Developing a Shared Task for Speech Processing on Endangered Languages

Gina-Anne Levow, Emily P. Ahn, Emily M. Bender
Department of Linguistics
University of Washington
Seattle, WA USA
levow, eahn, ebender@uw.edu

Abstract

Advances in speech and language processing have enabled the creation of applications that could, in principle, accelerate the process of language documentation, as speech communities and linguists work on urgent language documentation and reclamation projects. However, such systems have yet to make a significant impact on language documentation, as resource requirements limit the broad applicability of these new techniques. We aim to exploit the framework of shared tasks to focus the technology research community on tasks which address key pain points in language documentation. Here we present initial steps in the implementation of these new shared tasks, through the creation of data sets drawn from endangered language repositories and baseline systems to perform segmentation and speaker labeling of these audio recordings—important enabling steps in the documentation process. This paper motivates these tasks with a use case, describes data set curation and baseline systems, and presents results on this data. We then highlight the challenges and ethical considerations in developing these speech processing tools and tasks to support endangered language documentation.

1 Introduction

Language processing technologies have made dramatic strides, achieving sufficiently strong performance on tasks, ranging from text-to-speech transcription (Stolcke and Droppo, 2017) to machine translation (Läubli et al., 2018), to support increasingly broad deployment in commercial applications. However, these accomplishments have been attained on only a small subset of relatively high-resource languages, for example, English speech-to-text transcription and Mandarin-English machine translation in the cases above. At this time, such tools are available for only 100–200 languages.[1]

In contrast, there are few such tools for low-resource or endangered languages. These methods rely heavily on large quantities of labeled training data and computing power for these successes: resources that are unobtainable for such languages. Nevertheless, speech and language processing techniques hold great potential to support documentation and preservation of endangered languages. Speech communities and linguists are working with urgency on language reclamation and documentation, while 50–80% of languages currently spoken could disappear by the end of this century.[2] Field linguists frequently highlight the "funneling" process of language documentation where much more data is recorded than can be transcribed, and more data is transcribed than can be properly analyzed. Levow et al. (2017) reported a needs assessment workshop which aimed to understand the work process of and identify key pain points for researchers engaged in language documentation. This work presented key tasks and guiding principles for the design of shared tasks that bring computational language processing techniques to bear on the problems important to language documentation.

Shared tasks have helped drive development in a broad range of speech and language processing technologies (Belz and Kilgarriff, 2006), from document retrieval (Voorhees and Harman, 2005) to spoken dialog systems (Mesnil et al., 2013; Williams et al., 2016). Shared tasks raise the visibility of the problem of interest while providing valuable standardized training and test data sets, task settings, evaluation metrics, and venues for sharing results and methods. One set of shared tasks proposed by Levow et al. (2017) is the so-

[1] http://translate.google.com
[2] https://www.linguisticsociety.org/content/what-endangered-language

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

called "Grandma's Hatbox" task cascade, which aims to capture the processing steps needed to prepare a set of field recordings and partial transcriptions to facilitate analysis and archiving, such as might be required when a scholar's cache of field recordings is discovered. The cascade takes original unprocessed digital audio recordings and aims to perform preliminary automatic speaker diarization, speaker recognition, language identification, genre identification, and alignment of partial transcriptions with the recorded audio.

Here we focus on the first two stages of this cascade: speaker diarization and speaker identification. Speaker diarization takes a span of audio and segments it into spans at speaker change points and identifies those spans spoken by the same speaker. Then for each speaker span, speaker identification determines which, if any, of a set of known speakers spoke in that span. Creating a model for one of these known speakers is referred to as "enrollment". Our goal is to create baseline systems to support the implementation of shared tasks targeting speaker diarization and speaker identification. We will further assess these baselines on a set of 8 typologically diverse endangered language data sets.

Our baseline systems serve several purposes. They serve to validate the design of the tasks and the data set, highlighting important characteristics of both. They also help to establish baseline scores for the tasks on this data, assessing their difficulty and providing a basis for comparison. We also plan to make their implementation available as part of upcoming shared tasks to lower barriers of entry for participating teams. Additionally, they will allow investigation of the challenges of these tasks on endangered language data while highlighting the assumptions of systems built on standard corpora for high-resource languages. As part of establishing clear baselines and lowering barriers to entry, we leverage existing open-source implementations of speaker diarization and speaker identification. Furthermore, we focus on relatively "lightweight" implementations that do not require extensive, licensed training data, large scale computing resources, or extensive compute time.

