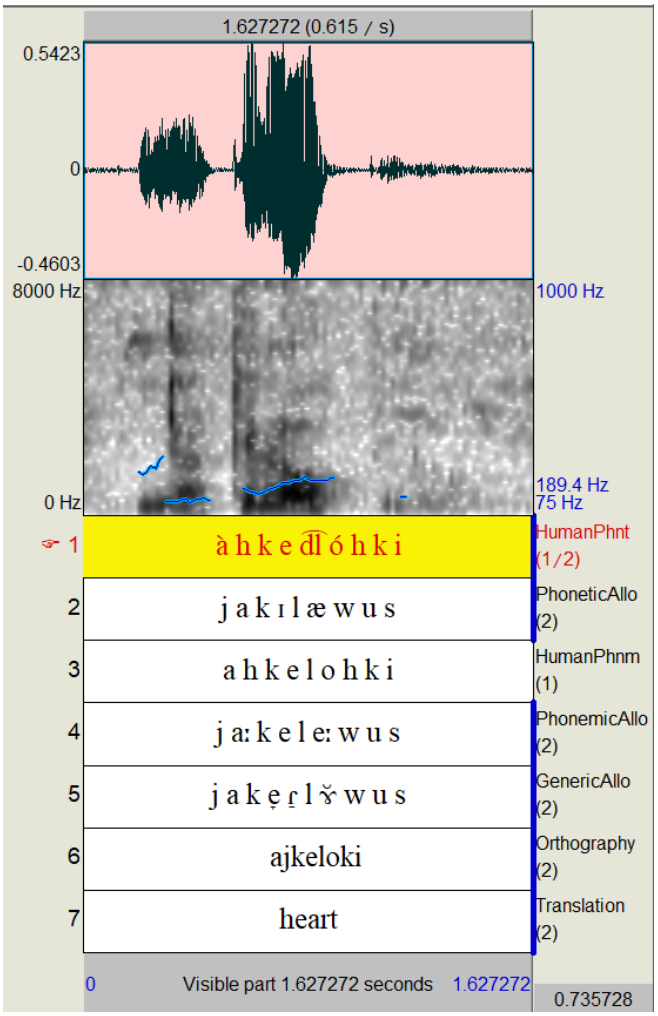

Fig. 4: An analysis of *ajkeloki* 'heart' in Chamikuro.

4.3. Chamikuro Results

Running Allosaurus on data from the language produced results as shown in Fig 4.

In the Chamikuro figures, the rows of transcription are as follows:

- HumanPhnt: a manual phonetic transcription
- PhoneticAllo: Allosaurus's transcription limited to all attested phonemes and allophones in the language

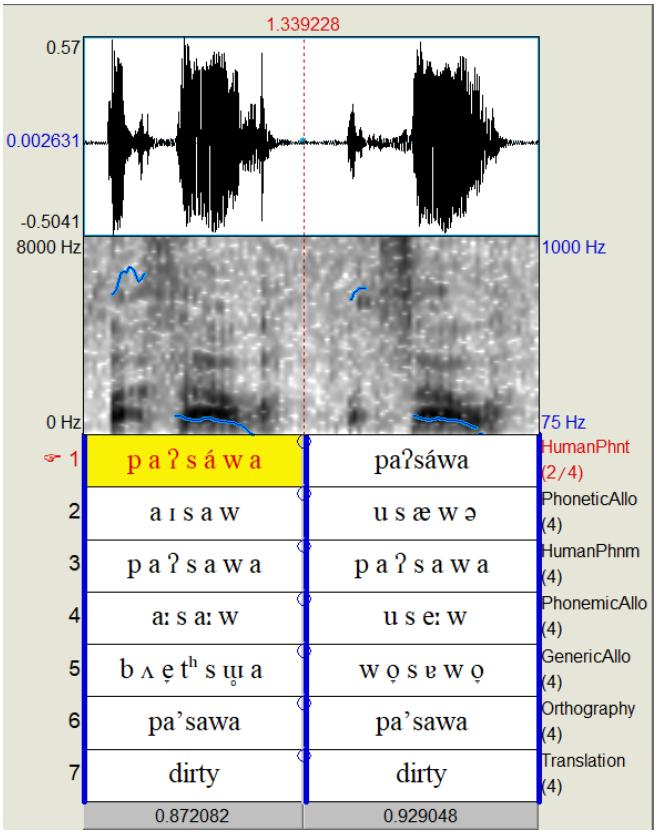

Fig. 5: An analysis of *pa'sawa* 'dirty' in Chamikuro.

- HumanPhnm: a manual phonemic transcription.
- PhonemicAllo: Allosaurus's transcription limited to attested phonemes in the language
- Orthography: An orthographic transcription.
- Translation: An English translation of the segment

4.3.1. The Good

In these results, vowels were quite consistently recognized. From the point of view of perception, vowels are especially resonant and most likely to be perceived, even at a distance. It makes sense that the most resonant sounds, like vowels would be the ones that would be maintained over the distance. These have plenty of robust harmonics that could be used to identify the phone.

Intervocalic consonants were consistently recognized in louder segments (see /akə/ in figure Figure 4 and notice the clear vertical line of the first /k/ in the spectrogram.

4.3.2. The Bad

Sibilants and affricates were hit-and-miss, but to be fair many of the retroflexed affricates were not in the training inventory ([is this true? –ML]).

4.3.3. The Ugly

Consonants, especially voiceless stops, were often conspicuously absent. The last syllable /ki/ of Figure 4, was never recognized by Allosaurus, even though it is clearly audible to the human ear. The initial /p/ in Figure 5 is also never realized by Allosaurus.

This should not be surprising as stops are some of the lowest phones on the sonority scale and they are likely to be hard to recognize from a distance from the noise that obscures it. See section 3.1.1 for more discussion of loss of phonetic detail. Without a large percentage of consonants, I had no hope of analyzing the syllable structure of the language.

See Wolf [**?**].

4.3.4. Conclusions

The Chamikuro results were disheartening as I have seen much better output for other languages. As discussed in section 3, it is clear that the recording was not acceptable for this use, and despite my best efforts to clean up the recording (see Appendix **??**) and to game the phonetic inventory with different variations, I was not able to improve the results.

Interestingly, the generic output of Allosaurus in Figure 5 was actually the only one to recover an initial consonant, even though it was incorrect. It is hard to say whether Allosaurus's Phonetic, Phonemic, or Generic annotation was the best for Chamikuro.

5. TRIAL 2: EWONDO

Ewondo is...

As the Chamikuro data did not provide the quality I had seen with other recordings and lan-

guages, I decided to return to Ewondo, where I had already seen some promising results.

5.1. Data Preparation

Allosaurus expects short segments of audio, and shorter segments are easier to transcribe, so the data needed to be segmented into clips. The 90-second clip was segmented in SayMore into 14 clips at pauses for easy transcription. I then manually transcribed the entire recording to have a basis for comparison.

The recording quality was quite good, and it seems that a nearby microphone was used, though probably not a headset microphone. As a result, some of the least sonorant consonants were less prominent, but still greatly improved from the Chamikuro data.

The Orthographic transcription was transcribed by my colleague NGONO Louis Pascal, a native Ewondo speaker. Some segments were not transcribed in the first pass, and I had to transcribe them later, so any items in brackets in the orthography section are my own unverified transcription.

