From Raw Data to Acoustic Analysis: A Roadmap based on Acquaviva Collecroce

Simon Gonzalez
The Australian National University
simon.gonzalez@anu.edu.au

Abstract

This paper presents a workflow framework of computational tools to be used in the process of forced alignment and analysis for endangered languages. We introduce a roadmap which uses established methodologies in the area of data processing and analysis, with a strong focus on socio-phonetic studies. The tools are organized into practical stages that can be followed systematically by researchers of under-resourced languages. We have implemented these tools in Acquaviva Collecroce, an endangered language from southern Italy and spoken by approximately 600 speakers. Alongside the tools, we also give suggestions based on our experience, which can contribute to the preservation and revitalization of endangered languages.

1 Introduction

The use of computational tools in endangered languages has proven critical for the revitalization and preservation of languages. There is an increase interest in using latest technologies to strengthen our understanding and processing of minority languages (Adams et al., 2018; Adams et al., 2020; Michaud et al., 2018; Levow, 2019; Levow et al., 2021), including speech to text (Foley et al., 2019; Michaud et al., 2018; Mitra, 2016), speech recognition (Amith et al., 2021; Foley et al., 2018; Hjortnaes et al., 2020; Matsuura et al., 2020; Shi et al., 2021; Thai et al., 2020), phonemic transcription (Adams et al., 2017; Amith and Castillo García, 2020), and forced alignment (Cavar et al., 2016; Coto-Solano, 2017; Gonzalez et al., 2018). The field of Automatic Speech Recognition (ASR) has strongly influenced this endeavor (Prud'hommeaux et al., 2021; Jimerson and Prud'hommeaux, 2018; Jimerson et al., 2018). One of the greatest contributions is that advanced technologies, which had traditionally been available only to major languages, can now be accessed by less resourced languages.

The implementation of computational techniques in language documentation has established a toolkit of skills that need to be met to access these technologies, which shows that the tasks carried out in these processes are complex in nature. These tasks are generally done by computational linguists with the required expertise, who can decide on what tools and techniques are used in any given project. In deciding what to choose, there are many options to select from, and the decision on the workflow depends on the resources available. Since there is no ultimate or perfect process, the decisions must be based on what works best, as long as the goal of language documentation is achieved. Also, given the increasing effectiveness of current algorithms developed, the documentation of endangered languages is in a crucial moment where the work done by computational linguistics can be maximized to its best potential. However, there is still more work needed to efficiently link long-established linguistic analysis traditions and advances in data processing.

Once the data is processed through computational techniques, the task is then to identify what are the best approaches for endangered languages to make the leap towards systematic analysis of the data available. One area that is a suitable test ground for this transition between computational outputs and linguistic analysis is the field of sociolinguistics. The relevance of sociolinguistics for endangered languages is that languages are better analyzed in their social context and not just as isolated entities. Sociolinguistics then helps interpret language patterns related to factors such as gender, age, ethnicity, for example. Therefore, an important contribution from computational linguists to endangered languages is to develop technologies that take computational outputs and allow researchers to analyze linguistic patterns following robust methodologies standard in the respective fields, all this, in relatively short periods of time. In this paper, we focus on technologies that are pertinent to the analysis of speech data, with a focus on socio-phonetics.

1.1 Speech Technologies and Data Size

One of the main challenges faced by languages with small amounts of speech data, is that the technologies available tend to require a minimum threshold of speech. This threshold is generally way more than what the vast majority of world languages cann afford to have. The reasoning behind this is that the more data available, the more robust the acoustic models are to accurately identify speech boundaries based on the phonetic features extracted. It does not mean that under-resourced languages cannot be processed, but rather that the results are not as reliable as those having more data available for training and testing their models. However, we argue that even smaller languages can be maximized by using all available material, and the results are still of great value for language researchers.

In this sense, computational tools used in under-resourced languages are not the means on their own, but rather they are the facilitators for quantifying speech data and identifying language patterns not available otherwise. It will then be the role of the linguist to use all the outputs and look at areas of interest, such as vowel spaces, allophonic variation, morpheme sequence occurrence, intonation, for example. In this sense, it is important to make the difference between what a computational linguist wants and what the field researcher needs. A clear example is about error accuracy. (Semi-) automatic computational models evaluate their performance based on their accuracy (or error rate). Higher accuracy is always desired, but even lower accuracy models can make a big difference in a researcher working with an under-resourced language.