In the remainder of this paper, we first briefly describe a use case for these tasks (§2). After presenting related work in §3, we introduce the endangered languages and data sets employed in this work (§4). The baseline systems and experimental results follow in §§5–6. We then discuss the challenges highlighted by these experiments, both in terms of the data itself and of the assumptions underlying existing systems (§7). We conclude with a discussion of ethical considerations in automating this annotation of endangered language materials (§8) as well as plans for future work (§9).

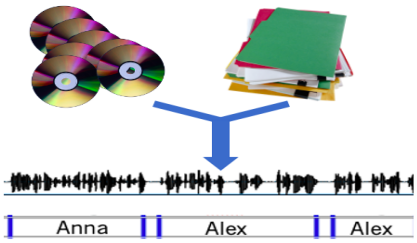

Figure 1: Diarization and speaker ID use case.

2 Exemplar Use Case

We anticipate that speaker diarization and speaker identification could benefit the workflows of both field linguists and managers of endangered language archives. Here we discuss a use case from the perspective of the field linguist; a schematic appears in Figure 1.

After recording a few sessions with their consultants, the linguist could run the speaker diarization system on the recordings, to create a first-pass turn segmentation by speaker in ELAN format. At this stage, turns would be aligned with the audio, but speakers would only be identified by their automatically assigned cluster id, e.g. speaker1, speaker2. The linguist would have the opportunity to correct any errors they observe. The linguist could then associate specific participants — for example, themselves or their primary consultant — with corresponding automatically generated speaker IDs. On this basis the speaker ID model would be trained and then applied to label all of the researcher's and consultant's speaking turns in subsequent recordings. This segmentation and labeling would allow the linguist to navigate more quickly through recordings and focus their analysis and annotation on only those speaker spans of interest. This semi-automatic segmentation and labeling could also be used to enrich the metadata in archive deposits and thus facilitate search through their resources.

3 Related Work

Both speaker diarization and speaker identification have been the subject of ongoing shared task evaluations (Ryant et al., 2018, 2019; NIST, 2016). Earlier work on diarization focused on telephone conversations (Godfrey et al., 1992), broadcast news, and multiparty meetings (Janin et al., 2003). Recent tasks and data sets have refocused attention on more varied and challenging interaction settings, such as child-directed speech, restaurant conversations, and courtroom speech, in the DIHARD (Ryant et al., 2018) task series. However, most diarization task data has been in English or other high resource languages, such as French.

The NIST Speaker Recognition Evaluation (SRE) (NIST, 2016) series has been active since 1996. The data has included both telephone and microphone speech and explored different training and test duration configurations. While earlier iterations focused on English test data, with a mix of languages in the training set, recent years have included test data from Cantonese, Tagalog, and Arabic. More varied speaker recognition data sets are now available, such as "Speakers in the Wild" (SITW) (McLaren et al., 2016) or VoxCeleb (Nagrani et al., 2020), which uses YouTube interviews. Speaker recognition systems have also been built for lower resource languages such as Bengali (Das and Das, 2018) and Uyghur (Rozi et al., 2015).

Endangered language data poses many challenges for speaker diarization and speaker identification. Diarization is sensitive to the style of interaction, e.g. broadcast news vs. courtroom vs. dinner-party conversation, and recordings collected by documentary linguists span diverse domains from structured elicitations to sermons and ceremonies. Recording conditions for documentary linguistic data are also potentially more variable than those in prior studies, many of which have focused on telephone or wideband laboratory recording settings. In addition, we consider endangered languages with areal and typological diversity. Finally and crucially, documentary linguistic data is typically much more limited in quantity, precluding techniques which rely on large amounts of in-language training data.

4 Data

We have run experiments on data collections from 8 different languages all deposited with Endangered Languages Archive (ELAR) at SOAS University of London.[3] Table 1 provides an overview of the data, listing for each language the language family it belongs to and the location (country name) where the fieldwork took place, per the information indicated in the ELAR deposit. In this section, we briefly summarize the available information about the genres represented in each deposit and basic phonological typological information about each language, the latter being summarized in Table 2. Where available, we provide the ISO639-3 language codes.[4]

In addition to genre and phonological typological profile, we expect further kinds of variation across these languages and data sets. For example, speaker communities likely vary in their conventions around overlap between turns of different speakers (how much overlap is allowed before it is considered a rude interruption? do speakers need leave some silence after another's turn? are listeners expected to provide audible backchannels? (Clancy et al., 1996; Levow et al., 2010; Duncan et al., 2010; Tannen, 1994; Goldberg, 1990; Laskowski, 2010)), and recordings likely vary in terms of the amount and type of ambient noise included (animals, traffic, wind). Unfortunately, we do not have access to either type of information and so will have to assume that these data sets do vary along these dimensions.