5.2. Phonetic and Phonemic Inventory

The Phonemic Inventory used for Ewondo was:
/a/, /b/, /d/, /dz/, /e/, /ɛ/, /ə/, /f/, /g/, /gb/, /i̯/, /j/, /k/, /kp/, /l/, /m/, /mb/, /ŋv/, /n/, /nd/, /ndz/, /ɲ/, /ŋ/, /ŋg/, /ŋmgb/, /o/, /ɔ/, /s/, /t/, /ts/, /u/, /v/, /w/, /z/ [**?**]

The Phonetic Inventory used for Ewondo was:
/a/, /æ/, /ɐ/, /b/, /b̥/, /d/, /dz/, /dʒ/, /d͡ʒ/, /e/, /ə/, /ɛ/, /ɘ/, /f/, /g/, /gb/, /h/, /ɦ/, /i/, /j/, /k/, /k͡p/, /k͡pˈ/, /l/, /m/, /m͡b/, /mv/, /ŋ̟/, /ŋ̟v/, /n/, /nd/, /ndz/, /nj/, /nʲ/, /ɲ/, /ŋ/, /ŋ͡g/, /ŋ͡m/, /ŋmgb/, /o/, /ɔ/, /p/, /r/, /ɾ/, /s/, /t/, /ts/, /t͡s/, /t͡ʃ/, /u/, /ɯ/, /v/, /w/, /z/ [**?**]

The following phones are not in the training set, but most of them are complex segments. Allosaurus doesn't really distinguish most affricates as combined units, and is more likely to find the individual elements instead (i.e. /n/ and /d/).
/æ/, /dz/, /gb/, /kp/, /mb/, /mv/, /ŋv/, /nd/, /ndz/, /nj/, /ŋg/, /ŋm/, /ŋmgb/

5.3. Results

All of the Ewondo transcribed segments are available at the end of this paper. For the Ewondo annotations, the rows in each annotation are defined as follows:

- Clip: The reference number of the audio clip.
- HumanPhnt: a manual phonetic transcription
- PhoneticAllo: Allosaurus's transcription limited to all attested phonemes and allophones in the language
- PhonemicAllo: Allosaurus's transcription limited to attested phonemes in the language
- Orthography: An orthographic transcription, items in brackets were not transcribed by a native speaker.
- Translation: An English translation of the segment

Phone Levenshtein: 393 Number of Phones in Reference: 751 52% Phone Error rate

Character Levenshtein: 414 Number of Chars in Reference: 776 53% Character error rate.

6. MAKING ALLOSAURUS BETTER

One obvious future improvement for Allosaurus would be the addition (or improvement) of suprasegmental identification, such as tone, stress, and length.

6.1. Timing Data

All (or nearly all) timing data is normalized (lost) in the output of Allosaurus. This is with the exception of the order of phones. If a /k/ was recognized, the system cannot report a specific time-range within which it is located. This is unfortunate, as this information would be useful in identifying segment errors, but one already has the data needed to align the transcribed text with the audio. A standard method is to use a text-to-speech system such as eSpeak [**?**] to generate a waveform from text, and then use a forced-alignment tool to align the intensity spikes of the two audio segments.

6.1.1. Ambiguous Segments

Related to this phenomenon are some inconsistencies in the training data. Transcription is a process where many decisions are made based on non-phonetic information, such as morae (segment length) or syllable structure. In the list of 180+ phones in the training set, and there are many language-specific distinctions made on the grounds of timing or syllable structure, and it is my understanding that the system is largely blind to timing changes. If this is true, it does not make sense to output unreliable information that was in the training set but that the system cannot evaluate.

I know that Allosaurus cannot recognize any segment not attested in the training data, and with my custom inventory, both /t s/ and /t͡s/ are being output in the Ewondo transcriptions. This means that there are exemplars of both in the dataset, and the system is attempting to distinguish between them. Unfortunately, the difference between /t s/ and /t͡s/ is timing-related at best, and at worst a language specific convention used when one consonant is needed instead of two. It does not seem helpful to make this distinction at a language-independent the phonetic level, and no language is likely to need to distinguish /t s/ from /t͡s/. Instead, all exemplars trained for either /t s/ or /t͡s/ should be treated as the same segment, and "displayed" as one or the other based on the provided inventory. I propose normalizing the transcription of the training data to remove all tie bars and treat all consecutive phonemes as sequences. This will effectively give more exemplars of each individual segment.

The next problematic group contains sequences like /gʷ/, a labialized voiced velar plosive. Like the ambiguous sequences above, the distinction between /gʷ/ and /g w/ is mostly language specific, but may be based on slight changes in timing and intensity or phonololgy. Again, a language is unlikely to distinguish phonemically between /gʷ/ and /g w/, and the system is unlikely to be able to successfully reliably distinguish the segemnt. I propose a trial where all segments with the superscript notation (palatalization, prenasalization, labialization, etc) are normalized to their full-size IPA representation before training the model. If an "equivalent" grouping is given in the language inventory, then the output should display the output (/gʷ/ or /g w/) in the required format.

6.1.2. Vowel and Consonant Length

Vowel and consonant length is another scalar suprasegmental that is dependant on context, prosody, language and speaker and can not be reliably determined without intimate knowledge of the language. The system does not reliably distinguish between long and short vowels, and doing so from a single utterance is not feasible without a reference for comparison. I propose a trial setup where all segments tagged with length in the dataset /ː/ are normalized to remove length distinctions, as well as removing syllabic marks, if they are included in the training set. This would prevent the system from analyzing vowel length, but I don't think this would actually introduce any new errors. If this is unacceptable, a second pass armed with the timing data from section 6.1 could be devised to run over large datasets to measure all segments of each type (/a/, /e/, etc) that allow length and classify them by length only if length is included in the language inventory. This would likely require a cached corpus of previously transcribed data, as there would not be enough data in a single utterance to make any distinction.

6.1.3. Further "equivalent" segments.

One of the most common unusual phones in the full-repertoire output is /b̥/, a devoiced voiced bilabial stop. To begin with, this is an odd segment, and most would cancel the double negative and call it /p/, a voiceless bilabial stop. It is most likely transcribed when a phonemic /b/ is devoiced in a specific environment, and the transcriber wishes to maintain the underlying distinction. From a phonetic standpoint, /b̥/ and /p/ should be considered phonetically equivalent and normalized to /p/ in the training set, and the /b̥/ should only be output when requested in the phonetic inventory. The same is true for other equivalent devoiced consonant segments such as /d̥/ and /t/.

Devoicing in vowels, while phonetically faint, is sometimes phonemically relevant and is only

marked in one way, so voiceless vowels should be maintained in the system.

6.2. Allophones and Environments

At this point, Phone inventories for a language are flat, and there is no way to describe an allophonic relationship. I'm not sure how allophonic relationships between could be taken into account, or whether this should be at the main analysis or decoding stage.

Allophonic forms are often restricted to a specific an environment, such as intervocalically or a the end of a prosodic word. Some of the striking oddities of the output are finding allophones that are normally restricted to a specific environment outside of that environment. While Allosaurus cannot be expected to predict those language-specific environments, an interface could be developed where one with significant understanding of the language could create rules that specify the appropriate environment. Specification of these environments (or the simpler solution in the previous paragraph) could also be reversed to work backwards to a phonemic representation. For example, if the phoneme /i/ is devoiced after a an unvoiced stop and at the end of the prosodic word, the rule would be:

$$/\text{i}/ \rightarrow [\underset{\circ}{\text{i}}] \Big/ \begin{bmatrix} -continuant \\ -voice \end{bmatrix} \_ \#$$

6.3. Tone

It is my understanding that in most automated speech recognition systems, the first formant (f0) is ignored, as it says more about the vocal tract geometry (and by extension, size and pitch range) of the person than it does about the phoneme being uttered. ASR Technicians consider the first formant to be misleading noise. This poses a problem for tonal languages, where tonal variation (as opposed to the starting point) is necessary.