1.2 Phonetic Analysis and Endangered Languages

Among the areas of linguistic interest is the acoustic/phonetic study of under-resourced languages, and forced alignment has played a crucial role in the way (and amount of data) phoneticians analyze smaller languages. The forced alignment process (See more details in Section 4) takes audio files and their corresponding time-stamped transcriptions, generally at the sentence level, and segments the data into the corresponding individual phonological segment (e.g. vowels and consonants). This tool has sped up processes that would otherwise take more time, by exponential differences. This is especially meaningful when language researchers are working against the clock in languages that unfortunately do not have much time to be analysed. Forced alignment has allowed smaller languages to be fully analyzed as it has been done in major languages. The way it works by current workflows is by taking the automatically aligned segments and extracting the relevant acoustic features, such as duration, formants, centre of gravity, to name a few. Sociophonetic research has exploited this by extracting acoustic features and finding correlations with social and geographic factors, especially in the area of vowel spaces.

2 Aim of Paper

In this paper, we combine these overlapping fields and develop an efficient roadmap that can be implemented in endangered languages with at least time-stamped orthographic transcriptions. The nature of the paper is then a hybrid one. On the one hand, it proposes a methodological approach brings together different techniques, and on the other hand, it provides resource materials that can be freely used under open-source frameworks. This roadmap includes the testing and implementation of a socio-phonetic computational workflow, from data processing to data analysis. All this is developed following best practices in the field of sociolinguistics and creates a single toolkit that can be adapted to any language.

The algorithms and instructions are placed on a GitHub repository for public use. The final output is an ordered set of code files and instructions. It is our intention to bring more systematicity and data normalization that combines the power of computational tools and linguistic analysis traditions. We believe that the implications can be many-fold. First, tools like these can shed more light into language patterns never observed before. Second, it makes data from under-resourced languages comparable with other languages, including major ones. Finally, it equips a language community to have the starting tools for more advanced technologies, such as ASR and other (semi)-automated processes. All this will contribute to the ultimate goal of this type of work: language documentation, conservation, and revitalization.

3 Methodology

3.1 Forced Alignment and Endangered Languages

Forced alignment is strongly used in endangered languages. In initial approaches, when aligning a new language, researchers ran pre-existing acoustic models from a similar language for the new language (Coto-Solano, 2017). Though effective to some extent, the main flaw of this approach is that there are always features in a language that are not accurately captured by another language acoustic model. One of the main motivations for this approach was that new languages did not have the same amount of data, thus having less accurate alignments. In this sense, data size was a limitation in the forced alignment task. Then, with the emergence of more powerful data processing techniques such as neural network and deep learning, newer approaches became more robust and more efficient at dealing with lower amount of data (McAuliffe et al., 2017), to a point, that a threshold was reached, in which the adding more data would not significantly improve the acoustic model (Fromont and Watson, 2016). This opened the door to training and aligning new languages without the need for huge amounts of data. As expected, minority and endangered languages greatly benefited from these advances (Gonzalez et al., 2018; Gupta and Boulianne, 2020; Hildebrandt, 2017).

Across time, the processes became more streamlined to such a point that forced aligning a new language from scratch is more efficient and accurate than using a pre-trained language model. If compared to ten years ago, the process is simpler but without compromising accuracy. Despite these advances, there are still many stages to simplify the process of forced alignment and its practical applications. In this paper, we propose a more succinct yet efficient workflow of data alignment and analysis. Since the paper has a methodological approach, which can be followed step by step, we present the tools and solutions in sections.

3.2 Data Selection

The first task is to identify the language to be forced aligned. A good source available for use is Pangloss (Michailovsky et al., 2014), which is an open archive created to help in the preservation of world languages, with a strong focus on endangered and minority languages. Currently, it hosts over 170 languages with more than 700 hours of recordings. An approximate of half of the audiovisual material (video and audio) has annotated files. We then chose to work with Na-Našu (Molise Slavic) (Breu, 2020), which is a micro-language with three dialects, including *Acquaviva Collecroce*. The material available for this dialect comes from a village called *Kruč*, within the province of Campobasso, in the Molise region of southern Italy (See Figure 1 for reference). The dialect has been documented by Adamou and Breu (2013) and Breu (2017).

Figure 1: Location of Kruč, where the Acquaviva Collecroce is found.

The language material available on the website was a compilation of 27 audio recordings with their corresponding transcription files. The data was recorded in 2010 by Walter Breu, and the transcriptions have three main layers of information. The first one is a time-stamped transcription at the utterance level (described in the original documentation as orthographic, representing a broad phonological transcription). This time-stamp information is the one that is relevant for the current study, because it is used to create the TextGrids explained in section **4.2**.

The second layer was a phonetic transcription of all the words, which is not used in the current study. The motivation is to use the broad phonological transcription, which forms the basis for the forced-alignment process, as explained below. The third layer available includes morphemic breakdowns. Even though these are not used in the forced-alignment process, this information is relevant for the analysis of vowels, which can help identify whether there are morpho-syntactic effects of vowel formants, for example, running a model that measures whether there are differences between vowels that appear in stems or vowels that appear in affixes. This is a good example on how forced-alignment tools can help contribute to understand phonetic/phonological features and their relationship with other features in the language. The final annotation layer included translation into Italian and German. For the purposes of this study, they were not included in any stage of the process.

