Assessing Pre-Built Speaker Recognition Models for Endangered Language Data

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

Abstract

Significant research has focused on speaker recognition (SR), determining which speaker is speaking in a segment of audio. However, few experiments have investigated speaker recognition for very low-resource or endangered languages. Furthermore, speaker recognition has the potential to support language documentation and revitalization efforts, making recordings more accessible to researchers and communities. Since endangered language datasets are too small to build competitive speaker representations from scratch, we investigate the application of large-scale pre-built speaker recognition models to bridge this gap. This paper compares four speaker recognition models on six diverse endangered language data sets. Comparisons contrast three recent neural network-based x-vector models and an earlier baseline i-vector model. Experiments demonstrate significantly stronger performance for some of the studied models. Further analysis highlights differences in effectiveness tied to the lengths of test audio segments and amount of data used for speaker modeling.

Keywords: speaker recognition, endangered languages

1. Introduction

Recent advances have led to substantial improvements in many natural language and speech processing tasks. However, such systems are largely focused on and available for a few hundred, typically high-resource, languages. In contrast, a significant language technology gap remains for many of the world's languages, which may be lower-resource or endangered. At the same time, there are significant efforts to document, research, and revitalize these languages. Language technologies have potential to support these efforts.

Current speaker recognition (SR) models are developed on large datasets, such as VoxCeleb2 (Nagrani et al., 2020), with over 2k hours of recordings, over 1M utterances from 6k speakers. In contrast, our endangered language datasets range from 2 to 14.5 hours. The requirements for training data size and computational power preclude building such models from scratch for endangered languages. Fortunately, high-performing pre-built models have been released and can potentially be used to create good speaker representations for endangered language data. However, a mismatch remains between languages used to build the models and those we hope to apply them to.

This paper investigates the use of pre-built speaker recognition systems for endangered language data, which could support documentation efforts by automatically enriching metadata or facilitate access to recorded materials by community members. Figure 1 depicts this process. For example,

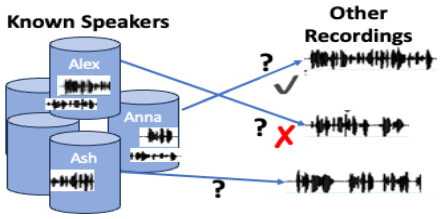

Figure 1: Illustration of speaker recognition

speaker recognition could allow community members to automatically identify recordings from a particular speaker in an audio collection, even in the absence of complete, manually created metadata. Similarly, such tools could allow endangered language archives to semi-automatically enrich metadata with speaker information for their deposits. Also, a field linguist could use such a system to identify speech from a particular consultant, and exclude the researcher's own speech, when prioritizing recordings for transcription.

This paper compares four speaker recognition models on six diverse endangered language data sets. Comparisons contrast three recent neural network-based x-vector models and an earlier baseline i-vector model. Experiments demonstrate significantly stronger performance for some of the studied models. Further analysis highlights differences in effectiveness tied to the lengths of test audio segments and amount of speaker modeling data.

SIGUL2024 Workshop, pages 27–32
21-22 May, 2024.

2. Related Work

Speaker recognition (or speaker identification) has long been an area of research interest. 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, as well as audio-visual settings. The Odyssey workshops have also promoted work on speaker recognition. Other large speaker recognition data sets are now available, such as "Speakers in the Wild" (McLaren et al., 2016) or VoxCeleb (Nagrani et al., 2020), which use YouTube interviews. Systems have also been built for lower resource languages such as Bengali (Das and Das, 2018) and Uyghur (Rozi et al., 2015).

A range of models for speaker recognition have been developed leveraging these resources and evaluation programs. i-vector models (Verma and Das, 2015), which dominated the field, have now largely been supplanted by x-vector models. X-vector models (Snyder et al., 2018) use neural networks pre-trained on large amounts of supervised speaker identification data to create embedding representations of new audio. A variety of modifications and improvements to the standard x-vector model have been developed (Desplanques et al., 2020; Li et al., 2020). In addition, enhancements over simple cosine similarity between vectors have been implemented, such as PLDA (Biswas et al., 2014), though cosine remains a strong approach.

Endangered language data presents a number of challenges for speaker recogntion. Documentary linguistic data may have significant variation in recording conditions, for instance due to background noise from public or outside settings. In contrast, most speaker recogntion data has focused on telephone or wideband laboratory recording settings, though datasets such as VoxCeleb include YouTube videos in a wide range of settings. Further, our endangered language datasets were chosen for 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

The experiments below follow Levow et al. (2021) in terms of data set and selection as well as pre-processing. Six different languages stored in the Endangered Language Archive, `http://elararchive.org`, were chosen to provide typological and areal variety. Gold-standard speaker segments for training and evaluation are derived from the recordings and accompanying time-aligned transcriptions in ELAN (Brugman and Russel, 2004) format. We note that this data is drawn from diverse genres, including greetings, narrative and ritual discourse, interviews, elicitations, folktales, and cultural practices.