Tone (pitch) variation can be extracted from an utterance, and could be overlaid if timing data was still present in (or was aligned with) the main phonetic output. A very similar second pass could take all utterances by a speaker, normalize the tone range to a standard minimum and maximum, and apply it only to tone-bearing phones. Limitation to a language-specific set of "expected" tonemes would limit the noisy output to only appropriate representations. Obviously this is non-trivial, but it seems a priori that this secondary analysis would be best left for a later pass over the data after the phonetic transcription, and only when requested.

7. REFERENCES

[1] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littel, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Florian Metz, "Universal Phone Recognition with a Multilingual Allophone System," 2020, p. 5.

[2] Steven Moran and Daniel McCloy, Eds., *PHOIBLE 2.0*, Max Planck Institute for the Science of Human History, Jena, 2019.

[3] Matthew Lee, "LTLDR19-speech/phoneInventory," 2019.

[4] David R Mortensen, Siddharth Dalmia, and Patrick Littell, "Epitran: Precision G2P for Many Languages.," in *LREC*, 2018.

[5] Anthony Rousseau, Paul Deléglise, and Yannick Esteve, "TED-LIUM: an Automatic Speech Recognition dedicated corpus.," in *LREC*, 2012, pp. 125–129.

[6] John J Godfrey, Edward C Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1992, vol. 1, pp. 517–520.

[7] Kikuo Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[8] Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff, "Hkust/mts: A very large scale

mandarin telephone speech corpus," in *Chinese Spoken Language Processing*, pp. 724–735. Springer, 2006.

[9] Xingyu Na Bengu Wu Hao Zheng Hui Bu Jiayu Du, "AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline," in *Oriental COCOSDA 2017*, 2017, p. Submitted.

[10] Zhiyong Zhang Dong Wang Xuewei Zhang, "THCHS-30 : A Free Chinese Speech Corpus," 2015.

[11] Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *INTERSPEECH-2005*, 2005.

[12] G N Clements, "Does sonority have a phonetic basis? Comments on the chapter by Vaux," *Contemporary Views on Architecture and Representations in Phonological Theory*, pp. 165–175, 2009.

[13] Amanda Beth Silberer, Ruth Bentler, and Yu Hsiang Wu, "The importance of high-frequency audibility with and without visual cues on speech recognition for listeners with normal hearing," *International Journal of Audiology*, vol. 54, no. 11, pp. 865–872, 2015.

[14] Steve Parker, "The sonority grid in Chamicuro phonology," *Linguistic Analysis*, vol. 19, pp. 3–58, 1991.

A. AUDIO CLEANUP

Several challenges related to the quality of audio were identified in the section on Data Preparation. While I attempted to clean up the Chamikuro data to achieve a better result, my efforts were ultimately fruitless. Nevertheless, this section contains discussion on attempts to repair audio data.

A.1. Repairing Volume

As one would expect, garbage in garbage out. When recording for language documentation, it is expected that one will use a quality microphone close to the speaker. This creates the ideal situation for less noise and high fidelity of the speaker's voice. Increasing distance creates negative effects such as Echo and greater noise. In the case where volume is not ideal, adjusting the volume is straightforward. I applied in normalization that searches for the highest peaks in the recording and brings them down to an appropriate level. I alternately tried compression, which will only quiet the loudest sounds in a recording. The result, as I said before, seemed to be a perfectly acceptable volume level.

A.2. Noise Reduction

Typically noise is removed by analyzing several seconds of non-speech in the environment, and removing similar sound from the rest of the recording. This works very well in quiet segments, but often noise is either not removed from the speech data or desired speech data is partially removed along with the noise. If the noise is removed before compression, some of the data that you might have been able to recover is lost. If the noise is removed after compression, some of the noise that has been amplified may not be completely removed. This is a catch-22.

A.2.1. Repairing Echo

While it is superficially easy to introduce Echo into a recording, it is very difficult to reliably remove. If I had a clean copy of the audio, I could generate the echo to get a profile and then subtract that Echo from the original recording. Of course, if I had a

clean copy of the recording, I wouldn't be doing this process. With the available toolset, all I could expect was to hope to remove some of the echo in the quiet moments with standard noise reduction.

A.2.2. Repairing Attentuation

If the high frequencies are quiet but still recoverable, it might be possible to boost them using an equalizer. In a humid environment, and that part of Peru is a humid environment, the curve of attenuation is almost a straight line. Thus, I tried decreasing the low frequencies while increasing the high frequencies to simulate being nearer to the speaker.