3.3 Speakers

The Acquaviva Collecroce dialect is estimated to have just over 600 speakers as for 2019, according to the Italian National Institute of Statistics (ISTAT). There were over 2200 speakers at the beginning of 1950s, with sharp decreases since then due to migration. The speakers in the corpus were two females and four males, born between 1932 and 1960 (See Table 1).

Speaker	Gender	Recordings Duration (Min)	Speech Duration (Min)
GN	Male	16.3	15.6
GR	Male	10.2	9.5
PG	Female	0.7	0.5
PG	Male	3.5	2.9
PL	Male	9.9	9.6
SN	Female	13.2	11.7

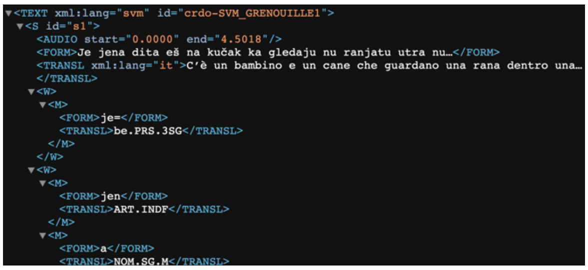
Table 1: Speakers in the corpus with their corresponding durations.

Since this is a first analysis on this dataset, we have focused on Gender to identify socio-phonetic differences. Age is another relevant factor that can be analysed to understand phonetic differences. This can be done in further stages of the research.

Speakers were recorded narrating stories, which is a good source of naturalistic data. This is a relevant characteristic in this study, since it is the type of data that is generally available for endangered languages and much suitable for socio-phonetic analyses, as compared to more controlled data such as wordlists and isolated tokens (e.g. Hay and Foulkes, 2016; Grama et al., 2020; Catherine E. Travis and Ghina, 2021).

3.4 Data format

The structure of the transcription files varies according to the format given by corpus developers. In the current case, the transcription format is available as XML files (See Figure 2 for reference). In the original recordings there were at least two speakers per file: one interviewer and a speaker, but the transcriptions provided included the transcription for the speakers only.



```
▼<TEXT xml:lang="svm" id="crdo-SVM_GRENOUILLE1">
  ▼<S id="s1">
     <AUDIO start="0.0000" end="4.5018"/>
     <FORM>Je jena dita eš na kučak ka gledaju nu ranjatu utra nu…</FORM>
     <TRANSL xml:lang="it">C'è un bambino e un cane che guardano una rana dentro una…
     </TRANSL>
    ▼<W>
      ▼<M>
         <FORM>je=</FORM>
         <TRANSL>be.PRS.3SG</TRANSL>
       </M>
     </W>
    ▼<W>
      ▼<M>
         <FORM>jen</FORM>
         <TRANSL>ART.INDF</TRANSL>
       </M>
      ▼<M>
         <FORM>a</FORM>
         <TRANSL>NOM.SG.M</TRANSL>
```

Figure 2: XML file from the source file.

The transcription files were processed in R, using a script developed by the main author. The script first identifies the sentence ID (`<S id="s1">` in Figure 2), under which three dependent sections are extracted: the start time, end time (`<AUDIO start="0.0000" end="4.5018"/>`), and the transcribed sentence (`<FORM>Je jena dita eš na kučak ka gledaju nu ranjatu utra nu…</FORM>`). The audio files were available in MP3 format, sampled with 44.1 kHz. They had different durations, with the shortest file being 38 seconds and the largest 7.5 minutes, and the mean duration being 2 minutes in length.

4 Forced-Alignment Process

The forced-alignment process involves four main stages, presented in Figure 3. Each stage is expanded in the section below. One important observation for these stages is that investing time in the pre-processing of the files would ensure better outputs and dealing with less bugs in future stages. We present some recommendations in each section.

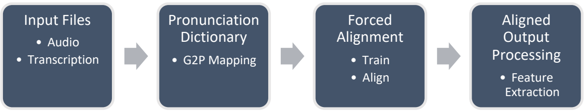


Figure 3: Main stages in the forced alignment process.

4.1 Pronunciation Dictionary from Input Files

First, a pronunciation dictionary must be created. In some approaches, these dictionaries are created from a lexicon file available for the language. However, for languages without curated lexicon files, pronunciation dictionaries can be created from the orthographic transcriptions. In this study, we took the available raw transcription of the data and then tokenized the transcriptions to have unique individual words.

