Interspeech 2018
2-6 September 2018, Hyderabad

The ACLEW DiViMe: An easy-to-use diarization tool

Adrien Le Franc[1,2], *Eric Riebling*[3], *Julien Karadayi*[1,2], *Yun Wang*[3], *Camila Scaff*[1], *Florian Metze*[3], *Alejandrina Cristia*[1]

[1] LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[2] Institut national de recherche en informatique et en automatique, Paris, France
[3] Language Technologies Institute; Carnegie Mellon University; Pittsburgh, PA, U.S.A.

adrien.le-franc@eleves.enpc.fr,alecristia@gmail.com

Abstract

We present "DiViMe", an open-source virtual machine aimed at packaging speech technology for real-life data, and developed in the context of the "Analyzing Children's Language Environments across the World" Project. This first release focuses on Speech Activity Detection, Speaker Diarization, and their evaluation. The present paper introduces the set of included tools and the current workflow, which is focused on making minimal assumptions regarding users' technical skills. Additionally, we show how the current DiViMe tools fare against three sets of challenging data. In a first experiment, we look at performance with samples extracted from daylong recordings gathered using the LENA^{TM} system from English-learning children. We find that the performance of the tools currently in DiViMe is not far from that achieved by the LENA^{TM} proprietary software. In a second experiment, we generalize to other samples of child-centered daylong files, gathered with non-LENA^{TM} hardware from non-English-learning children, showing that performance does not degrade in this condition. Finally, we report on performance in the DiHARD 2018 Challenge Test Data. Originally conceived in the "Speech Recognition Virtual Kitchen", DiViMe is a promising platform for packaging speech technology tools for widespread re-use, with potential impact on both fundamental and applied speech and language research.

Index Terms: speech activity detection, speaker diarization, virtual machine, language acquisition, children DiHARD Challenge

1. Introduction

Research projects in linguistics, speech pathology, and other language sciences often collect and compare ecological data from different cultures and settings with a diverse set of acquisition devices. The resulting heterogeneous speech corpora truly deserve the "in the wild" label, and have been shown to test the limitations of even state-of-the-art of speech processing algorithms. The difficulties in processing data such as child speech in a daily-life environment have been highlighted at the 2017 JSALT Summer Workshop at CMU [1], where it became apparent that unconventional speech containing mumble, cry, overlaps and other artifacts required finer models and motivated the organization of the 2018 DIHARD Challenge [2].

As the field advances in solving "hard diarization", there is another limitation that should not be forgotten: the difficulty to deploy cross-platform and user-friendly softwares for linguistic projects. Many recent speech processing algorithms with high performance offer open-source implementations. However, installing and running such code is not always straightforward. In particular, integrating an open-source project into a local processing pipeline is a challenging task since file formats and environment settings might differ from one tool to another. This technical hurdle is a threat to the reproducibility of experiments. Complex tools might lead to excellent performance, but do not benefit the larger scientific community as they should if they cannot be easily applied to reproduce experiments and to build on top of them. These observations motivated us to develop the ACLEW Diarization Virtual Machine - DiViMe for short. DiViMe follows in the Speech Recognition Virtual Kitchen's [3][4] footsteps in that it is a virtual machine (VM) gathering speech processing tools inside a unified computational environment. As a result, it can be deployed on most host computer systems and offers a simple interface to run the integrated models within a global pipeline.

Our main goal is to bring these systems within the reach of the general language scientist, requiring only minimal computing power and programming skills. We are ideally positioned to contribute this because we are part of a large international collaboration grant, "ACLEW: Analyzing Child Language Experiences Around The World" [5]. The scientific goal of this grant is to document patterns of variation and stability in young children's language experiences, and their subsequent development, as documented via daylong recordings. Daylong recordings are particularly interesting for the present project because they present a difficult diarization problem (and in the case of acquisition data, probably the hardest case imaginable), and they are a natural test case for VM use because these data are typically difficult or impossible to share broadly, and thus must be analyzed *in situ*. Additionally, our collaborator network includes some members with very limited or no previous programming experience, allowing us to beta test that instructions are clear and usable. Moreover, much research in this field employs a unified recording device and software toolkit for automatic speech processing developed by the LENA^{TM} Foundation. While this product is not open source, it provides an interesting benchmark to compare our work against since it was specifically designed to process children's speech.