Cicipu (ISO639-3:awc) The deposit for Cicipu includes "greetings, conversations, hortative discourse, narratives, procedural, and ritual discourse" (McGill, 2012), and elicitation activities. The language is analyzed as having 27 consonants, 6 basic monophthongal vowels all allowing length and nasalization contrasts, and 4 diphthongs, almost always nazalised. It is a tonal language, contrasting high v. low tones and contour tones on long vowels/diphthongs (McGill, 2014).[5]

Effutu (ISO639-3:awu) The deposit for Effutu (Agyeman, 2016b) includes interviews, prompted narratives, elicitations, naturally occurring linguistic events (e.g. the language taking place around fishing activities), natural conversations, folks songs, fables, and radio programming. According to the analysis of Agyeman (2016a),

[3] https://elar.soas.ac.uk/

[4] https://www.iso.org/standard/39534.htm

[5] McGill (2014) doesn't specify which contour tones there are. Given the basic H and L tones, we assume the presence of HL and LH.

Language	Family	Location	Hours	Turn	Sampling
Cicipu	Niger-Congo	Nigeria	3.3	1.9 (1.3)	48k
Effutu	Niger-Congo	Ghana	2.0	3.4(11.1)	44.1k
Mocho'	Mayan	Mexico	4.3	2.0 (1.5)	48k
Northern Prinmi	Sino-Tibetan	China	3.2	5.1(19.0)	48k
Sakun	Afro-Asiatic	Nigeria	9.2	2.7 (2.3)	44.1k
Upper Napo Kichwa	Quechuan	Ecuador	10.0	2.9 (4.6)	48k
Toratán	Austronesian	Indonesia	14.5	2.1 (2.2)	48k
Ulwa	Keram	Papua New Guinea	3.2	3.6 (5.1)	48k

Table 1: Overview of the 8 data sets used in this work. Hours refers to quantity of speaker-annotated, time-aligned data. Turn length specifies the mean (standard deviation) in secs; each data set has a right-skewed distribution with a long tail. Audio was recorded at the Sampling rate in the right column.

Language	# C	# V	# Tones
Cicipu	27	28	4
Effutu	26	17	4
Mocho'	27	10	2
Northern Prinmi	42	13	4
Sakun	39	3	2
Upper Napo Kichwa	20	8	–
Toratán	21	5	–
Ulwa	13	8	–

Table 2: Basic phonological typological properties per language: size of inventory of consonants, vowels, and tones.

Effutu has 26 consonants, 17 vowels (9 oral, 8 nasal), and contrasts 4 tones (high, low, falling, downstep).

Mocho' (ISO639-3:mhc) The deposit for Mocho' (Pérez González, 2018) includes both biographical and non-biographical narratives (the latter including historical events, myths, local beliefs, traditional building practices, witchcraft, etc.), one prayer, conversation, elicitation sessions, and one session involving the translation of a text from another village. According to the analysis by Palosaari (2011), Mocho' has 27 consonants, 5 vowel qualities each contrasting short and long variants, and a tone contrast on long vowels only (unmarked v. falling).

Northern Prinmi (ISO639-3:pmi) The Northern Prinmi deposit is designed to document oral art, from 12 different locations. It includes rituals, traditional songs, folktales and conversations (Daudey and Pincuo, 2018). Northern Prinmi refers to a family of varieties spoken in the area studied. We draw on Daudey's (2014) analysis of one of them, Wǎdū Pǔmǐ, for typological information. She finds 42 consonants, 6 single oral vowels, 4 nasal vowels and 3 diphthongs; and 4 tones (H level, H falling, L level, L rising).

Sakun (ISO639-3:syk) The Sakun deposit (Thomas, nd) is a collection of recordings of discourses (which we assume to be multi-party) pertaining to community cultural practices. According to Thomas (2014), the language has 39 consonants, 3 vowels and contrasts H vs. L tones.