Ewondo Data

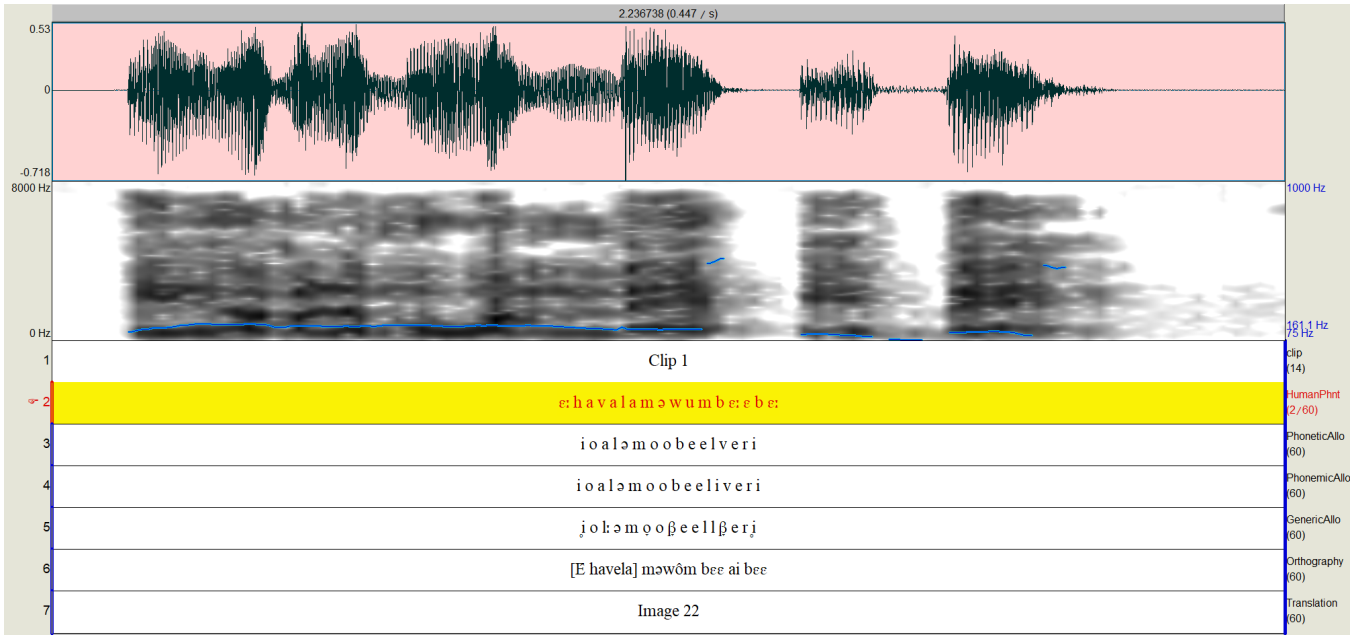

Fig. 6: Ewondo Clip 1

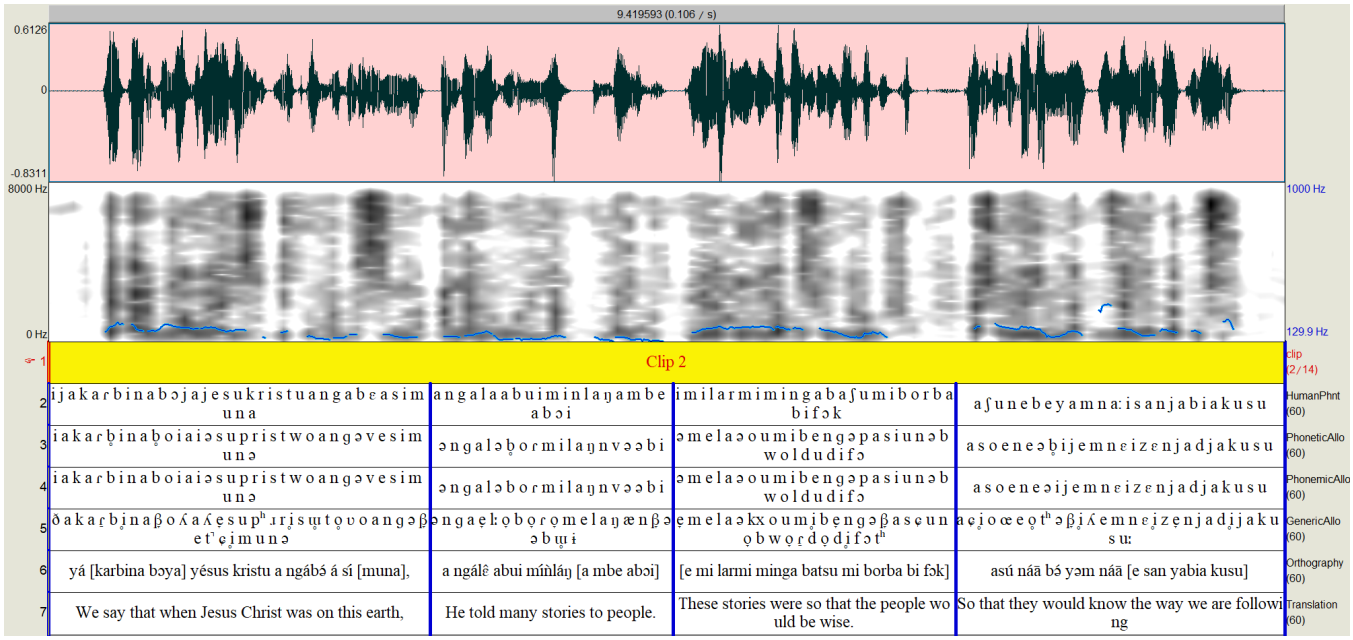

Fig. 7: Ewondo Clip 2

Fig. 8: Ewondo Clip 3

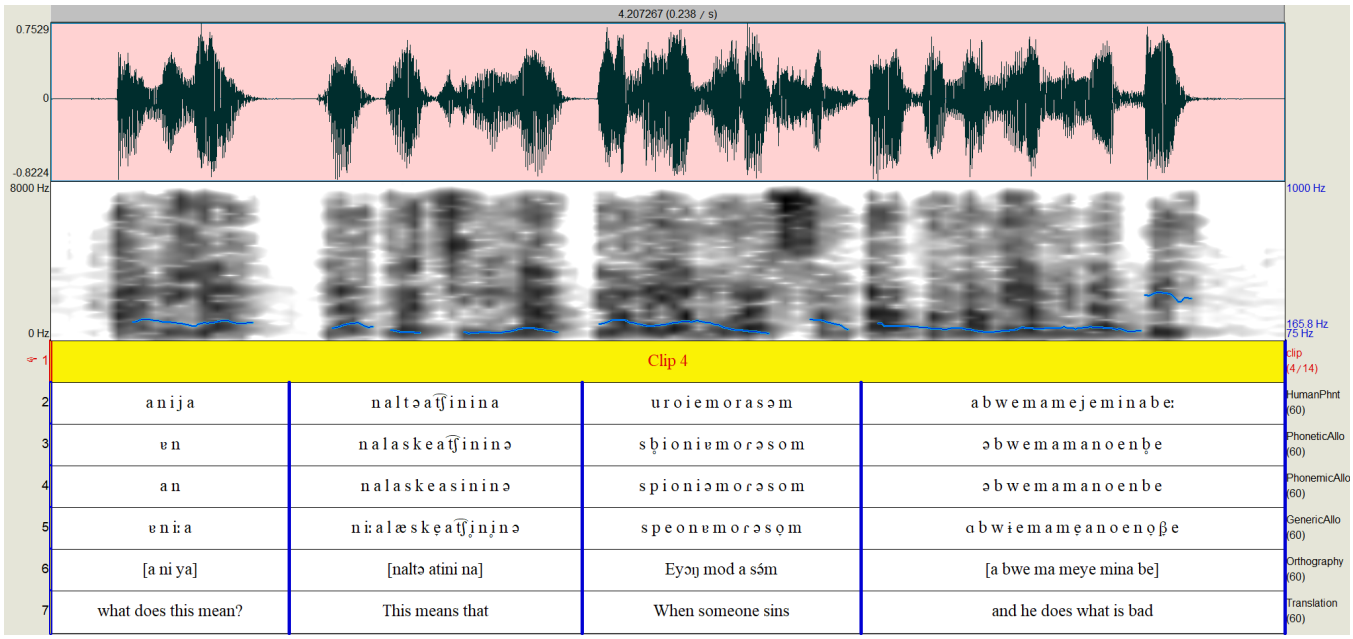