In this paper, we summarize our current progress. At the time of submission, DiViMe contains a set of algorithms which were designed to automatically detect and label speaker turns in naturalistic audio recordings. Two main tasks are distinguished to achieve this goal. A first category of tools perform *Speech Activity Detection* (SAD). The output of such tools is typically a file of time labels with 'speech' or 'non-speech' tags (although for one tool other classes such as 'music' or 'noise' can also be recognized). Once the speech is located in the audio files, a second category of tools can be applied to attribute each occurrence of speech to a specific speaker. This second task is named *Talker Diarization* (TD).

10.21437/Interspeech.2018-2324

2. Description

2.1. Workflow

2.1.1. Installation and application

The VM is designed with Vagrant [6], which is a tool enabling to build and manage virtual machine environments. It comes with a Vagrantfile script which contains the core architecture of the computing system to be deployed. Based on this file, Vagrant runs the virtual environment on top of usual providers such as VirtualBox [7] or Docker [8]. We provide a stable Vagrantfile which enables us to easily build and run a Ubuntu virtual machine isolated from the hosting computer system. The resulting environment runs on any local machine regardless of the hosting OS. It installs all required dependencies to have the speech processing tools introduced in this paper working inside the VM. The only way to commute files between the VM and the local supporting machine is a synced folder enabling to transfer data from the host to the VM and results from the VM back to the host. The basic workflow of the VM is summarized in the schematic diagram of Figure 1.

Once the installation is complete, the tools that the VM provides can be applied to data files on the user's host machine with a series of simple shell commands (e.g., `vagrant ssh -c "tools/TOOLNAME data/"`). We provide users with a detailed README available on our public repository where the software can be downloaded `https://github.com/aclew/DiViMe`

2.1.2. Input and output files

Audio files are expected to be in .wav format. If the user has annotations at either the speech activity or diarization levels, for simplicity we only require the RTTM [9] format. That is, if the user wants to evaluate the SAD performance, then he/she will need to provide the RTTM label for each wav file containing the human-annotated reference annotation. Notice that this gold RTTM can also be provided for the diarization tools, so as to assess talker diarization performance in the absence of SAD errors.

The system returns all annotations in the RTTM format, with the name of the tool that produced them appended to the original file name. Evaluations are returned in a dataframe format, with wavs as rows, and metrics as columns.

2.2. Tools in the current DiViMe release

The current DiViMe builds exclusively on tools that have been developed, documented, and made available by independent researchers. We therefore keep the descriptions very short, and instead provide links to the original resources, where readers will be able to find the full technical descriptions.

We currently provide two options for Speech activity detection (SAD) tools. The first is the **LDC SAD** [10], which relies on HTK [11] to band-pass filter and extract PLP features, prior to applying a broad phonetic class recognizer trained on the Buckeye Corpus [12] using a GMM-HMM model. An official release by the LDC is currently in the works, and should be ready by the time Interspeech is held.

Our second SAD tool will be referred to as **Noiseme SAD** because it draws from a broader "noiseme classifier" [13], a neural network that can predict frame-level probabilities of 17 types of sound events (called "noisemes" [14]), including speech, singing, engine noise, etc. The network consists of one single bidirectional LSTM layer with 400 hidden units in each direction. It was trained on 10 h of HAVIC data [15] with the Theano toolkit which we will change in the future since this framework is no longer maintained. The OpenSMILE toolkit [16] is used to extract 6,669 low-level acoustic features, which are reduced to 50 dimensions with PCA. For our purposes, we summed the probabilities of the classes "speech" and "speech non-english" and labeled a region as speech if this probability was higher than all others.

We currently provide one Talker Diarization (TD) tool. The **DiarTK** model imported in the VM is a C++ open source toolkit [17]. The algorithm first extracts MFCC features, then performs non-parametric clustering of the frames using agglomerative information bottleneck clustering [18]. At the end of the process, the resulting clusters correspond to identified speakers. The most likely Diarization sequence is computed by Viterbi realignement.

