Investigating Speaker Diarization of Endangered Language Data

Gina-Anne Levow
Department of Linguistics
University of Washington
Seattle, WA USA
levow@uw.edu

Abstract

The task of speaker diarization aims to determine which speakers spoke when in a recording. Such functionality could help to accelerate work in endangered languages by facilitating transcription and semi-automatically extracting useful meta-data to enrich language archives. However, there has been little work on speaker diarization for low-resource or endangered languages. This work explores three neural approaches to speaker diarization applied to data sets drawn from endangered language archives. We find consistent improvements for recent neural x-vector models over earlier approaches. We also assess the factors which impact performance across models and data sets, with a focus on the challenging characteristics of endangered language recordings.

1 Introduction

The task of speaker diarization aims to determine which speakers spoke when and for how long in a recording. Such functionality could facilitate work in endangered language documentation and revitalization by accelerating transcription and semi-automatically extracting useful meta-data to enrich language archives, making such materials more accessible to researchers and speaker communities. It could also help to alleviate the so-called "transcription bottleneck." The DIHARD series of shared task evaluations (Ryant et al., 2018) has brought renewed attention to the challenges of speaker diarization in a variety of interaction scenarios. However, there has been little work on speaker diarization for low-resource or endangered languages. This work explores three neural approaches to speaker diarization applied to data sets drawn from eight geographically and typologically diverse endangered language archives. We find consistent improvements for recent neural x-vector models over earlier approaches. We also assess the factors which impact performance across models and data sets, with a focus on the characteristics of endangered language recordings which prove challenging.

2 Related Work

Speaker diarization has been the subject of ongoing shared task evaluations (Ryant et al., 2018, 2019, 2020). 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 or Chinese.

Some prior work has explored speaker diarization on endangered language data. Le Franc et al. (2018) investigated an easy-to-use diarization tool on child speech recordings, including datasets of Tsimane, an endangered Bolivian language, and Tseltal in Mexico. Both child speech data and recording environments posed challenges. Most closely related is work in Levow et al. (2021), which creates data sets and baselines for planned shared tasks on speech processing for endangered languages and reports diarization results from a LIUM (Rouvier et al., 2013) baseline. The current work investigates a range of stronger baselines and further analysis.

Endangered language data poses many challenges for speaker diarization. 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.

3 Data

Our experiments employ data following the language and data set selection and pre-processing of Levow et al. (2021), described briefly below. The data sets are drawn from eight different languages deposited in the Endangered Language Archive, now housed at `http://elararchive.org`. Recordings and accompanying time-aligned transcriptions in ELAN (Brugman and Russel, 2004) format form the basis for both the experimental data and the requisite gold-standard speaker segmentation for evaluation. For each language, we provide information about its language family, the ISO639-3 language codes where available[1], location of the fieldwork, and the genres represented in the collection. In addition, the total duration as well as the mean and standard deviation of turn lengths for recordings in the experimental data sets are presented.

Cicipu (ISO639-3:awc) The deposit for Cicipu a language of the Niger-Congo family, was collected in Nigeria and includes "greetings, conversations, hortative discourse, narratives, procedural, and ritual discourse" (McGill, 2012). The experimental data set comprises 3.3 hours of audio with an average turn length of 1.9 seconds, with a standard deviation of 1.3 seconds.

Effutu (ISO639-3:awu) The deposit for Effutu (Agyeman, 2016), a language of the Niger-Congo family, was collected in Ghana and includes interviews, prompted narratives, and elicitations, among others. The experimental data set comprises 2.0 hours of recordings, with mean turn length of 3.4 seconds, and standard deviation of 11.1s.

Mocho' (ISO639-3:mhc) The deposit for Mocho' (Pérez González, 2018), a Mayan family language, was recorded in Mexico and includes both biographical and non-biographical narratives (the latter including historical events, myths, etc.), a prayer, conversation, elicitation sessions, among others. The experimental set comprises 4.3 hours of recordings, with an average turn length of 2.0s (1.5s standard deviation).

Northern Prinmi (ISO639-3:pmi) The Northern Prinmi deposit aims to document oral art, including rituals, traditional songs, folktales and conversations (Daudey and Pincuo, 2018), from multiple locations. The data set used for experimentation includes 3.2 hours of audio, with turns averaging 3.2s, standard deviation of 19.0s.

Sakun (ISO639-3:syk) The Sakun deposit (Thomas, nd), of the Afro-Asiatic language family, is a collection of recordings of discourses typically among 2-5 participants relating to community cultural practices, collected in Nigeria. The experimental data set spans 9.2 hours, with mean turn length of 2.7s and standard deviation of 2.3s.

Upper Napo Kichwa The Upper Napo Kichwa deposit (Grzech, 2018) describes a Quechuan family language of Ecuador[2]. It includes grammatical elicitation and life interviews. The resulting experimental data set includes 10 hours of audio, with mean turn duration of 2.9s and standard deviation of 4.6s.