Fig. 9: Ewondo Clip 4

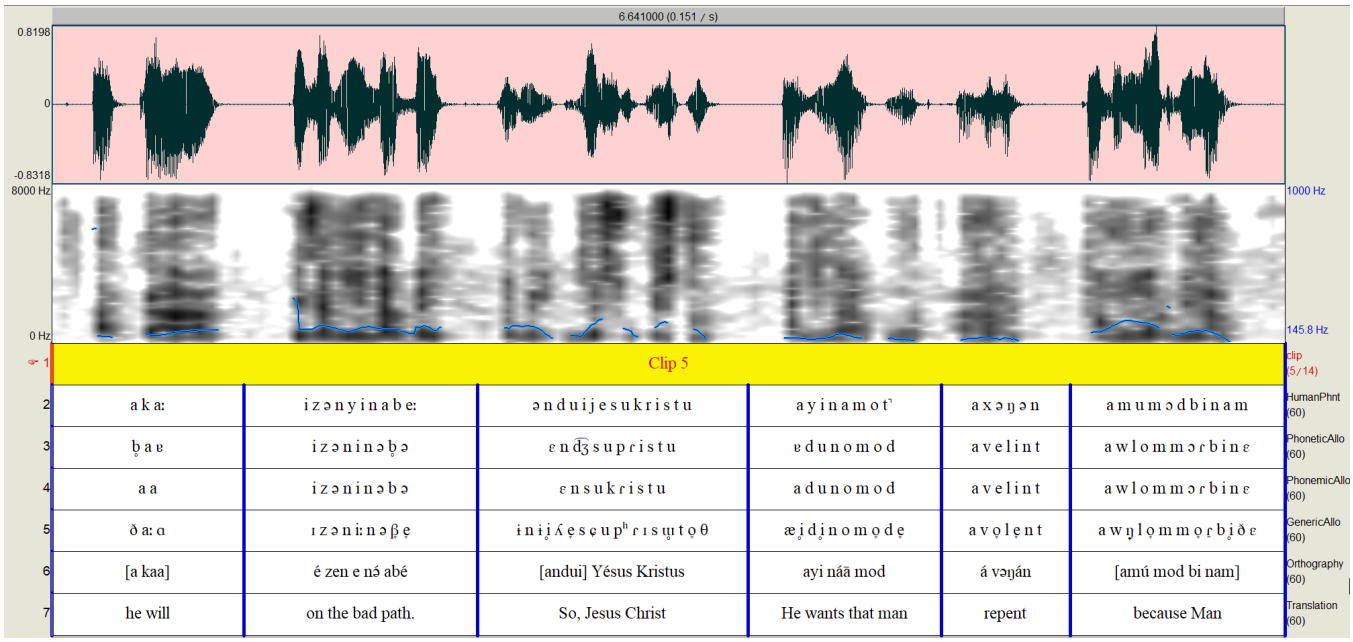

Fig. 10: Ewondo Clip 5

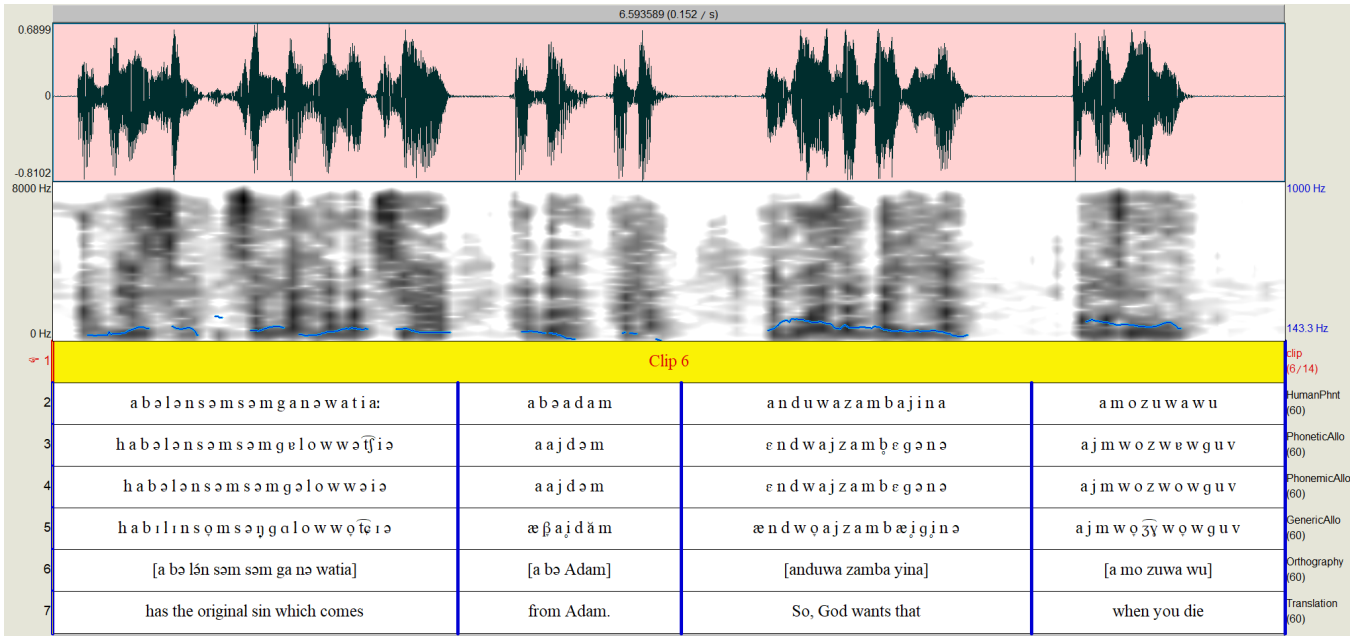

Fig. 11: Ewondo Clip 6

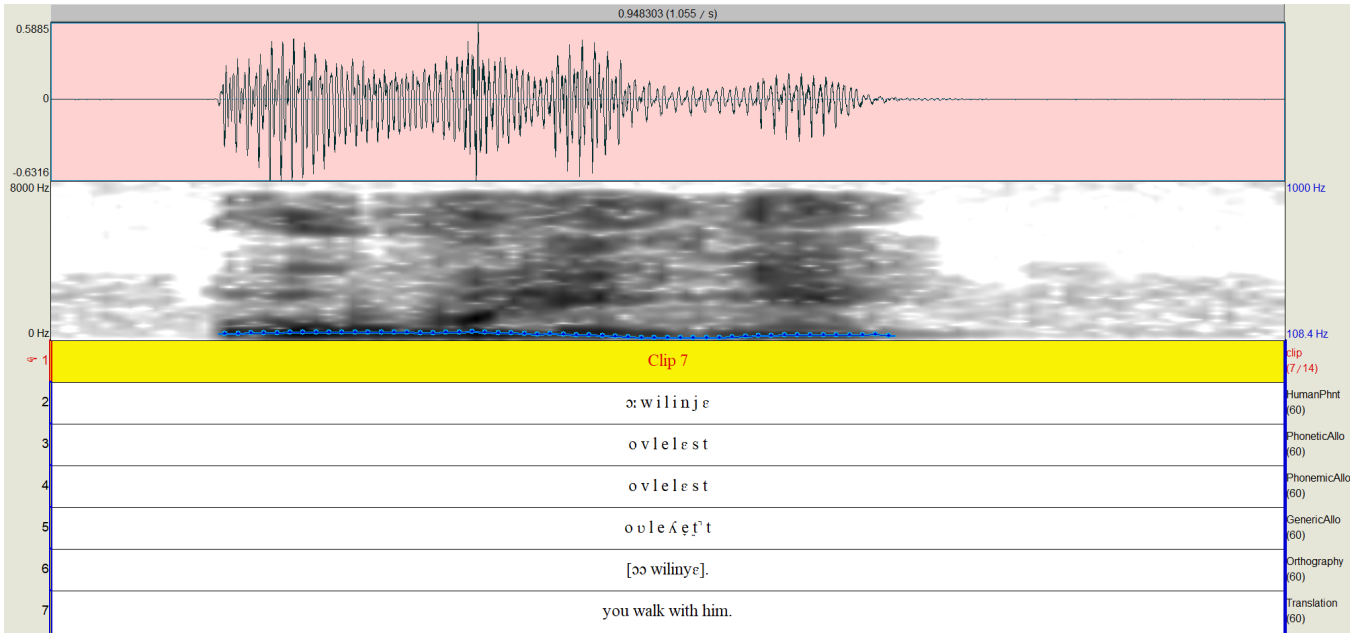

