Aikuma: A Mobile App for Collaborative Language Documentation

Steven Bird[1,2], **Florian R. Hanke**[1], **Oliver Adams**[1], **and Haejoong Lee**[2]
[1]Dept of Computing and Information Systems, University of Melbourne
[2]Linguistic Data Consortium, University of Pennsylvania

Abstract

Proliferating smartphones and mobile software offer linguists a scalable, networked recording device. This paper describes *Aikuma*, a mobile app that is designed to put the key language documentation tasks of recording, respeaking, and translating in the hands of a speech community. After motivating the approach we describe the system and briefly report on its use in field tests.

1 Introduction

The core of a language documentation consists of primary recordings along with transcriptions and translations (Himmelmann, 1998; Woodbury, 2003). Many members of a linguistic community may contribute to a language documentation, playing roles that depend upon their linguistic competencies. For instance, the best person to provide a text could be a monolingual elder, while the best person to translate it could be a younger bilingual speaker. Someone else again may be the best choice for performing transcription work. Whatever the workflow and degree of collaboration, there is always the need to manage files and create secondary materials, a data management problem. The problem is amplified by the usual problems that attend linguistic fieldwork: limited human resources, limited communication, and limited bandwidth.

The problem is not to collect large quantities of primary audio in the field using mobile devices (de Vries et al., 2014). Rather, the problem is to ensure the long-term interpretability of the collected recordings. At the most fundamental level, we want to know what words were spoken, and what they meant. Recordings made in the wild suffer from the expected range of problems: far-field recording, significant ambient noise, audience participation, and so forth. We address these problems via the "respeaking" task (Woodbury, 2003). Recordings made in an endangered language may not be interpretable once the language falls out of use. We address this problem via the "oral translation" task. The result is relatively clean source audio recordings with phrase-aligned translations (see Figure 1). NLP methods are applicable to such data (Dredze et al., 2010), and we can hope that ultimately, researchers working on archived bilingual audio sources will be able to automatically extract word-glossed interlinear text.

We describe *Aikuma*, an open source Android app that supports recording along with respeaking

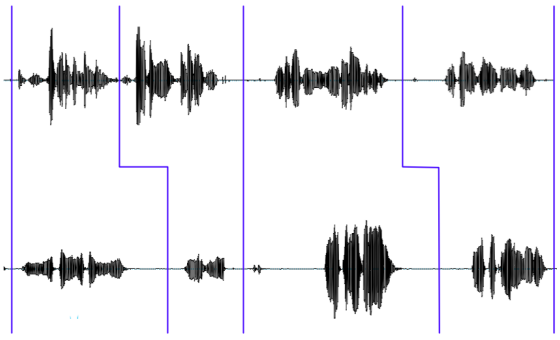


Figure 1: Phrase-aligned bilingual audio

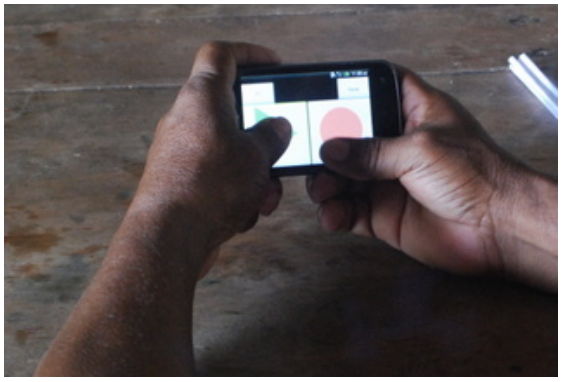


Figure 2: Adding a time-aligned translation

Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 1–5,
Baltimore, Maryland, USA, 26 June 2014.

and oral translation, while capturing basic metadata. Aikuma supports local networking so that a set of mobile phones can be synchronized, and anyone can listen to and annotate the recordings made by others. Key functionality is provided via a text-less interface (Figure 2). Aikuma introduces social media and networked collaboration to village-based fieldwork, all on low-cost devices, and this is a boon for scaling up the quantity of documentary material that can be collected and processed. Field trials in Papua New Guinea, Brazil, and Nepal have demonstrated the effectiveness of the approach (Bird et al., 2014).