These are then used to create the g2p (grapheme to phoneme) mapping. The amount of processing for creating this dictionary varies from language to language. For example, in Spanish there is a closer letter to phoneme mapping, where there is an almost full mapping between orthographic letters and phonemes, except for silent ‘h’ and digraphs (‘ll’, ‘ch’) (Gonzalez, 2022). This is different from English, where the mapping cannot always follow the orthographic spelling. As an example, the orthographic letter ‘a’ can have different phonemic representations, e.g. /eɪ/, /ə/, /a:/. The latter case would present a more challenging task for the mapping. For the case of Acquaviva Collecroce, the transcriptions done by the original creators was a broad phonological representation. This facilitated the g2p task and we decided to split words into individual letters, which are then considered the phonemes for each entry, as shown in Figure 4 below.

```
1 baliže   b a l i ž e
2 balun    b a l u n
3 balunič b a l u n i č
4 baluniča     b a l u n i č a
5 bane     b a n e
6 banu     b a n u
7 baratol b a r a t o l
8 baštunam     b a š t u n a m
9 baštunič     b a š t u n i č
```

Figure 4: Sample entries for the pronunciation dictionary.

In this case, the g2p mapping had a one-to-one correspondence. However, this is not always that case. In cases where no such correspondence exists in the transcription file, as in the Spanish example, the recommendation is to assign a phonemic symbol that does not overlap with other symbols. This must be done a priori before creating the dictionary so in the final product each grapheme or grapheme sequence is accounted for.

4.2 Transcriptions in TextGrid Format

The first processing of the text involves text normalization, which includes identifying spelling mistakes, non-speech annotations (e.g. notes from the transcribers, alternative pronunciations, etc.). This ensures that all entries can be mapped to the same word and not having multiple forms for the same entry. Another step here is to identify whether there are special characters that should not be included in the text, such as parenthesis, brackets, and slashes. Once the text has been normalized, the next step is to convert the text into a time-stamped file, since available forced aligners read transcriptions with time-stamped formats.

An R script was developed to create transcription files in TextGrid files, a format used in Praat (Boersma and Weenink, 2022). This format is widely used in linguistics, with strong emphasis for acoustic phonetic analysis. TextGrids are files containing time-stamped texts. The content is divided into tiers, where the text can be split into smaller sections with their respective boundaries. This is very useful when researchers need to break the content into different categories, such as identifying different speakers or annotating different linguistic layers, such as words, segments, features, for example. A sample TextGrid file from our data is shown in Figure 5, together with its corresponding audio file represented in the waveform above.

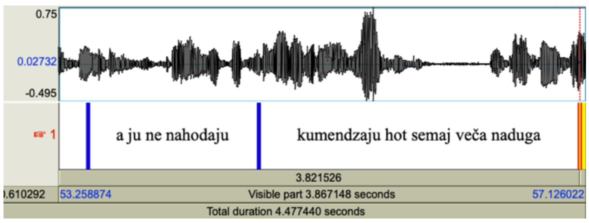


Figure 5: Sample TextGrid and audio files, with the transcription tier.

The figure shows the transcription for one speaker. The blue lines represent the time boundaries which correlate with the time information from the audio file. Based on our experience, we have identified that the size of the intervals has an impact on the output of the forced aligned file.

Since aligners analyze the acoustic signal as linear in time, if there are alignment errors at the beginning of an interval, they will likely roll the error over the following segment boundaries in the same interval. For example, if the aligner marks the beginning of a stop sound earlier than the actual start (e.g., due to a spike in the acoustic signal caused by a cough or a mouse click), then this will also influence where the boundaries of the following segments are placed. If the error is at the start of a long interval, then it will most likely render the full interval inaccurate. However, if the error takes place at the beginning of a shorter interval, less data will be compromised, because the acoustic mapping restarts at the beginning of each interval. Thus, we recommend the intervals are closely mapped with natural pauses and speech boundaries. This will also facilitate the mapping of words into natural speech units.

4.3 Running the Forced Alignment

Once we have prepared the pronunciation dictionary and transcription files with the corresponding audio files, the next step is to run the forced aligner. Previous studies have shown that the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), based on Kaldi (Povey et al., 2011), is one of the most accurate aligners currently available, especially used in sociophonetic studies (Gonzalez et al., 2020). We used the MFA following the instructions from the source website https://montreal-forced-aligner.readthedocs.io/en/latest/. The main challenge here is to have the correct setup to ensure that the aligner runs though the data without any bugs. For this, it is recommended to have all audio files in the same format, including, bit rate, sampling rate, and following good labelling practice for the files (which is mainly relevant for feature extraction in future stages).

4.4 Forced alignment outputs

MFA provides the aligned outputs as TextGrid files with two tiers, one for the forced-aligned words and another for the forced-aligned phonemic segments. We have found it efficient to recombine this output with the original input in the same TextGrid to include the utterance-level transcription. This is especially important when examining features such as intonation and prosodic patterns, where whole utterances would be relevant for analysis and not just words and segments on their own. The output would then be as shown in Figure 6 below.