Fig. 12: Ewondo Clip 7

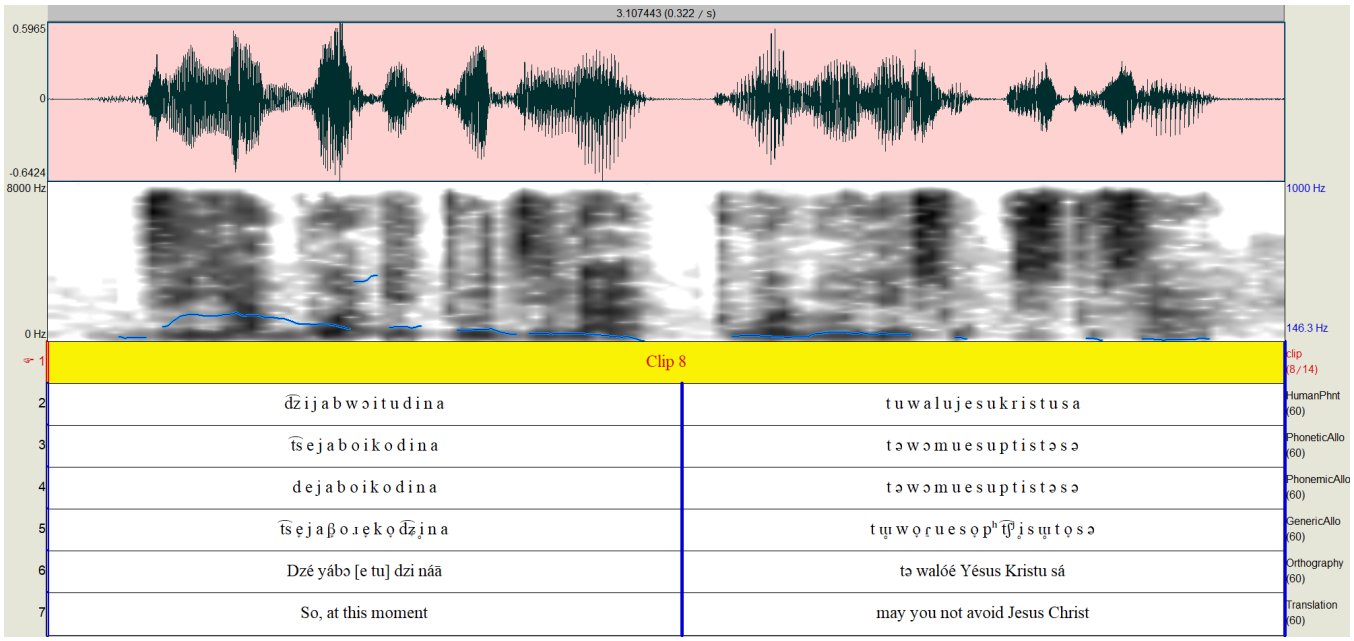

Fig. 13: Ewondo Clip 8

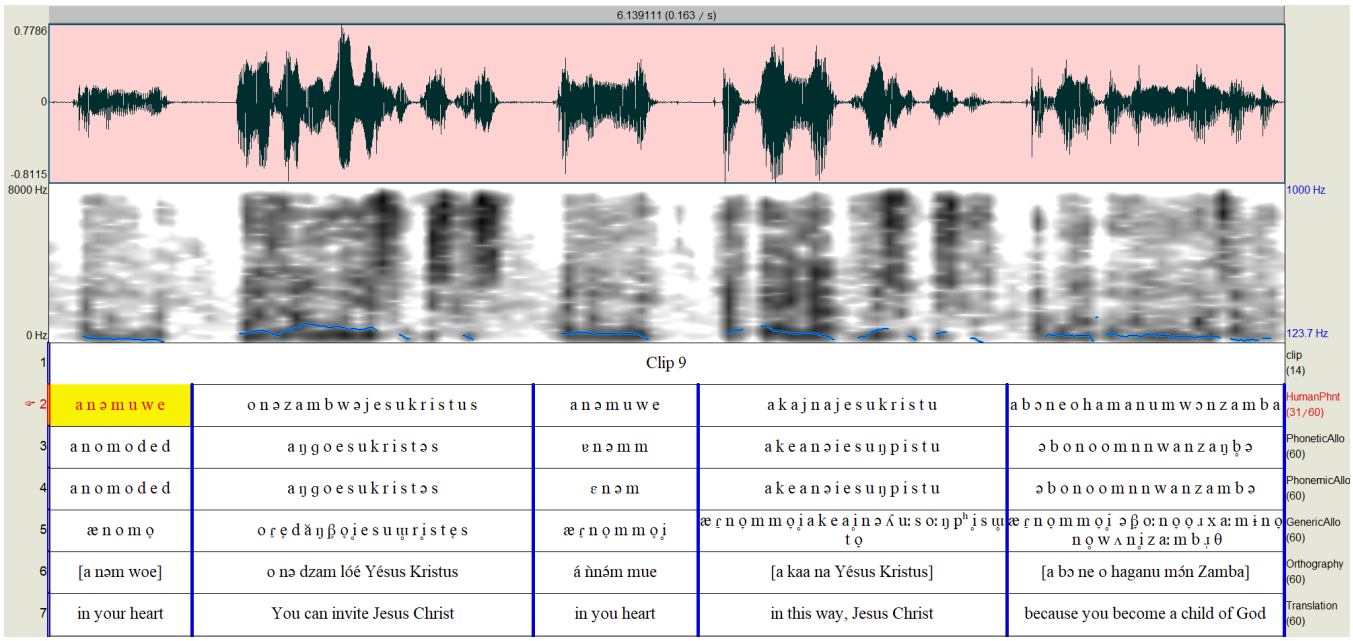

Fig. 14: Ewondo Clip 9-1