Upper Napo Kichwa The Upper Napo Kichwa deposit (Grzech, 2018) includes grammatical elicitation and life interviews. We believe this language to be the same as or closely related to Tena Quechua (ISO639-3:quw) analyzed by O'Rourke and Swanson (2013), who describe it as having 20 consonants, 3 vowel qualities, a length contrast and 2 diphthongs (au, ai), for 8 total vowels. There is no mention of any tones.

Toratán (ISO639-3:rth) The Toratán deposit (Jukes, nd) includes conversational data, elicitation sessions, and narratives (personal history, folk tales). According to Himmelmann and Wolff (1999), the language has 21 consonants and 5 vowels. There is no mention of any tones.

Ulwa (ISO639-3:yla) The Ulwa deposit (Barlow, 2018a) includes conversational data, traditional and personal stories, and one video of traditional singing and dancing. According to Barlow (2018b), Ulwa has 13 consonants and 8 vowels (including 2 diphthongs), and no tones.

5 Methodology

We first describe steps taken to preprocess the data for the shared task. Then we discuss the configurations of the baseline systems for speaker diarization and identification that pertain to our linguist's use case.

5.1 Data Preprocessing

All files are downloaded from ELAR in two formats: .wav (audio) and .eaf (annotated ELAN files in XML format). As we focus our tasks on types of speaker recognition, we select only files that are annotated for speaker turns, i.e. when a speaker starts and ends their utterance. Each language deposit was created by a different field linguist, and annotation conventions vary across data sets, as we see in the varied means and standard deviations of turn length in column 5 of Table 1. Figure 2 captures the quantity of annotated audio by number of speakers present in the recording. For example, we can see that Mocho' and Ulwa only have recordings with at most 2 speakers while a majority of Northern Prinmi recordings have at least 3 speakers present.

5.2 Baseline Systems

We choose state-of-the-art systems that are lightweight and convenient for distribution, in order to prevent barriers of access for researchers without strong compute power. While all 8 of our data sets are suitable for creating training and test data in a shared task on speaker diarization, only 6 can be used for speaker identification, since that task requires uniform labeling of speakers across different recordings. Thus, Northern Prinmi and Sakun are not used in the speaker ID task.

Speaker Diarization For this task, the goal is to determine who spoke when in a given audio file. LIUM[6] is a lightweight system that uses Integer Linear Programming (ILP) clustering techniques to determine segments of audio that are from the same speaker (Meignier and Merlin, 2010; Rouvier et al., 2013). It was originally developed for broadcast news. LIUM does not require any additional training before use, so we do not split the data into train and test, but rather evaluate on the full data set.

Speaker Identification (ID) In this task that follows diarization, the goal is to identify who the speakers are in new speech segments by comparing them to audio from a set of pre-enrolled speakers. For this task, we leverage the Kaldi toolkit (Povey et al., 2011), specifically the sre08 recipe for speaker ID which implements a strong i-vector model. The recipe usually relies on several large English language speech corpora to train the Gaussian Mixture Models (GMMs) for the Universal Background Model (UBM) and to support i-vector extraction. However, we do not have analogous large corpora available for our endangered languages. Furthermore, the resources typically used for training are not publicly available, are extremely large, and impose substantial computational requirements. Instead, similar to Rozi et al. (2015), we adopt a transfer learning strategy and pre-train these models on the following smaller English corpora: a subset of the Fisher corpus,[7] the NIST SRE 2005 and 2006 training data,[8] and the NIST SRE 2005 test data.[9] The pre-training process required approximately 1 week on a single CPU; however, the resulting system using that pre-trained model is relatively lightweight, requiring only 1-2 hours per individual data set experiment, to train new known speaker models and test the system. Furthermore, we divide the audio files into an enrollment set and test set with a ratio of 80%/20%, ensuring all speakers in the test set have appeared in the enrollment set.

Summary We have employed endangered language data that was already labeled with speaker turn information to create training and test data sets with gold standards. These data sets allow us to build baseline systems as a proof-of-concept for the planned shared task and to assess their effectiveness in controlled experiments. Ultimately, we expect that new techniques developed in the shared tasks will be applied to automatically annotate new field recordings.

6 Experimental Results

Below, we present the results of our speaker diarization and speaker identification baseline systems on the data sets derived from endangered language resources.