For each language, we provide information about its language family, the ISO639-3 language codes where available, location of the fieldwork, as well as overall statistics about recording and turns lengths in the experimental data.

Cicipu (ISO639-3:awc) is a Niger-Congo family language, and the material for this deposit was collected in Nigeria (McGill, 2012). 3.3 hours of audio form the experimental data set, with an average turn length of 1.9 seconds, with a standard deviation of 1.3 seconds.

Effutu (ISO639-3:awu) (Agyeman, 2016) is a Niger-Congo family language, with data collected in Ghana. 2.0 hours of recordings form the experimental data set, with mean turn length of 3.4 seconds, and standard deviation of 11.1s.

Mocho' (ISO639-3:mhc) (Pérez González, 2018) is a Mayan family language, and the data for this deposit recorded in Mexico. 4.3 hours of recordings are available in the experimental data set, with an average turn length of 2.0s (1.5s standard deviation).

Upper Napo Kichwa (Grzech, 2018) (U. N. Kichwa in tables.) is a Quechuan family language, and the material for this deposit was collected in Ecuador. 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) (Jukes, nd) is an Austronesian language, and the material for this deposit was collected in Indonesia. 14.5 hours of audio are included in the experimental data; mean turn length is 2.1s, and standard deviation 2.2s.

Ulwa (ISO639-3:yla) (Barlow, 2018) is a Keram family language, with data collected in Papua New Guinea The experimental dataset includes 3.2 hours of audio, with mean turn length of 3.6s and standard deviation of 5.1s.

4. Speaker Recognition Models

All approaches share a comparable overall architecture. They employ a pre-trained model that creates vector representations from new input audio. These models are trained on large-scale external speech datasets, distinct from the current endangered language data. Representations of audio samples are then compared. The details of the different models are presented below.

4.1. Kaldi

This approach is based on the sre08 (v1) recipe in the Kaldi (Povey et al., 2011) speech processing toolkit. Following the baseline system presented in (Levow et al., 2021), this approach builds a strong i-vector model, using data from a subset of the Fisher corpus (Cieri, Christopher, et al., 2004), NIST SRE 2005 (NIST Multimodal Information Group, 2011c) and 2006 (NIST Multimodal Information Group, 2011a) training datasets, and NIST SRE 2005 test data (NIST Multimodal Information Group, 2011b). This represents a subset of the full sre08 recipe and was chosen due to resource limitations. This data enables the creation of the Gaussian Mixture Models (GMM) for the Universal Background Model (UBM) which support i-vector extraction.

4.2. Pyannote

We employed the pyannote (Bredin et al., 2020; Coria et al., 2020) embedding model from Hugging Face[1]. This embedding uses a standard x-vector TDNN (Time Delay Neural Network) (Snyder et al., 2018) enhanced with trainable SincNet features replacing filterbank features. TDNN approaches apply statistic pooling to create fixed dimension representations from variable length input audio. The model is trained on the VoxCeleb dataset (Nagrani et al., 2020). It achieves a 2.8% Equal Error Rate (EER) on the standard VoxCeleb 1 test set.

4.3. SpeechBrain (xvec)

We also applied the SpeechBrain x-vector model (Ravanelli et al., 2021) from Hugging Face[2] to create x-vector embeddings. This model also employs a pre-trained TDNN-based model. This model was trained on the VoxCeleb 1 and 2 training datasets, and reaches an EER of 3.2% on the VoxCeleb 1 test set.

4.4. SpeechBrain (ECAPA)

Finally, we compared the above models to the SpeechBrain ECAPA-TDNN pre-trained model, using the implementation on Hugging Face[3]. ECAPA (Emphasized Channel Attention, Propagation, and Aggregation) (Desplanques et al., 2020) incorporates improvements to the basic TDNN architecture with factors such as frame-level attention and more effective exploitation of hierarchical features. This model was also trained on VoxCeleb 1 and 2, achieving an EER of 0.8%.

[1] https://huggingface.co/pyannote/embedding

[2] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[3] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

Language	# Known Spkrs	# Seg Spkrs	# Files	Total Tests
Cicipu	27	5	10	1906
Effutu	15	6	4	514
Mocho'	8	5	7	1576
U. N. Kichwa	69	9	17	6768
Toratán	18	7	9	8686
Ulwa	6	6	4	654

Table 1: Statistics of evaluation data

For all the neural models, we used default settings for the pre-trained models with no additional training or parameter tuning.