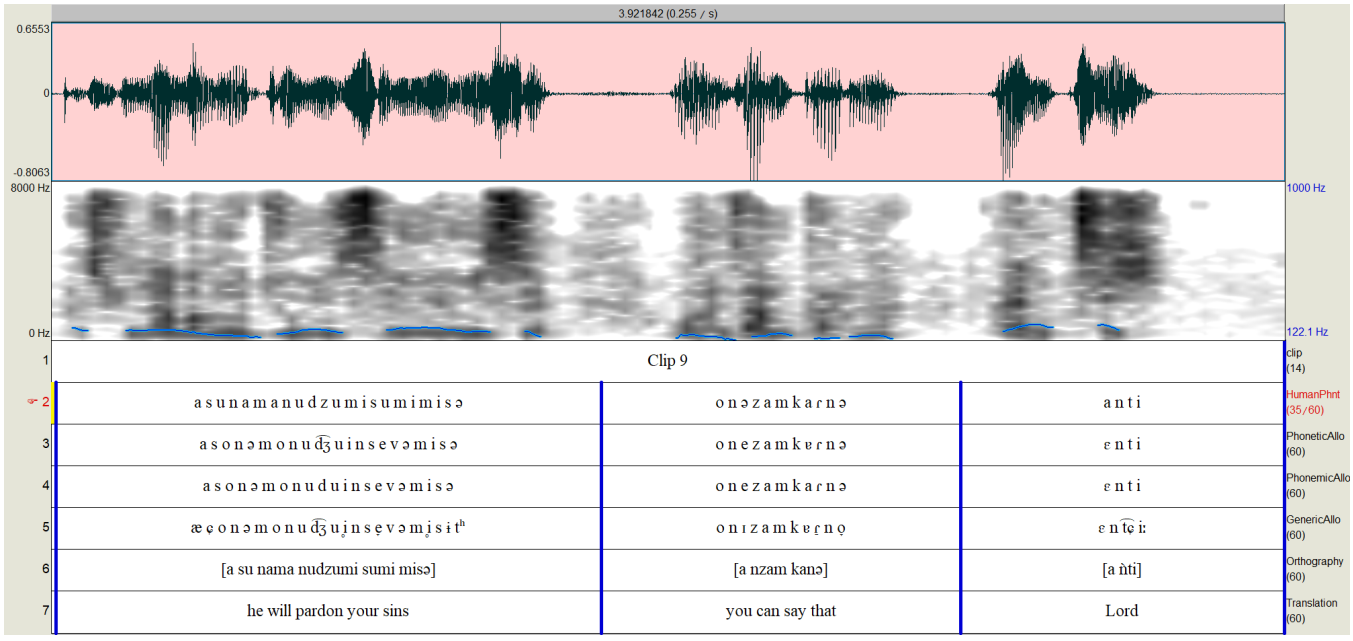

Fig. 15: Ewondo Clip 9-2

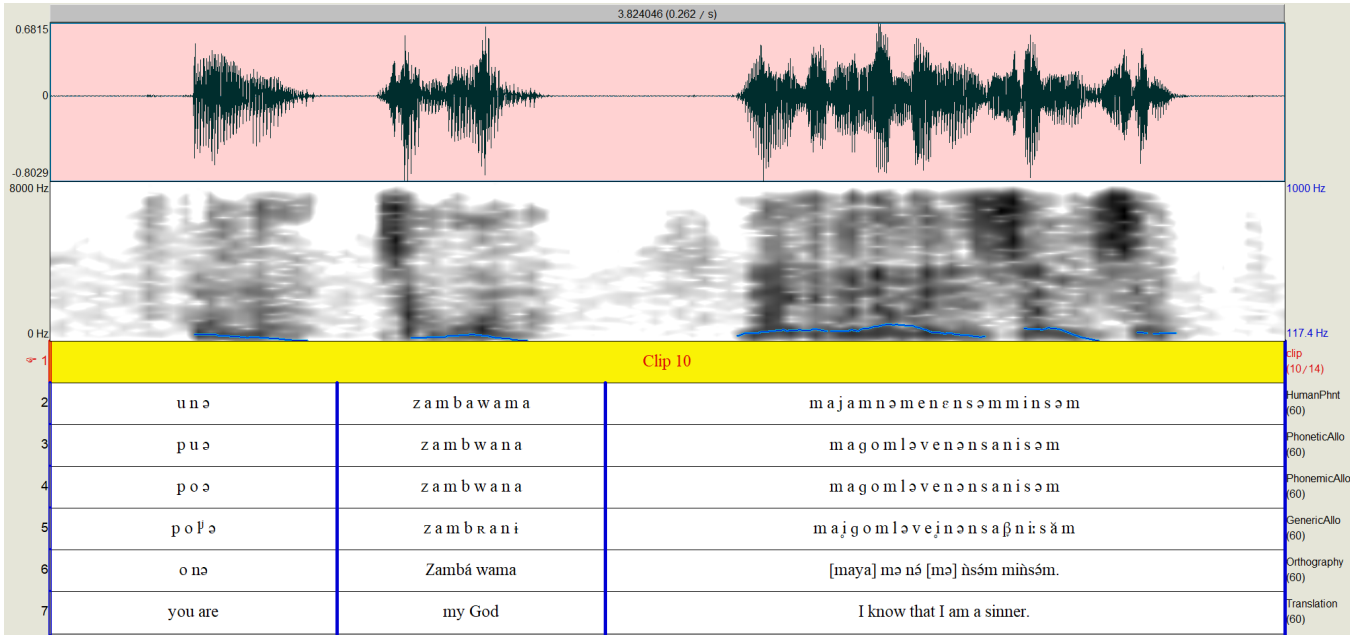

Fig. 16: Ewondo Clip 10

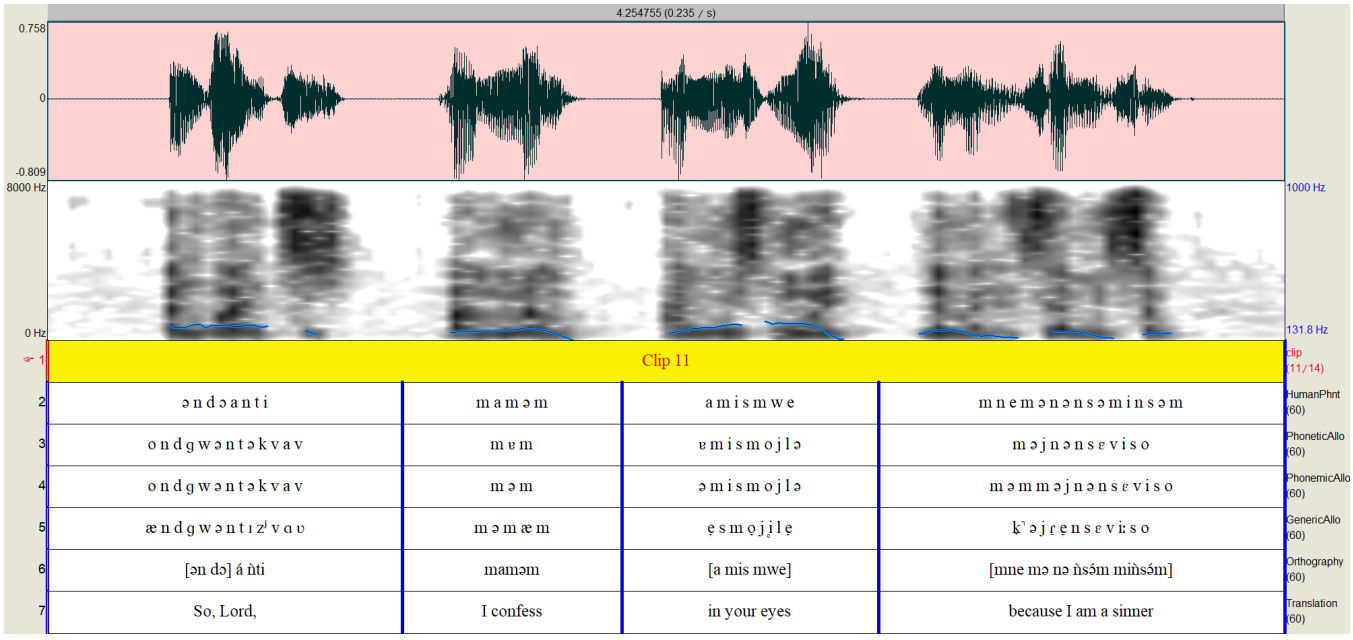

Fig. 17: Ewondo Clip 11