2 Thought Experiment: The Future Philologist

A typical language documentation project is resource-bound. So much documentation could be collected, yet the required human resources to process it all adequately are often not available. For instance, some have argued that it is not effective to collect large quantities of primary recordings because there is not the time to transcribe it.[1]

Estimates differ about the pace of language loss. Yet it is uncontroversial that – for hundreds of languages – only the oldest living speakers are well-versed in traditional folklore. While a given language may survive for several more decades, the opportunity to document significant genres may pass much sooner. Ideally, a large quantity of these nearly-extinct genres would be recorded and given sufficient further treatment in the form of respeakings and oral translations, in order to have archival value. Accordingly, we would like to determine what documentary materials would be of greatest practical value to the linguist working in the future, possibly ten to a hundred or more years in future. Given the interest of classical philology in ancient languages, we think of this researcher as the "future philologist."

Our starting point is texts, as the least processed item of the so-called "Boasian trilogy." A substantial text corpus can serve as the basis for the preparation of grammars and dictionaries even once a language is extinct, as we know from the cases of the extinct languages of the Ancient Near East.

[1]E.g. Paul Newman's 2013 seminar *The Law of Unintended Consequences: How the Endangered Languages Movement Undermines Field Linguistics as a Scientific Enterprise*, `https://www.youtube.com/watch?v=xziE08ozQok`

Our primary resource is the native speaker community, both those living in the ancestral homeland and the members of the diaspora. How can we engage these communities in the tasks of recording, respeaking, and oral interpretation, in order to generate the substantial quantity of archival documentation?

Respeaking involves listening to an original recording and repeating what was heard carefully and slowly, in a quiet recording environment It gives archival value to recordings that were made "in the wild" on low-quality devices, with background noise, and by people having no training in linguistics. It provides much clearer audio content, facilitating transcription. Bettinson (2013) has shown that human transcribers, without knowledge of the language under study, can generally produce phonetic transcriptions from such recordings that are close enough to enable someone who knows the language to understand what was said, and which can be used as the basis for phonetic analysis. This means we can postpone the transcription task – by years or even decades – until such time as the required linguistic expertise is available to work with archived recordings.

By interpretation, we mean listening to a recording and producing a spoken translation of what was heard. Translation into another language obviates the need for the usual resource-intensive approaches to linguistic analysis that require syntactic treebanks along with semantic annotations, at the cost of a future decipherment effort (Xia and Lewis, 2007; Abney and Bird, 2010).

3 Design Principles

Several considerations informed the design of Aikuma. First, to facilitate use by monolingual speakers, the primary recording functions need to be text free.

Second, to facilitate collaboration and guard against loss of phones, it needs to be possible to continuously synchronise files between phones. Once any information has been captured on a phone, it is synchronized to the other phones on the local network. All content from any phone is available from any phone, and thus only a single phone needs to make it back from village-based work. After a recording is made, it needs to be possible to listen to it on the other phones on the local network. This makes it easy for people to annotate each other's recordings. This also en-

ables users to experience the dissemination of their recordings, and to understand that a private activity of recording a narrative is tantamount to public speaking. This is useful for establishing informed consent in communities who have no previous experience of the Internet or digital archiving.

Third, to facilitate trouble-shooting and future digital archaeology, the file format of phones needs to be transparent. We have devised an easily-interpretable directory hierarchy for recordings and users, which permits direct manipulation of recordings. For instance, all the metadata and recordings that involve a particular speaker could be extracted from the hierarchy with a single filename pattern.