Finally, we have evaluation tools for both tasks. For **SAD**, we employ the evaluation included in the LDC SAD [10], which returns the false alarm (FA) rate (proportion of frames labeled as speech that were non-speech in the gold annotation) and missed speech rate (proportion of frames labeled as non-speech that were speech in the gold annotation). For **TD**, we employ the evaluation developed for the DiHARD Challenge [19], which returns a Diarization error rate (DER), which sums percentage of speaker error (mismatch in speaker IDs), false alarm speech (non-speech segments assigned to a speaker) and missed speech (unassigned speech).

One important consideration is in order: What to do with files that have no speech to begin with, or where the system does not return any speech at the SAD stage or any labels at the TD stage. This is not a case that is often discussed in the literature because recordings are typically targeted at moments where there is speech. However, in naturalistic recordings, some extracts may not contain any speech activity, and thus one must adopt a coherent framework for the evaluation of such instances. We opted for the following decisions.

If the gold annotation was empty, and the SAD system returned no speech labels, then the FA $= 0$ and M $= 0$; but if the SAD system returned some speech labels, then FA $= 100$ and M $= 0$. Also, if the gold annotation was not empty and the system did not find any speech, then this was treated as FA $= 0$ and M $= 100$.

As for the TD evaluation, the same decisions were used above for FA and M, and the following decisions were made for mismatch. If the gold annotation was empty, regardless of what the system returned, the mismatch rate was treated as 0. If the gold annotation was empty but a pipeline returned no TD labels (either because the SAD in that system did not detect any speech, or because the diarization failed), then this was penalized via a miss of 100 (as above), but not further penalized in terms of talker mismatch, which was set at 0.

3. Experiments

We conducted several experiments to test and benchmark the SAD and TD tools currently included in DiViMe. To this end, we used 4 datasets, as follows.

- ACLEW Starter-English Plus (ASE+; 3h): The ACLEW Starter dataset [20] contains 11 5-minute clips extracted from as many daylong recordings gathered with a LENA$^{\text{TM}}$ device from English-speaking children growing up in urban areas in the UK [21], the US [22, 23], and Canada [24]. Melanie Soderstrom's team additionally annotated 8 5-minute clips

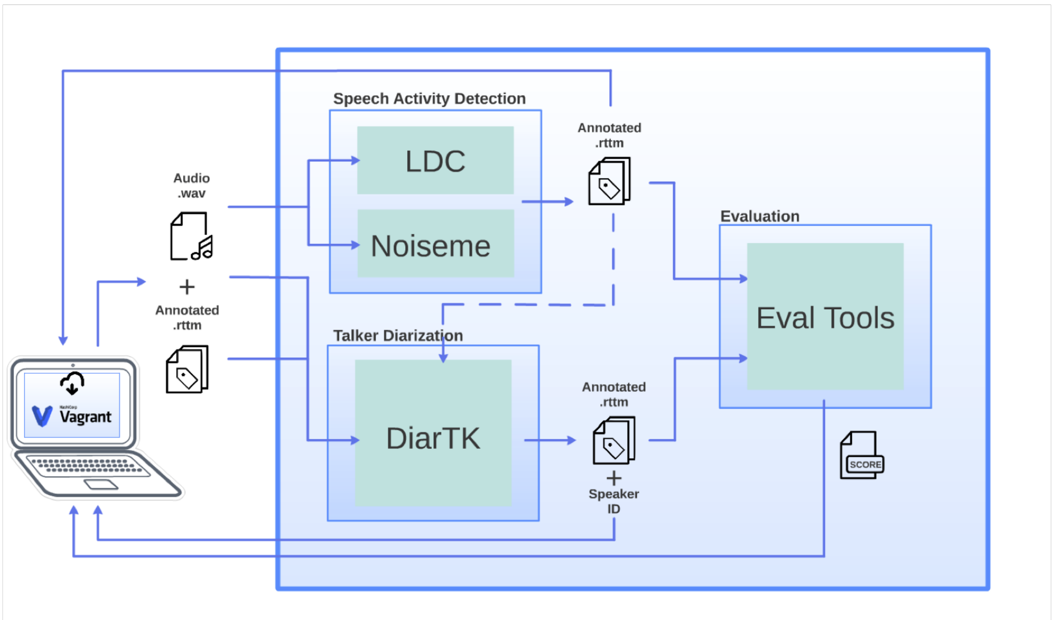

Figure 1: *Schematic diagram of data flow in DiViMe. Significant inputs and outputs (as well as log files) of the individual tools are being read from and written to synced folders of the host computer. Processing is triggered via shell commands, on the host machine.*

from as many recordings [24]. Clips were extracted from regions with a lot of speech. Annotators attempted to label speakers as a function of their individual identity, although they did not know the recorded families.