[6] https://github.com/StevenLOL/LIUM

[7] https://catalog.ldc.upenn.edu/LDC2004S13

[8] https://catalog.ldc.upenn.edu/LDC2011S01, https://catalog.ldc.upenn.edu/LDC2011S09

[9] https://catalog.ldc.upenn.edu/LDC2011S04

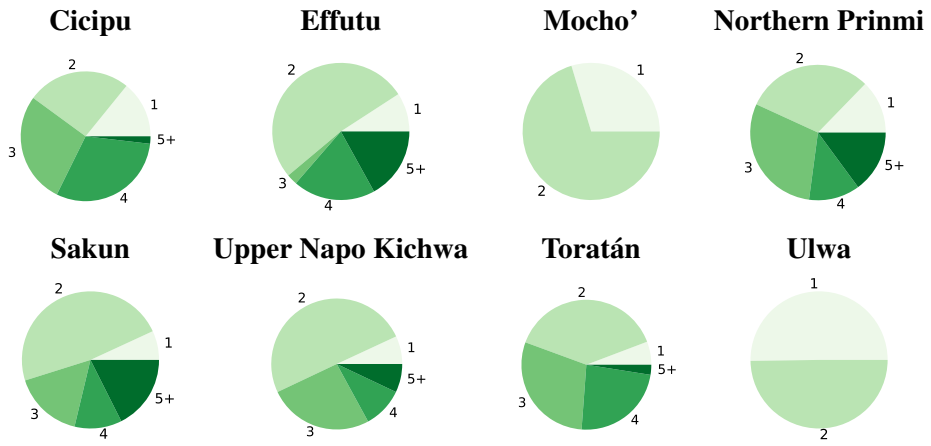

Figure 2: Distribution of annotated audio for each data set by number of speakers in each recording.

6.1 Speaker Diarization

The Diarization Error Rate (DER) (NIST, 2003) is a standard metric which accounts for speaker error, false alarm speech, and missed speech in speaker diarization tasks. We use it to measure the performance of LIUM on each of our data sets and report DER scores in Table 3. DER ranges from 34.7% at best for Effutu to 62.6% for Sakun.

Data Set	DER [%]
Cicipu	44.5
Effutu	34.7
Mocho'	60.2
Northern Prinmi	37.8
Sakun	62.6
Upper Napo Kichwa	43.7
Toratán	55.6
Ulwa	57.9
DIHARD1 (LIUM baseline)	55.8
DIHARD1 (SOTA)	23.7

Table 3: Diarization Error Rate (DER) for the baseline LIUM system on each data set, as well as on English speech from the DIHARD1 challenge evaluation set (Ryant et al., 2018). The final row reports the DER of a state-of-the-art model specifically developed for the DIHARD task from Sell et al. (2018). Lower scores are desirable.

For comparison, we evaluate our baseline LIUM system on the evaluation data set from the DIHARD1 challenge—English speech collected from intentionally difficult settings such as restaurant conversations and clinical interviews (Ryant et al., 2018). The DER of 55.8% is in line with scores from our data, indicating that the endangered language data sets are similarly challenging to diarize as the currently identified diverse and challenging environments of the DIHARD1 data set. The last row in Table 3 shows the score on the same DIHARD1 data from a state-of-the-art model in (Sell et al., 2018) which has been tuned to the specific DIHARD1 task. As LIUM was developed for broadcast news speech, there is much room for researchers to improve upon the system through additional training and tuning.

6.2 Speaker Identification

For Speaker ID evaluation, we adopt the simple cosine-distance scoring method from the Kaldi recipe and report the Equal Error Rate (EER). EER finds the threshold at which the proportion of missed targets (e.g. utterances from Speaker A are not identified with Speaker A) equals the proportion of false acceptances (e.g. utterances from Speaker A are incorrectly identified with Speaker B). Column 3 of Table 4 shows the results of EER from our baseline system performed on each individual data set. Scores range from 12.3% at best for Mocho' to 48.6% for Upper Napo Kichwa. There appears to be a small positive correlation between number of enrolled speakers and EER.

Upon comparison with the NIST SRE 2008 (SRE08) test data[10] of multilingual telephone speech and English interview speech, most EER scores on our data sets are higher than the baseline system run on SRE08. This is predicted, due to the diverse settings of the endangered language data sets.