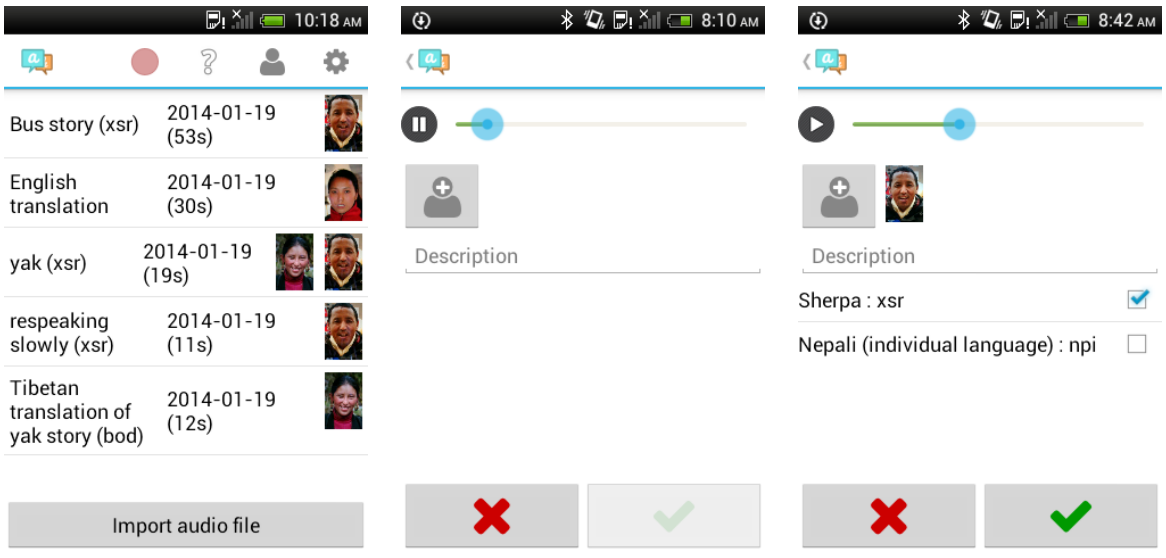
4 Aikuma

Thanks to proliferating smartphones, it is now relatively easy and cheap for untrained people to collect and share many sorts of recordings, for their own benefit and also for the benefit of language preservation efforts. These include oral histories, oral literature, public speaking, and discussion of popular culture. With inexpensive equipment and minimal training, a few dozen motivated people can create a hundred hours of recorded speech (approx 1M words) in a few weeks. However, adding transcription and translation by a trained linguist introduces a bottleneck: most languages will be gone before linguists will get to them.

Aikuma puts this work in the hands of language speakers. It collects recordings, respeakings, and interpretations, and organizes them for later synchronization with the cloud and archival storage. People with limited formal education and no prior experience using smartphones can readily use the app to record their stories, or listen to other people's stories to respeak or interpret them. Literate users can supply written transcriptions and translations. Items can be rated by the linguist and language workers and highly rated items are displayed more prominently, and this may be used to influence the documentary workflow. Recordings are stored alongside a wealth of metadata, including language, GPS coordinates, speaker, and offsets on time-aligned translations and comments.

4.1 Listing and saving recordings

When the app is first started, it shows a list of available recordings, indicating whether they are respeakings or translations (Figure 3(a)). These recordings could have been made on this phone, or synced to this phone from another, or downloaded from an archive. The recording functionality is accessed by pressing the red circle, and when the user is finished, s/he is prompted to add metadata to identify the person or people who were recorded (Figure 3(b)) and the language(s) of the recording (Figure 3(c)).



(a) Main list (b) Adding speaker metadata (c) Adding language metadata

Figure 3: Screens for listing and saving recordings

4.2 Playback and commentary

When a recording is selected, the user sees a display for the individual recording, with its name, date, duration, and images of the participants, cf. Figure 4.