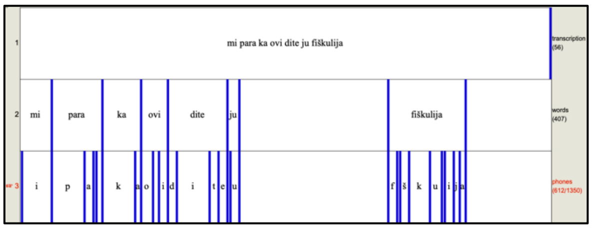


Figure 6: TextGrid with combined tiers: original transcription (Tier 1), and aligned words (Tier 2) and phonemes (Tier 3).

As with any automatic process, a sanity check is always important to assess the accuracy of the outputs. Previous studies have identified that the errors can be systematic, with some phonological contexts being more susceptible for more inaccuracies (Gonzalez et al., 2020). In this case, we propose an initial assessment where duration can be used to look at errors. This is based on durational differences, where outliers, too long or too short, can be considered errors in the alignment. It is also common practice in cases where there are enough resources to manually check a proportion of the outputs by trained phoneticians.

5 Data Wrangling (Data Processing)

In this stage, we gather all the data from the TextGrids, which also prepares them for the extraction of acoustic and phonetic features. This process is done in R (R Core Team, 2022), using a combination of libraries such as rPraat (Boril and Skarnitzl, 2016), dplyr (Wickham et al., 2022), tidyr (Wickham and Girlich, 2022), for example. The main frequency counts from the forced aligned outputs are shown in Table 2.

Speaker	Gender	Consonants	Vowels	Words
GN	Male	4298	3573	2012
GR	Male	3175	2588	1487
PG	Female	204	159	103
PG	Male	491	389	252
PL	Male	3569	2920	1668
SN	Female	2866	2449	1483
Total		**14603**	**12078**	**7005**

Table 2: Main frequency Counts from forced aligned outputs.

We extract all the information from the three tiers: utterance, word, and phoneme. This process takes phoneme labels, start and end time information, and phonological contexts (previous and following segments). Then, the same type of information is extracted for words and utterances. The final product is a full description of each phoneme with its environments, phonetic, phonemic, and lexical, as shown in Figure 7 below.

Speaker	Gender	Previous	Segment	Following	Duration	PrevWord	Word	FollWord	WordDur
GN	M	l	e	d	0.12	ka	gledaju	nu	0.52
GN	M	r	a	n	0.17	nu	ranjatu	utra	0.6
GR	M	r	i	v	0.09	je	riva	prije	0.17
GR	M	t	u	c	0.04	je	tuculala	di	0.44
PG	F	v	i	d	0.08	bi	vidila	ka	0.32
PG	F	d	i	p	0.1	nonda	di	parket	0.18

Figure 7: Sample output after data wrangling.

5.1 Acoustic Features

Acoustic features are a crucial component in socio-phonetic studies. There is a wide range of acoustic features that can be used, and here we focus on three, namely, Intensity (used in prosody), Pitch (prosody and tonality), and Formants (vowels and sonorant consonants). These features cover a wide range of areas of interest. We use Praat as the main program for extracting the acoustic values, taking as input the time-specified data wrangled in the previous stage.

For the acoustic information to be extracted, the first step is to convert each audio file into a formant file in Praat. From this file, we can then extract information from the F1 and F2 for vowel analysis. Based on some experimentation, we have identified that combining R and Praat can streamline the process more efficiently, by using each program to their best capacity. For example, R is very efficient at data wrangling and analysis, but Praat cannot efficiently dealt with the level of wrangling and dataset processing as in R, especially when dealing with multiple file formats. On the other hand, Praat is much more efficient at acoustic processing and querying phonetic features as compared to R. This is why we do the data wrangling in R and the feature extraction in Praat. We then do the data analysis in R again once all the necessary information has been collected from the audio, formant, and TextGrid files.

5.2 Populating Data from Praat

Once this step is finished, we have a fully annotated dataset with individual features and their corresponding acoustic features. This functions as the main data hub from which various analyses can be carried out from the dataset. In the following stages, we present the steps for processing vowels and prepare them for acoustic analysis (See Figure 8).

Segment	Speaker	Time	formant_1	pitchValue	intensityValue	mfcc_5
e	GN_M_1	0.324	435.644779	143.2858523	72.95193861	146.0820365
a	GN_M_1	0.678	540.9370843	128.096409	74.44932052	155.946977
i	GN_M_1	1.11	335.1292519	153.6110352	71.98936304	170.2450033
a	GR_M_1	52.1712	486.1003805	139.0686298	80.94875149	124.2078462
e	GR_M_1	52.2932	486.8311409	132.6412425	80.45326531	95.35667664
o	GR_M_1	52.3532	397.4000227	141.6490183	84.11465582	102.6605409
u	PG_M_1	10.0255	337.7970078	159.3213093	76.3552603	-100.4482523
i	PG_M_1	10.8055	639.8518027	136.3856407	68.11278488	-107.7933192
u	PG_M_1	10.8355	385.7087415	133.8610273	74.54216515	-118.1969266

Figure 8: Sample output after feature extraction.