[10] https://catalog.ldc.upenn.edu/LDC2011S08

It is important to note that in the Kaldi configuration, the speech data can undergo a gender identification step to automatically pre-sort audio segments into male or female, then use the corresponding male or female model. As our data sets do not include metadata on gender and scores were not largely varying across these configurations, we only report system performance on passing all speech through the male model after ignoring the gender ID step.

Data Set	# Spkrs	EER [%]
Cicipu	27	26.4
Effutu	15	42.3
Mocho'	8	12.3
Upper Napo Kichwa	69	48.6
Toratán	20	27.8
Ulwa	6	20.2
SRE08 (Baseline)	–	14.6
SRE08 (SOTA)	–	11.9

Table 4: Speaker Identification Equal Error Rate (EER) for each of the 6 data sets. Number of enrolled speakers is given in Column 2. The last 2 rows report the EER of our baseline Speaker ID system on the NIST SRE 2008 test data, with a state-of-the-art i-vector model from the full sre08 recipe of the Kaldi toolkit Povey et al. (2011). Lower scores are better.

7 Discussion

We assessed the applicability of speech processing techniques for endangered language documentation by creating data sets based on language archives, implementing baseline systems for speaker diarization and identification, and evaluating them against the documentary data. Our experiments demonstrated the feasibility of conducting such evaluations and the utility of the baseline systems and data sets. In addition, this process highlighted challenges in creating data and systems for these tasks.

One source of challenge was the data itself. As the distribution of speakers across recordings demonstrates (see Figure 2), there is significant variation in speech settings both within and across data sets. Differences in genre also drive many of these contrasts. This variation remains a significant challenge for speaker diarization, and motivated the development of the DIHARD tasks to focus on new domains. The performance of the baseline diarization system also varied significantly across our data sets. This variety suggests that endangered language data can be a rich testbed.

In addition, while we tried to exclude data sets with very significant noise, field recordings will inevitably have more variation in recording conditions than in more controlled settings. Compensating for such variation remains a core research challenge for speech processing. Furthermore, while speech processing data collection efforts have often aimed to carefully control for certain social or demographic factors, such as gender, constraints on the field linguistics process make such controls impractical. Several of the endangered language data sets had substantial imbalance in the reported genders of the contributors, potentially affecting these speaker-oriented tasks.

Aspects of transcription and annotation of the source data also impacted the creation of these evaluation data sets. The frequent use of ELAN .eaf for annotation and .wav files for audio was very helpful. However, there was also substantial variation across deposits in terms of naming conventions for ELAN tiers, anonymization or naming of speakers in the transcription files, and in coverage of transcriptions. We managed much of the variation in ELAN tier naming and structure through custom code. However, in cases where the deposits reused the same identifiers for different speakers, it was impossible to create a gold standard for speaker identification. Lastly, we noted different strategies for partial transcription of recordings, including transcribing only one or a few contiguous spans in the recording or transcribing only the consultant side in an elicitation. We managed the former through custom code, while we excluded the latter type in creating gold standards for diarization. Standardized processes could simplify these tasks.

Finally, the adaptation and application of the baseline systems to this new data highlighted some challenging assumptions and design decisions. In the case of speaker identification, the system we adapted assumed access to a large amount of licensed data and computing resources well beyond those of our own research lab, much less a field linguist's laptop or repository's data center. Exploiting such resources has yielded impressive results, but that reliance poses a significant barrier for both developers and users of technology for

endangered languages. Fortunately, we were able to use smaller amounts of high-resource language data with little loss in effectiveness. Also, while not strictly required, both baseline systems incorporated or expected to train a gender identification module. Such annotation is unlikely to be provided with endangered language data. Lastly, the systems which we leveraged were built for specific tasks, such as diarization of broadcast news, or speaker identification under particular noise conditions. Novel data will challenge these systems.

8 Ethical Considerations

There are two broad types of ethical considerations that we would like to raise with respect to this work: considerations to do with the usage of data and considerations to do with the technology that the shared task is meant to help develop.

The data we are using have been deposited with ELAR, an archive for endangered language data. Archives are conceptualized as repositories responsible for handling the long-term preservation of collections of data as well as managing the access to them. Data depositors can establish access rights in accordance with the wishes of the communities they are working with as well as their own considerations (Nathan, 2013; Drude et al., 2012). We have chosen data sets which are accessible to all registered users of the repository. However, we do not plan to redistribute the data. Rather, we ask shared task participants to register for their own ELAR account and download the data, using scripts we provide to preprocess it appropriately for the specific shared tasks. Our purpose in doing so is to ensure that the data sets we work with remain connected to their metadata and access rights as determined by the archive. Should these change, anyone accessing the data through our shared task will have up-to-date information.