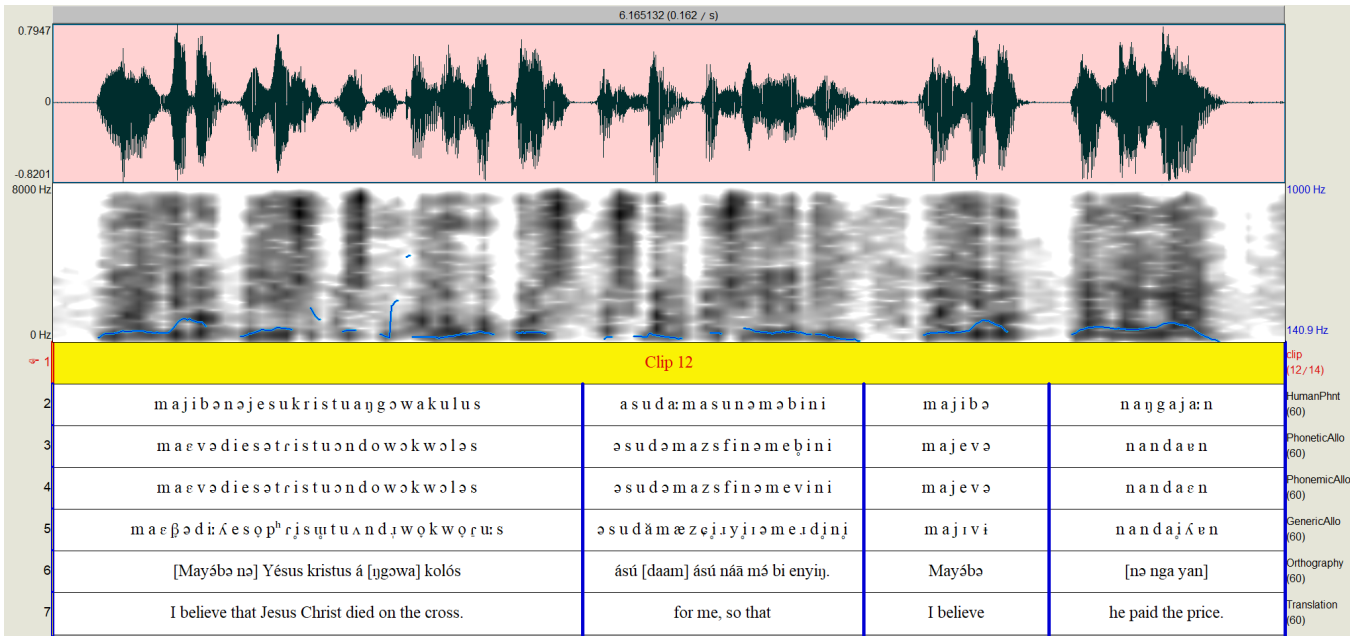

Fig. 18: Ewondo Clip 12

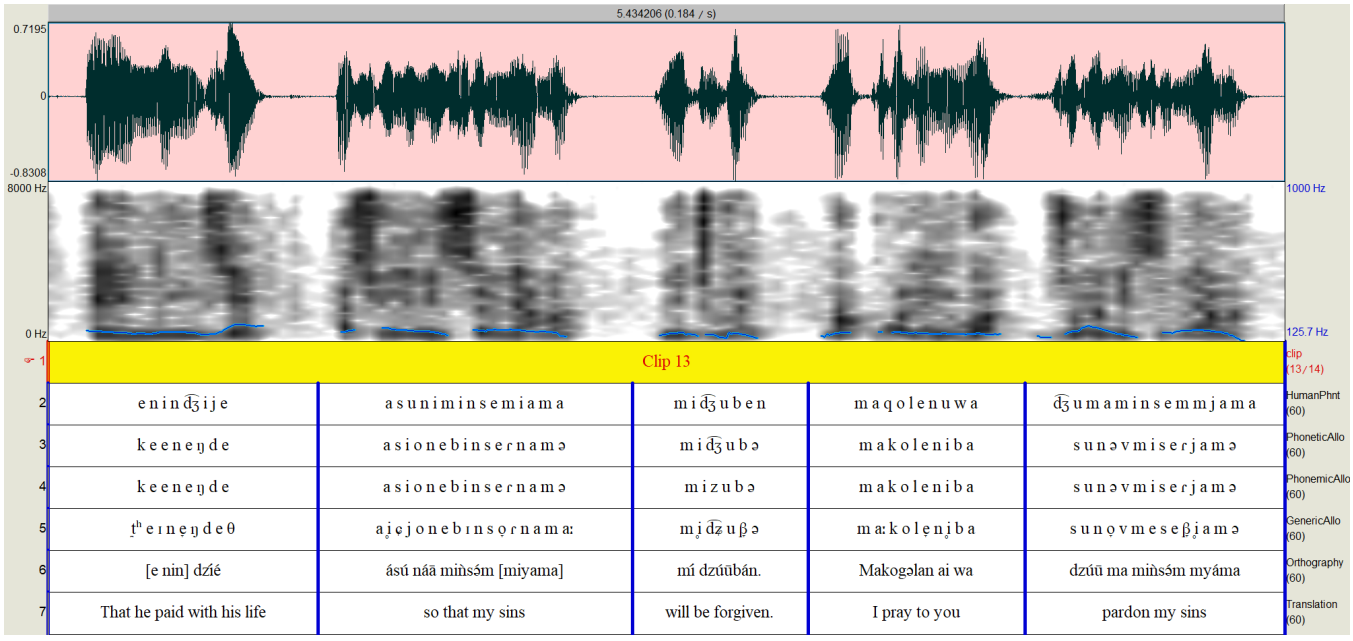

Fig. 19: Ewondo Clip 13

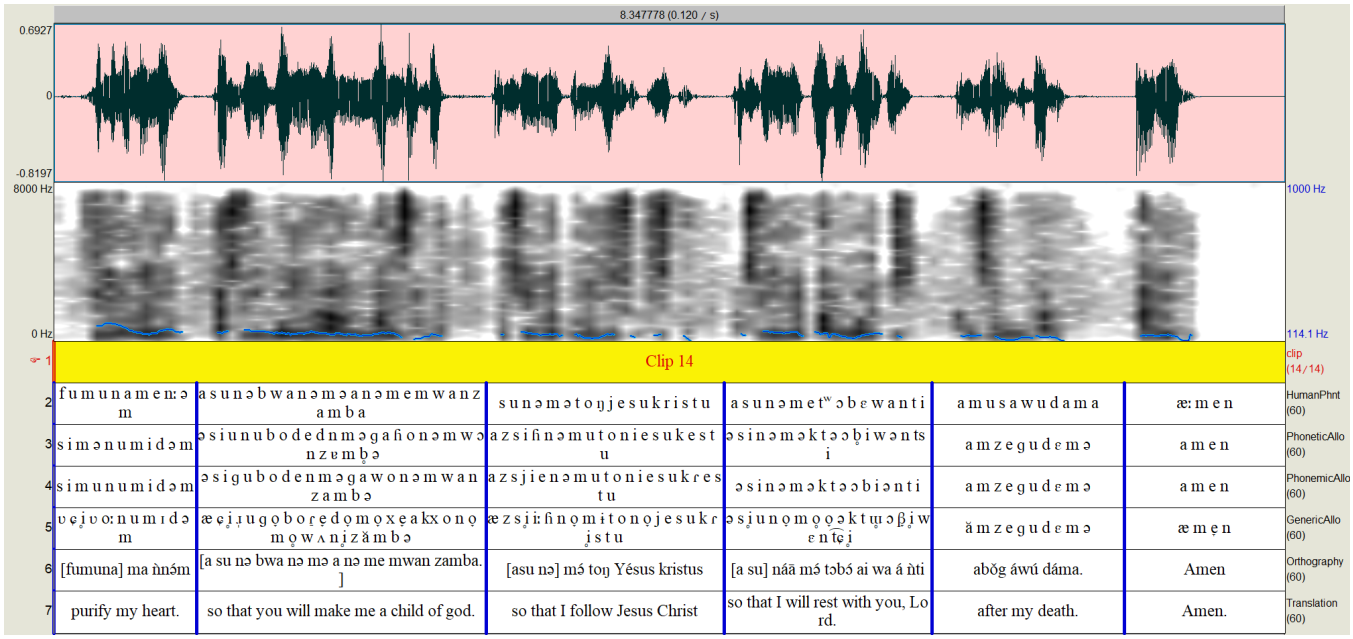

Fig. 20: Ewondo Clip 14