Toratán (ISO639-3:rth) The deposit for Toratán (Jukes, nd) (Austronesian, collected in Indonesia) spans conversational data, elicitation sessions, and various narratives. The experimental data set covers 14.5 hours of audio recordings, with mean turn lengths of 2.1s and standard deviation of 2.2s.

Ulwa (ISO639-3:yla) The deposit for Ulwa (Barlow, 2018), a Keram family language of Papua New Guinea, includes conversational data as well as traditional and personal stories. The experimental dataset comprises 3.2 hours of audio, with mean turn length of 3.6s and standard deviation of 5.1s.

4 Diarization Models

We compare three recent, high performing neural diarization models: the speaker diarization recipes (Snyder et al., 2018; Sell et al., 2018) for Kaldi (Povey et al., 2011), pyannote.audio (Bredin et al., 2020; Bredin and Laurent, 2021), and VBx (Landini et al., 2022). All models leverage a pre-trained neural network model to compute a so-called "x-vector" embedding representation of the

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

[2] It is likely closely related to Tena Quechua (ISO639-3:quw).

audio. These x-vectors are then employed in subsequent clustering stages to produce a global segmentation and assignment of speakers to segments. In all cases, we used the system's included or publicly released x-vector models, without additional task-specific fine-tuning or adaptation. Furthermore, all models were run with default hyperparameters in a fully unsupervised mode, where the true segmentation, voice activity detection, and number of speakers were unknown and needed to be determined by the model. Additional model details are below.

Kaldi The Kaldi toolkit (Povey et al., 2011) provides reference implementations for a range of state-of-the-art models for speech processing, in the form of "recipes" to reproduce published approaches. For the speaker diarization task, we use the pre-trained x-vector models (Callhome Diarization Xvector Model 1a) (Snyder et al., 2018) available from the Kaldi site [3] with the DIHARD_2018 (Sell et al., 2018) recipe [4]. After creating x-vector representations of spans of the audio stream, the recipe uses a probabilistic linear discriminant analysis (PLDA) backend to score the similarity between spans after which clustering is performed to produce the final diarization output.

pyannote.audio The `pyannote.audio` package (Bredin et al., 2020; Bredin and Laurent, 2021) provides a neural speaker diarization pipeline, which is available through Hugging Face [5]. The pipeline leverages the SpeechBrain (Ravanelli et al., 2021) implementation of the ECAPA-TDNN (Desplanques et al., 2020) model, based on its superior performance within this pipeline. The approach applies a local neural speaker segmentation over 5 second windows of speech, from which local speaker embeddings are computed for each speaker in the window. A global agglomerative hierarchical clustering is then performed prior to a final aggregation phase which yields the full diarization.

VBx The VBx (VBHMM x-vectors Diarization) model (Landini et al., 2022) also computes x-vector representations over input audio spans. It then performs agglomerative hierarchical clustering over those representations as an initialization phase for a Variational Bayes (VB) Hidden Markov Model over the x-vectors to create final diarization output. This approach requires a preliminary segmentation created using voice activity detection (VAD) to distinguish speech and non-speech regions; we employ the Kaldi VAD scripts for this purpose.

[3] https://kaldi-asr.org/models/m6

[4] For simplicity, we employ the model without Variational Bayes inference, though we expect that better effectiveness would be achieved with it as in prior work.

[5] https://huggingface.co/pyannote/speaker-diarization

4.1 Baseline Diarization Model: LIUM

We contrast the three neural diarization models above with the LIUM (Rouvier et al., 2013) system baseline presented in Levow et al. (2021), applied to the same data sets as in the current work. LIUM employs the previously popular i-vector model to represent speech spans and an Integer Linear Programming (ILP) approach to globally optimize the assignment of speech spans to particular speakers. LIUM's publicly available Java implementation [6] was used in an unsupervised setting, so that none of the endangered language data was used for tuning or training, and was all treated as test data.

5 Experiments & Results

All three neural speaker diarization models were applied to the eight endangered language data sets. The standard evaluation metric for speaker diarization is Diarization Error Rate (DER), which combines: 1) speaker error: the portion of scored time assigned to the wrong speaker, 2) false alarm speech: portion of scored time incorrectly labeled as speech, and 3) missed speech: portion of scored time incorrectly labeled as non-speech. All measures were computed with the `dscore` package [7].

5.1 Overall findings

Results are shown in Table 1. We can observe that all three neural diarization models outperform the LIUM i-vector baseline system. All pairwise system differences are significant by Wilcoxon signed rank test ($p < 0.05$), except for Pyannote/VBx. In addition, the best effectiveness for each of the languages is given either by Pyannote or VBx.

In addition to the overall differences in system effectiveness, there is also substantial variation across the different languages, with DER ranges of 25 or more across languages within each of the systems. The Sakun and Toratán data sets appear to be challenging for all of the systems, while Prinmi and Effutu conversely present less challenge overall.