A second set of data-level considerations involves speaker privacy and anonymity. When working with audio data, we are necessarily working with personally identifiable information (the speakers' own voices) as well as any personal information included in the recorded speech itself. Our pre-processing scripts include an anonymization step, where speaker names in the metadata are replaced with non-personal identifiers.[11] However, anyone accessing the data for our shared task will also have access to the data sets as deposited in the archive. We rely, here, on the data set depositors to have clearly communicated to the speakers they worked with about what it means for their recordings to be in an internet-accessible archive and to have received consent for the archiving.

At the technology level, we would like to acknowledge that speaker identification and speaker diarization technology, which our shared task aims to help develop, change what it means to have recorded data in an internet-accessible archive. Should we succeed in spurring the development of successful speaker identification or diarization technology for endangered languages, and should there be other recordings of the same individuals available elsewhere on the web, such technology would make it possible to link the different recordings to the same speaker identity, potentially deanonymyzing otherwise anonymous deposits. Beyond the harms to privacy that this represents, it also opens up further risks, such as the potential to amass sufficiently large amounts of data to create deep fakes in the voice of the recorded speakers.

We find that the potential for *dual use* (Hovy and Spruit, 2016) is inherent in speech technology developed for endangered language documentation. That is, alongside the positive value the technology can bring by facilitating language documentation and revitalization, there is also the risk of harmful use. Different speaker communities and indeed different speakers may view these risks differently. What is particularly vexing is that the development of new technology can reshape what risks a speaker is taking on when consenting to be recorded for archival data. We recommend that computational linguists and archivists communicate about the state of technology. In this way, the computational linguists can support the archivists in designing appropriate modifications to the access rights systems and appropriate explanations for data curators—and ultimately speech communities—about the effects of archiving data in a changing technological landscape.

9 Conclusion

We have created a suite of typologically diverse data sets based on endangered language resources, and used that data to build and evaluate baseline systems for speaker diarization and identification. These steps support our planned shared tasks in speech processing for endangered languages,

[11] Furthermore, the original data set producers may have recorded pseudonyms instead of speakers' actual names.

which we hope will spur development of systems to accelerate language documentation. This work has also highlighted the challenges of and ethical considerations for developing such technology. We will bear these in mind as we move forward with deployment of these shared tasks.

Acknowledgments

This work was partially supported by NSF #: 1760475. The authors also thank Isaac Manrique and Cassie Maz for their contributions to this work.

References

Nana Ama Agyeman. 2016a. *A descriptive grammar of Efutu (southern Ghana) with a focus on serial verb constructions: A language documentation study*. Ph.D. thesis, SOAS University of London.

Nana Ama Agyeman. 2016b. Documentation of Efutu. https://elar.soas.ac.uk/Collection/MPI1029692, Accessed on 12 Oct 2020.

Russell Barlow. 2018a. Documentation of Ulwa, an endangered language of Papua New Guinea. https://wurin.lis.soas.ac.uk/Collection/MPI1035105, Accessed on 12 Oct 2020.

Russell Barlow. 2018b. *A Grammar of Ulwa*. Ph.D. thesis, University of Hawaii at Mānoa.

Anja Belz and Adam Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 133–135, Sydney, Australia. Association for Computational Linguistics.

P. Clancy, S. Thompson, R. Suzuki, and H. Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26:355–387.

Shubhadeep Das and Pradip K. Das. 2018. Analysis and comparison of features for text-independent Bengali speaker recognition. In *Proceedings of SLTU 2018*, pages 274–278.

Henriëtte Daudey. 2014. *A Grammar of Wadu Pumi*. Ph.D. thesis, La Trobe University.

Henriëtte Daudey and Gerong Pincuo. 2018. Documentation of Northern Prinmi oral art, with a special focus on ritual speech. https://elar.soas.ac.uk/Collection/MPI1083424, Accessed on 12 Oct 2020.

Sebastian Drude, Daan Broeder, Paul Trilsbeek, and Peter Wittenburg. 2012. The Language Archive — a new hub for language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3264–3267, Istanbul, Turkey. European Language Resources Association (ELRA).