5. Experiments & Findings

We follow the basic structure of the NIST Speaker Recognition Evaluation (SRE) tasks. A set of known speakers are enrolled by providing one or more instances of their recorded speech. During evaluation, an unseen audio segment is presented along with a known speaker identity. In a "target" pair, that known speaker's speech is present in the new audio sample; in a "non-target" pair, it is not. The system must assign a score to each speaker-segment pair. Equal Error Rate, computed based on that score and gold-standard target/non-target label, provides a single figure of merit, balancing between false alarms and misses.

We leveraged the data pre-processing and training/test splits for each of the six endangered language data sets from (Levow et al., 2021). The evaluation data is evenly split between target and non-target instances, and all test segments are drawn from held-out recording session files. Statistics of the data are shown in Table 5[4].

We applied all three new neural network models to that data, and compare to the results for the baseline i-vector model reported in (Levow et al., 2021). In each of the neural x-vector models, we extracted an embedding for each audio segment. We evaluated two configurations. In one set of experiments, we used those embeddings directly, computing the representation for a known speaker as the average of the individual training sample x-vectors and scoring each speaker-segment pair with cosine distance computed using *scipy cdist* function. In the second set, we applied (in-domain adapted) ADT PLDA [5] with hyperparameters tuned on a small development set to create the segment representations, again averaging to create known speaker models, and scoring with likelihood ratio.

[4] Due to model constraints, test segments were a minimum of 0.75 secs.

[5] https://github.com/RaviSoji/plda/

	Kaldi	Pyan	SB (xvec)	SB (ECAPA)
Cicipu	26.0	12.97	17.83	**5.98**
Effutu	42.0	21.7	32.29	**15.56**
Mocho'	11.5	**8.375**	12.30	9.39
U. N. Kichwa	49.2	**40.25**	46.69	42.17
Toratán	27.3	19.52	30.43	**16.96**
Ulwa	19.9	15.36	19.87	**11.62**
With PLDA				
Cicipu		11.41	18.57	**7.87**
Effutu		18.97	29.96	**7.74**
Mocho		7.42	**7.23**	8.12
U. N. Kichwa		**37.77**	45.5	38.06
Toratan		19.19	25.12	**6.19**
Ulwa		**8.10**	13.76	9.39

Table 2: Equal Error Rates (EER) for Pyannote), SpeechBrain (SB) (xvec), and SpeechBrain (SB) (ECAPA) compared to a baseline Kaldi system for six endangered language data sets. X-vector&cosine above; x-vector&PLDA&likelihood ratio below. Lower scores are better; best results for each language/block are in bold.

5.1. Overall Findings

The EER values for each model applied to each of the six endangered language data sets appear in Table 2. The best overall effectiveness was found for the Pyannote and SpeechBrain ECAPA models, in both configurations, with the best performance for each language being reached by one of these two models (shown in bold in Table 2), except for Mocho' PLDA. The Kaldi i-vector and SpeechBrain (xvec) models did not perform as strongly, with the Kaldi model having the weakest average EER scores. With cosine, all pairwise system differences were significant by Wilcoxon test ($p < 0.05$), except for Kaldi vs. SpeechBrain (xvec) and Pyannote vs. SpeechBrain (ECAPA). With PLDA, although numerically better - sometimes substantially - in all but three cases, only the improvement for Pyannote reached significance ($p < 0.05$), and cross-model differences did not reach significance. The difference between best and worst models reached a factor of four for some languages. It is also important to note that there were large differences between languages as well as across models. The Upper Napo Kichwa data set was challenging for all models with EERs near or above 40%. In contrast, the EER for the best performing data set overall, Mocho', had 75% lower EER. Finally, all EERs remain substantially higher than for the same models on the VoxCeleb test set.

5.2. Analysis

To better understand the source of the variations in data set and model performance, we conduct further analysis. In particular, we focus on two factors relating to sample size: (1) duration of test audio segments and (2) amount of data used train known speaker representations.

Audio segment length has been used as a contrastive factor in prior NIST SRE tasks (NIST, 2016), and can impact tasks such as language identification (Styles et al., 2023). We also note that the annotated speaker segments for the endangered language data sets average only 2-5 seconds. To assess the impact of test audio segment duration, we broke down results by length into 0.5s bins, using the threshold associated with EER to compute accuracy. We focus on the "target" instances, where the new segment and speaker representation should have high similarity. For each of the models, we find a highly significant correlation[6] of accuracy with segment duration, ranging from correlation of 0.69 ($p < 0.0001$) for SpeechBrain (xvec) to 0.22 ($p < 0.01$) for ECAPA, both with and without PLDA.