5.3 Vowel Analysis and Visualization

Identifying Vowels in the Dataset: The analysis of vowels must account for important differences in each speaker's vocal tract. To have interpretable and robust comparisons, there must be a process of normalization techniques that give more credibility to analysis. In this study, we apply vowel normalization based on the Lobanov (Lobanov, 1971) technique. This allows the analysis of both static and dynamic measurements to be compared across speakers. Again, this gives researchers of endangered languages quick access to the vocalic spaces in the data. In this process, we use the vowel package (Kendall and Thomas, 2018) for vowel normalization and ggplot2 (Wickham, 2016) for data visualization.

Visualization and Analysis: The visualization gives importance access to vowel behaviors in the data, and this can be split into the sociolinguistic factors available, in this case, Gender. Figure 9 shows the vowel duration of a selection of five landmark vowels and their differences based on Gender. The data indicates that there is an increasing mean duration starting from /a/, then /e/, /i/ and /u/, ending in /o/, which is the longest vowel. The mean durations are similar for both Genders, but with more distinctions for /o/ and /u/. Further statistical differences can reveal whether ther are significant differences based on phonological contexts.

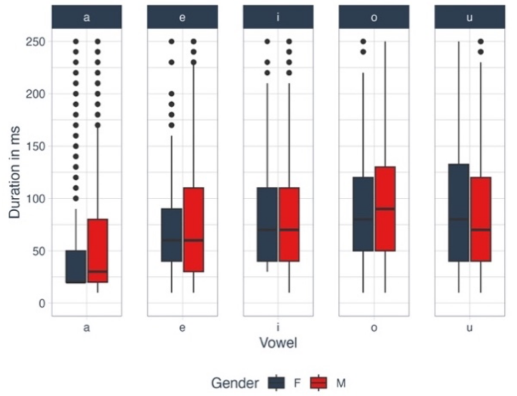


Figure 9: Vowel durations and Gender Differences.

Different from duration analysis, vocalic space analysis reveals important differences for Genders in Figure 10. First, the selection of the five vowels shows a different picture from the location of /u/, as compared to other languages such as French, English and Spanish, where the /u/ is higher and more retracted. In terms of the spread, it shows that Males are producing more compressed vowels than Females, especially for the Front non-Low vowels /i/ and /e/. Mean durations, represented by point size, shows that the main durational differences are observed for /a/. This is an indication that if there is a first potential area to examine socio-phonetic differences would be the formant and duration differences between Males and Females.

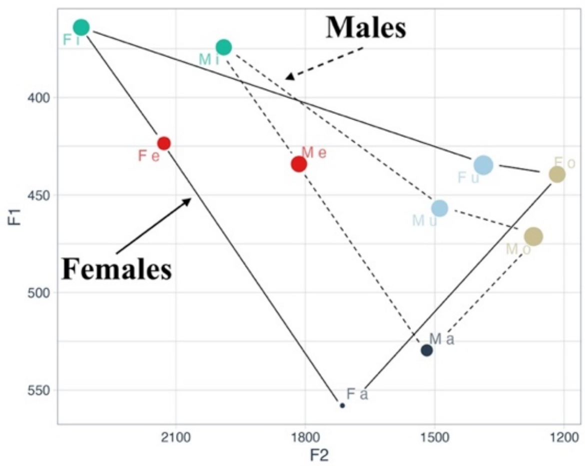


Figure 10: Vowel space for Males and Females from normalized formant values. Vowel size represents mean durations.

5.4 Assessing Consonantal Analysis

For the consonant analysis, we look at duration differences for the Coronal fricatives /s, z/ (alveolar) and /š, ž/ (Post-Alveolar), split by Gender. Two main observations can be drawn from Figure 11 below. First, durations are similar, but with Post-Alveolars having wider spread than Alveolars. Second, Females are producing mean larger durations than Males, except for /ž/. This indicates that the differences for these consonants are likely more based on Gender differences rather than phonological factors, a question than can be further studied with in-depth analysis.