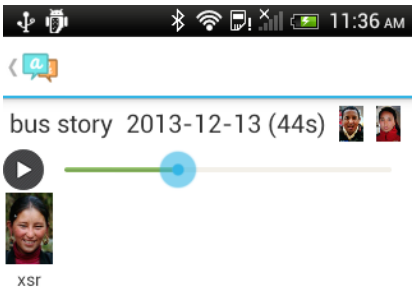


Figure 4: Recording playback screen

The availability of commentaries is indicated by user images beneath the timeline. Once an original recording has commentaries, their locations are displayed within the playback slider. Playback interleaves the original recording with the spoken commentary, cf Figure 5.

Figure 5: Commentary playback screen

4.3 Gesture vs voice activation

Aikuma provides two ways to control any recording activity, using gesture or voice activation. In the gesture-activated mode, playback is started, paused, or stopped using on-screen buttons. For commentary, the user presses and holds the play button to listen to the source, and presses and holds the record button to supply a commentary, cf Figure 2. Activity is suspended when neither button is being pressed.

In the voice-activated mode, the user puts the phone to his or her ear and playback begins automatically. Playback is paused when the user lifts the phone away from the ear. When the user speaks, playback stops and the speech is recorded and aligned with the source recording.

4.4 File storage

The app supports importing of external audio files, so that existing recordings can be put through the respeaking and oral translation processes. Storage uses a hierarchical file structure and plain text metadata formats which can be easily accessed directly using command-line tools. Files are shared using FTP. Transcripts are stored using the plain text NIST HUB-4 transcription format and can be exported in Elan format.

4.5 Transcription

Aikuma incorporates a webserver and clients can connect using the phone's WiFi, Bluetooth, or USB interfaces. The app provides a browser-based transcription tool that displays the waveform for a recording along with the spoken annotations. Users listen to the source recording along with any available respeakings and oral translations, and then segment the audio and enter his or her own written transcription and translation. These are saved to the phone's storage and displayed on the phone during audio playback.

5 Deployment

We have tested Aikuma in Papua New Guinea, Brazil, and Nepal (Bird et al., 2014). We taught members of remote indigenous communities to record narratives and orally interpret them into a language of wider communication. We collected approximately 10 hours of audio, equivalent to 100k words. We found that the networking capability facilitated the contribution of multiple members of the community who have a variety of linguistic aptitudes. We demonstrated that the platform is an effective way to engage remote indigenous speech communities in the task of building phrase-aligned bilingual speech corpora. To support large scale deployment, we are adding support for workflow management, plus interfaces to the Internet Archive and to SoundCloud for long term preservation and social interaction.

Acknowledgments

We gratefully acknowledge support from the Australian Research Council, the National Science Foundation, and the Swiss National Science Foundation. We are also grateful to Isaac McAlister, Katie Gelbart, and Lauren Gawne for field-testing work. Aikuma development is hosted on GitHub.

References

Steven Abney and Steven Bird. 2010. The Human Language Project: building a universal corpus of the world's languages. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, pages 88–97. Association for Computational Linguistics.

Mat Bettinson. 2013. The effect of respeaking on transcription accuracy. Honours Thesis, Dept of Linguistics, University of Melbourne.

Steven Bird, Isaac McAlister, Katie Gelbart, and Lauren Gawne. 2014. Collecting bilingual audio in remote indigenous villages. under review.

Nic de Vries, Marelie Davel, Jaco Badenhorst, Willem Basson, Febe de Wet, Etienne Barnard, and Alta de Waal. 2014. A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, 56:119–131.

Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church. 2010. NLP on spoken documents without ASR. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 460–470. Association for Computational Linguistics.

Florian R. Hanke and Steven Bird. 2013. Large-scale text collection for unwritten languages. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1134–1138. Asian Federation of Natural Language Processing.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Anthony C. Woodbury. 2003. Defining documentary linguistics. In Peter Austin, editor, *Language Documentation and Description*, volume 1, pages 35–51. London: SOAS.

Fei Xia and William D. Lewis. 2007. Multilingual structural projection across interlinearized text. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 452–459. Association for Computational Linguistics.