Susan Duncan, Gina-Anne Levow, Tizzian Baldenebro, and Atoor Lawandow. 2010. Multi-modal analysis of interactional rapport in three language/cultural groups. In *HCD 2010*.

J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics,Speech and Signal Processing, vol. 1*, page 517–520.

J. A. Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power and rapport-oriented acts. *Journal of Pragmatics*, 14:883–903.

Karolina Grzech. 2018. Upper Napo Kichwa: A documentation of linguistic and cultural practices. https://elar.soas.ac.uk/Collection/MPI849403, Accessed on 12 Oct 2020.

Nikolaus P Himmelmann and John U Wolff. 1999. *Toratán (Ratahan)*, volume 130. Lincom Europa.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, and A. Stolcke. 2003. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1*.

Anthony Jukes. nd. Documentation of Toratán (Ratahan). https://elar.soas.ac.uk/Collection/MPI87803, Accessed on 12 Oct 2020.

K. Laskowski. 2010. Modeling norms of turn-taking in multi-party conversation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 999–1008.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Gina-Anne Levow, Emily M. Bender, Patrick Littell, Kristen Howell, Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, David Inman, Michael Maxwell, Michael Tjalve, and Fei Xia. 2017. STREAMLInED challenges: Aligning research interests with shared tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47, Honolulu. Association for Computational Linguistics.

Gina-Anne Levow, Susan Duncan, and Edward King. 2010. Cross-cultural investigation of prosody in verbal feedback in interactional rapport. In *Interspeech 2010*, pages 286–289.

Stuart McGill. 2012. Cicipu documentation. https://elar.soas.ac.uk/Collection/MPI97667, Accessed on 12 Oct 2020.

Stuart McGill. 2014. Cicipu. *Journal of the International Phonetic Association*, 44(3):303–318.

M. McLaren, L. Ferrer, D. Castan, and A. Lawson. 2016. The speakers in the wild SITW speaker recognition database. In *Proceedings of Interspeech 2016*, pages 818–822.

S. Meignier and T. Merlin. 2010. LIUM SpkDiarization: An open source toolkit for diarization. In *Proceedings of CMU SPUD Workshop*.

Gregoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech 2013*, pages 3771–3775.

A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 60.

David Nathan. 2013. Access and accessibility at ELAR, a social networking archive for endangered languages documentation. *Oral Literature in the Digital Age: Archiving Orality and Connecting with Communities*, pages 21–40.

NIST. 2003. The rich transcription spring 2003 (RT-03S) evaluation plan.

NIST. 2016. 2016 NIST Speaker Recognition Evaluation Plan. https://www.nist.gov/file/325336. Downloaded October 8, 2016.

Erin O'Rourke and Tod D Swanson. 2013. Tena Quichua. *Journal of the International Phonetic Association*, 43(1):107–120.

Naomi Elizabeth Palosaari. 2011. *Topics in Mocho' Phonology and Morphology*. The University of Utah.

Jaime Pérez González. 2018. Documentation of Mocho' (Mayan): Language preservation through community awareness and engagement. https://elar.soas.ac.uk/Collection/MPI1079685, Accessed on 12 Oct 2020.

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.

M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier. 2013. An open-source state-of-the-art toolbox for broadcast news diarization. In *Interspeech 2013*, pages 1477–1481.

Askar Rozi, Dong Wang, Zhiyong Zhang, and Thomas Fang Zheng. 2015. An open/free database and benchmark for Uyghur speaker recognition. In *2015 International Conference Oriental COCOSDA*, pages 81–85.

Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2018. First DIHARD challenge evaluation plan. *2018, tech. Rep.*

Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2019. The second DIHARD diarization challenge: Dataset, task, and baselines. In *Proceedings of Interspeech 2019*, pages 978–982.

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, et al. 2018. Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Interspeech*, pages 2808–2812.

Andreas Stolcke and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. In *Proceedings of Interspeech 2017*, page 137–141.

D. Tannen. 1994. *Gender and Discourse*. Oxford University Press, New York.

Michael Thomas. nd. Sakun (Sukur) language documentation. https://elar.soas.ac.uk/Collection/MPI184105, Accessed on 12 Oct 2020.

Michael F Thomas. 2014. *A Grammar of Sakun (Sukur)*. Ph.D. thesis, University of Colorado at Boulder.

Ellen M. Voorhees and Donna K. Harman, editors. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Digital libraries and electronic publishing series. The MIT Press, Cambridge, MA.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.