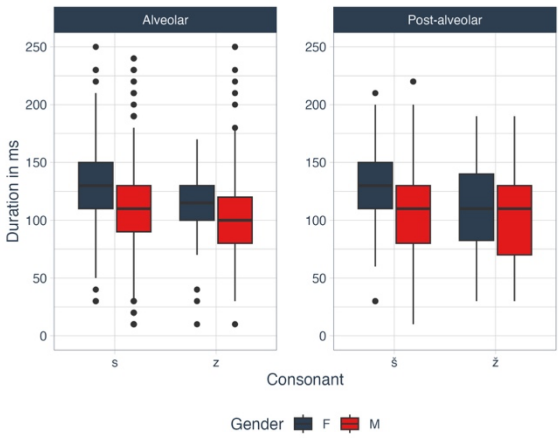


Figure 11: Coronal Fricative Duration Differences across Place of Articulation and Gender.

5.5 Assessing Prosodic Features

Finally, we look at pith as a suprasegmental feature. Figure 12 shows the pitch tracks for a section of the recording of speaker GN Male. There are six main utterances with their intonations shown in the blue lines. The arrows in each number represent the trajectory of the intonation, with all having a falling pattern, except from 5 having a slight rising pattern. These intonation patterns can further be examined with the output and prepared data.

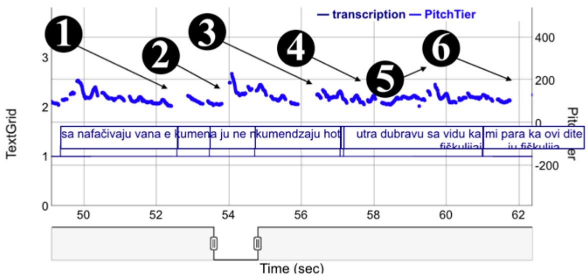


Figure 12: Pitch tracks used to identify intonation patterns in the language.

6 Discussion

This paper presents a roadmap of tools, from data processing to socio-phonetic analysis. We have taken Acquaviva Collecroce, an endangered language and whose data can be freely accessible.

This work has put together a range of computational tools and packages that can facilitate data processing and analysis in a simple, yet efficient way. Table 3 shows a summary of the tools. It is not our intention to present an ultimate workflow, but rather a practical toolkit that allows users to implement it in endangered language studies. The resource materials are open source and can be adapted an expanded to the required needs of the users.

Oder	Stage	Program	Description
1	Pre-Processing	R	Data gathering
2	Pre-Processing	Praat	TextGrid creation
3	Alignment	Python	Running the forced alignment
4	Post-Processing	R	Wrangling outputs
5	Acoustic Features	Praat	Extracting phonetic features
6	Analysis	R	Data visualization and main analyses

Table 3: Main stages of the workflow, with the corresponding program languages.

7 Conclusions

The field of computational linguistics is making invaluable contributions to the perseveration and revitalization of endangered languages. In this paper, we have a presented a set of relevant computational tools developed to help researchers from forced alignment to acoustic phonetic studies, including segmental and suprasegmental analysis. We have developed the tools for an endangered language, Acquaviva Collecroce, which is a practical example of the power and applicability of the tools presented here.

8 Future Work

Our future work will include an online application where these steps are streamlined and automated from user inputs to visualizing results and carrying out linguistic analysis. This is work in progress and we hope this contributes to the technologies developed to help endangered languages globally.

References

Oliver Adams, Trevor Cohn, Graham Neubig, & Alexis Michaud. 2017. Phonemic transcription of low-resource tonal languages. In *Proceedings of the Australasian Language Technology Association Workshop*, Brisbane, 6–8 December, 53–60. (https://www.aclweb.org/anthology/U17-1006/)

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, et al. 2020. User-friendly automatic transcription of lowresource languages: Plugging ESPnet into Elpis. In *ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 51–62.

Jonathan D. Amith, Jiatong Shi and Rey Castillo García. 2021. End-to-End Automatic Speech Recognition: Its Impact on the Workflow for Documenting Yoloxóchitl Mixtec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80 June 11, 2021. ©2021 Association for Computational Linguistics

Jonathan D. Amith and Rey Castillo García. 2020. *Audio corpus of Yoloxóchitl Mixtec with accompanying time-coded transcriptons in ELAN.* http://www.openslr.org/89/. Accessed: 2021-03-05.

Evangelia Adamou and Walter Breu. 2013. Présentation du programme EUROSLAV 2010. Base de données électronique de variétés slaves menacées dans des pays européens non slavophones. In: S. Kempgen, M. Wingender, N. Franz, M. Jakiša (eds.), Deutsche Beiträge zum 15. *Internationalen Slavistenkongress Minsk* 2013. München, 13–23.

Paul Boersma and David Weenink. 2022. *Praat: doing phonetics by computer* [Computer program]. Version 6.3, retrieved 15 November 2022 from http://www.praat.org/.

Tomas Boril and Radek Skarnitzl. 2016. Tools rPraat and mPraat. In Sojka P, Horák A, Kopeček I, Pala K (eds.), *Text, Speech, and Dialogue: 19th International Conference, TSD* 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings, 367–374. ISBN 978-3-319-45510-5, doi:10.1007/978-3-319-45510-5_42, http://dx.doi.org/10.1007/978-3-319-45510-5_42.