We also observe in our data sets that there is substantial variation in the amount of enrollment training data for the known speaker models. One speaker has only a single instance of roughly 1 second, while another reaches almost 11000 instances for a total of more than 5 hours. Here we compute the total duration of enrollment training data for each speaker. We then check the correlation of the target and non-target accuracies for each speaker. We find a significant negative correlation of amount of speaker data with non-target accuracy, under all models. In other words, speakers modeled with less total audio data are less likely to be mistakenly matched to a new audio segment. Possibly, larger amounts of modeling data can capture too much within-speaker variation, making it harder to exclude incorrect matches. This observation suggests the need for alternate strategies to incorporate speaker modeling audio data.

6. Conclusion & Future Work

This paper has investigated the effectiveness of three pre-built neural x-vector models and a baseline i-vector model for speaker recognition on six endangered language datasets. Experimental results indicate better effectiveness for the SpeechBrain (ECAPA) and Pyannote models, while highlighting substantial variation across data sets. Analysis showed the impact of test segment duration and amount of speaker modeling data.

These experiments highlight the need for better modeling of short segments and integration of

[6] Correlation is computed with *scipy.stats.spearmanr*

speaker enrollment data. Future work will also explore approaches to fine-tune existing models to better match the endangered language data.

7. Acknowledgements

This work was supported by NSF #1760475. Any opinions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We are grateful to ELAR for their invaluable work. We acknowledge the helpful feedback of anonymous reviewers. Many thanks also to Emily M. Bender for her guidance in this project, and to Emily Proch Ahn, Siyu Liang, Isaac Manrique, and Cassandra Maz for their contributions.

8. Ethical Considerations

Speech is intrinsically personally identifying information. Speaker names are anonymized during data set preprocessing, but speaker recognition links audio to speaker identities. Thus models of these speakers could possibly be linked to non-anonymized speech samples elsewhere on the Web. Furthermore, work risks "dual use" where models designed to support research or community access could instead be exploited for harmful purposes, such as spoofing.

9. Bibliographical References

S. Biswas, J. Rohdin, and K. Shinoda. 2014. i-Vector selection for effective PLDA modeling in speaker recognition. In *Proceedings Odyssey 2014—The Speaker and Language Recognition Workshop*, page 100–105.

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*.

Juan M. Coria, Hervé Bredin, Sahar Ghannay, and Sophie Rosset. 2020. A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification. In *Statistical Language and Speech Processing*, pages 137–148. Springer International Publishing.

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.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834. ISCA.

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*.

Xu Li, Jinghua Zhong, Jianwei Yu, Shoukang Hu, Xixin Wu, Xunying Liu, and Helen Meng. 2020. Bayesian x-vector: Bayesian Neural Network based x-vector System for Speaker Verification. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, pages 365–371.

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

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

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 Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

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.

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.

Suzy J. Styles, Victoria Y. H. Chua, Fei Ting Woon, Hexin Liu, Leibny Paola Garcia, Sanjeev Khudanpur, Andy W. H. Khong, and Justin Dauwels. 2023. Investigating model performance in language identification: beyond simple error statistics. In *Proc. INTERSPEECH 2023*, pages 4129–4133.

Pulkit Verma and Pradip K. Das. 2015. i-Vectors in speech processing applications: a survey. *International Journal of Speech Technology*, 18:529–546.

10. Language Resource References

Agyeman, Nana Ama. 2016. *Documentation of Efutu*. Endangered Languages Archive.

Barlow, Russell. 2018. *Documentation of Ulwa, an Endangered Language of Papua New Guinea*. Endangered Languages Archive.

Cieri, Christopher, et al. 2004. *Fisher English Training Speech Part 1 Speech LDC2004S13*. Linguistic Data Consortium.

Grzech, Karolina. 2018. *Upper Napo Kichwa: A Documentation of Linguistic and Cultural Practices*. Endangered Languages Archive.

Jukes, Anthony. nd. *Documentation of Toratán (Ratahan)*. Endangered Languages Archive.

McGill, Stuart. 2012. *Cicipu Documentation*. Endangered Languages Archive.

M. McLaren and L. Ferrer and D. Castan and A. Lawson. 2016. *The speakers in the wild SITW speaker recognition database*.

NIST Multimodal Information Group. 2011a. *2006 NIST Speaker Recognition Evaluation Training Set LDC2011S09*. Distributed by Linguistic Data Consortium.

NIST Multimodal Information Group. 2011b. *2005 NIST Speaker Recognition Evaluation Test Data LDC2011S04*. Distributed by Linguistic Data Consortium.

NIST Multimodal Information Group. 2011c. *2005 NIST Speaker Recognition Evaluation Training Data LDC2011S01*. Distributed by Linguistic Data Consortium.

Pérez González, Jaime. 2018. *Documentation of Mocho' (Mayan): Language Preservation through Community Awareness and Engagement*. Endangered Languages Archive.