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

[7] https://github.com/nryant/dscore

	LIUM	Kaldi	Pyannote	VBx
Cicipu	44.5	43.78	**32.45**	33.31
Effutu	34.7	31.46	34.88	**26.55**
Mocho'	60.2	51.85	**27.37**	33.36
Northern Prinmi	37.8	29.10	23.3	**22.97**
Sakun	62.6	57.37	**48.8**	52.45
Upper Napo Kichwa	43.7	35.26	**26.15**	30.34
Toratán	55.6	44.46	**40.13**	44.16
Ulwa	57.9	37.94	41.04	**29.23**

Table 1: Diarization Error Rates for Kaldi (DER), Pyannote, and VBx compared to a baseline LIUM system for eight endangered language data sets. Lower scores are better; best results for each language are in bold.

Mocho' and Ulwa yield quite variable performance across systems.

5.2 Analysis

Since we observe substantial variation in DER across languages and across files within languages, we investigate factors which might contribute to this disparity. One dimension of variation across datasets and recordings is the number of speakers per recording, ranging between 1 and 18, with a mean of 2.55. Speaker error is also a significant element of DER. So, we assess correlation between number of speakers and per-recording DER. For all systems, there was a highly significant ($p < 0.0001$) positive correlation between the true number of speakers and DER (Spearman: Kaldi: 0.28; VBx, Pyannote: 0.42)[8]. Thus, having more speakers in a recording was associated with more errors. If we look further at how the difference between the predicted and true speaker counts relates to DER, we find a significant correlation between predicting too few speakers and increased DER for Pyannote (0.22, $p < 0.0001$), but conversely a significant correlation with over-predicting speakers and DER for Kaldi. No such correlation was found for VBx. It seems reasonable that distinguishing among larger numbers of speakers would be intrinsically more difficult, and that models also differ in terms of over- or under-segmenting speech. Further, this difficulty can be exacerbated in endangered language recordings where a single microphone is often used and the distance from the speaker to the microphone may vary substantially. If available, supervision in the form of the number of known speakers could help to mitigate this problem.

Missed and false alarm speech also contribute to DER, and, while the data sets were constructed to exclude particularly poor or noisy recordings, background and environmental noise are likely to be more prevalent and variable in endangered language recordings than in more typical diarization data. This is another dimension along which models may differ. To investigate the effects of voice activity detection, we compute, for each experimental file, the total speech duration in the diarization output for each system and for the reference segmentation. Then we compute the correlation between the disparity in the system and the reference durations with the DER score for each file. We find a highly significant correlation between missing speech and DER for Pyannote (0.31, $p < 0.0001$), while, for VBx and Kaldi, false alarm speech is associated with higher DER. These associations highlight the challenges of accurate speech detection in varied and variable fieldwork recordings, and point to the importance of developing better and more adaptable techniques.

6 Ethical considerations

Speech remains intrinsically personally identifying information and can also expose personal information in the spoken content. Data pre-processing removes individual speaker names from the experimental data. However, the potential still exists to link data with other, potentially identifying information elsewhere on the web. As a result, these systems do raise the risk to privacy or harm from use in deep fakes. Furthermore, while these tools are being developed to support language documentation and revitalization, the same speaker-linking that these models create could itself increase the potential privacy risks. It is important that researchers communicate these possibilities and work with language archives and communities to craft appropriate access and use.

7 Conclusion & Future Work

We have evaluated three recent neural models on a diverse suite of endangered language archive data, with no special tuning, demonstrating improvements over a baseline i-vector model and potential increases in utility for researchers, archivists and community members. Further analysis highlights

[8] All correlations are Spearman correlation in the `scipy.stats` package

the challenges of the variation in this data, and the differences in how the models address them. Future work will explore adaptation to the endangered language data, and overall improvements to speech detection in these sorts of varied environments.

Acknowledgements

This work has been supported by NSF: #1760475. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government. Many thanks to Emily M. Bender, Emily P. Ahn, Cassandra Maz, Siyu Liang, and Isaac Manrique for their valuable contributions this work and the broader Streamlined project.

References

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

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

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

H. Brugman and A. Russel. 2004. Annotating multimedia/ multi-modal resources with ELAN. In *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*.

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.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proceedings of Interspeech 2020*.

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.

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.

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.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.

Adrien Le Franc, Eric Riebling, Julien Karadayi, Yun Wang, Camila Scaff, Florian Metze, and Alejandrina Cristia. 2018. The ACLEW DiViMe: An Easy-to-use Diarization Tool. In *Proc. Interspeech 2018*, pages 1383–1387.

Gina-Anne Levow, Emily P. Ahn, and 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*.

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

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.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva,

François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, , and Yoshua Ben-gio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

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.

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.

Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2020. The third dihard diarization challenge.

G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, M. Villalba, J. a nd Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and J. Khuda npur. 2018. Diarization is hard: Some experiences and lessons learned for the jhu team in th e inaugural dihard challenge. *Interspeech*.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

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