Walter Breu. 2020, online 1. “Molise Slavic”, in: *Encyclopedia of Slavic Languages and Linguistics Online*, Editor-in-Chief Marc L. Greenberg. Consulted online on 20 August 2020. First published online: 2020 http://dx.doi.org/10.1163/2589-6229_ESLO_COM_034736

Walter Breu. 2017. Slavische Mikrosprachen im absoluten Sprachkontakt. Band I. Moliseslavische Texte aus Acquaviva Collecroce, Montemitro und San Felice del Molise. *Glossierte und interpretierte Sprachaufnahmen aus Italien, Deutschland, Österreich und Griechenland*. Wiesbaden.

Malgorzata Cavar, Damir ´ Cavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4004–4011.

Rolando Coto-Solano and Sofía Flores Solórzano. 2017. Comparison of two forced alignment systems for aligning Bribri speech. *CLEI Electronic Journal* 20(1). 2:1–2:13.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, E Mark, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Ben Foley, Alina Rakhi, Nicholas Lambourne, Nicholas Buckeridge, and Janet Wiles. 2019. ELPIS: An accessible speech-to-text tool. *Proc. Interspeech 2019*, pages 4624–4625.

Robert Fromont and Kevin Watson. 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. Corpora 11(3). 401–431.

Simon Gonzalez, Catherine Travis, James Grama, Danielle Barth, and Sunkulp Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. 145-148. In *17th Australasian International Conference on Speech Science and Technology*.

Simon Gonzalez, James Grama, and Catherine E. Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, vol. 6, no. 1, 2020, pp. 20190058. https://doi.org/10.1515/lingvan-2019-0058

Simon Gonzalez. 2022. The Development of a Comprehensive Spanish Dictionary for Phonetic and Lexical Tagging in Socio-phonetic Research (ESPADA). In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022*, pages 8–14, Marseille, 24 June 2022

James Grama, Catherine E. Travis, and Simon Gonzalez. 2020. Ethnolectal and community change ov(er) time: Word-final (er) in Australian English. *Australian Journal of Linguistics*, vol 40, no. 3, pp. 346-368. https://doi.org/10.1080/07268602.2020.1823818

Vishwa Gupta and Gilles Boulianne. 2020. Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367.

Jen Hay and Paul Foulkes. 2016. The evolution of medial /t/ over real and remembered time. *Language* 92(2): 298-330. http://dx.doi.org/10.1353/lan.2016.0036.

Kristine A. Hildebrandt, Carmen Jany, and Wilson Silva. 2017. Documenting variation in endangered languages. *Language Documentation & Conservation Special Publication 14*. University of Hawai’i Press.

Nils Hjortnaes, Niko Partanen, Michael Rießler, and Francis M Tyers. 2020. Towards a speech recognizer for Komi: An endangered and low-resource Uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37.

Robbie Jimerson and Emily Prud’hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Robert Jimerson, Kruthika Simha, Ray Ptucha, and Emily Prud’hommeaux. 2018. Improving ASR output for endangered language documentation. In *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Tyler Kendall and Erik R. Thomas. 2018. *vowels: Vowel Manipulation, Normalization, and Plotting.* http://blogs.uoregon.edu/vowels/.

B. M. Lobanov. 1971. Classification of Russian vowels spoken by different listeners. *Journal of the Acoustical Society of America* 49:606-08.

Gina-Anne Levow. 2019. Promoting Language Technology for Endangered Languages with Shared Tasks. In *Proceedings of the Language Technologies for All (LT4All),* pages 116–119, Paris, UNESCO Headquarters, 5-6 December, 2019.

Gina-Anne Levow, Emily P. Ahn, Emily M. Bender. 2021. Developing a Shared Task for Speech Processing on Endangered Languages. In

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages, 1:96–106, Online, March 2–3, 2021.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020. Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2622–2628.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *INTERSPEECH-2017*, Stockholm Sweden, 498–502.

Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation & Conservation*, 8:119–135.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages: The case of Yoloxóchitl Mixtec (Mexico). In *Proc. Interspeech 2016*, pages 3076–3080.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on Automatic Speech Recognition and Understanding, CONF*, pages 1–4. IEEE Signal Processing Society

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation,* 15:491-513

R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yoloxóchitl Mixtec. *arXiv* preprint arXiv:2101.10877.

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully convolutional ASR for less-resourced endangered languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 126–130.

Catherine E. Travis and Inas Ghina. 2021. Gender, mobility and contact: Stability and change in an Acehnese dialect. *Asia-Pacific Language Variation*. 7(2): 142-167. https://doi.org/10.1075/aplv.20007.tra

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Hadley Wickham and Maximilian Girlich. 2022. *tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. 2022. *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.