- Tsimane (9h): A total of 537 1-minute clips were extracted from 1-2 daylong recordings gathered from 27 children learning Tsimane in rural Bolivia [25]. Of these, 227 came from LENA^{TM} recordings (henceforth Tsi-LENA), and the remaining 310 from other devices (USB or Olympus; henceforth Tsi-other). Clips were sampled periodically throughout the day to avoid sampling bias. Speakers were labeled using broad classes (children, female adults, male adults), with the exception of the child wearing the recorder and the most common female adult voice. The annotator did not know the recorded families.

- Casillas (10h): A total of 190 1-, 5-, or 6-minute clips were extracted from daylong recordings gathered from 10 children learning Tseltal in rural Mexico using an Olympus recorder. Some of the clips were extracted randomly throughout the day; others targeted regions with a lot of speech by the child, or a lot of conversational interactions. Annotators knew the recorded families well and were able to label speakers as a function of their individual identity.

- DiHARD (21h): The DiHARD Evaluation data set contains 5-10 minute clips sampled from heterogeneous corpora including recordings similar to those in ASE+ but also meeting data, and many others. More details can be found on the Challenge website [1]. To our knowledge, annotators attempted to label the (unknown) speakers as a function of their individual identity.

Results for SAD at the time of final submission are shown on 1 and 2; those for TD are shown on 3.[2] Recordings not collected with LENA^{TM} hardware cannot be analyzed with the LENA^{TM} software, and thus such combinations are shown as NA below.

Dataset	LDC	Noiseme	LENA
ASE+	46%	9%	16%
Tsi-LENA	63%	17%	76%
Tsi-other	45%	8%	NA
Casillas	34%	8%	NA
DiHARD	15%	15%	NA

Table 1: *False alarm (FA) rates in SAD as a function of the dataset and the SAD tool. Lower is better.*

Dataset	LDC	Noiseme	LENA
ASE+	29%	70%	71%
Tsi-LENA	9%	35%	1%
Tsi-other	12%	33%	NA
Casillas	22%	68%	NA
DiHARD	19%	44%	NA

Table 2: *Miss (M) rates in SAD as a function of the dataset and the SAD tool. Lower is better.*

Dataset	Gold	LDC	Noiseme	LENA
ASE+	53%	182%	114%	143%
Tsi-LENA	106%	174%	121%	263%
Tsi-other	97%	161%	126%	NA
Casillas	89%	186%	146%	NA
DiHARD	58%	65%	72%	NA

Table 3: *DER in TD as a function of the dataset. The LENA column indicates diarization performance for the LENA algorithm as a whole. For all other columns, diarization was done with DiarTK, and the column label indicates the SAD annotation used as input. Gold column gives the results of applying DiarTK to the human annotated SAD. The DiHARD results are as provided by the Challenge organizers. Lower is better.*

3.1. Experiment 1: How well do we fare against the current field standards?

The LENA^{TM} software performs joint segmentation and classification with acoustic models trained with 150 hours of hand-annotated data from English-learning American children growing up in urban settings. It returns a segmentation of the audio into categories: key child, other children, female adult, male adult, TV noise, other noise, silence, and overlap (which is overlap between any of the non-silence categories). For the purposes of our experiments, we declared as non-speech all the non-human categories as well as the speech categories that the

[1] http://coml.lscp.ens.fr/dihard/data.html

[2] At the time of final submission, errors were found in a few of the clips' gold annotation. We are in the process of correcting these, and will post revised results on `https://osf.io/kwdhq/`.

system classified as "far" from their acoustic models, because in pilot analyses the SAD performance was better without than with these "far" items.

To focus on differences that were stable rather than averages like the ones reported on the Tables above, we fit a mixed regression model (in R [26], package lme4 [27]), declaring corpus, system, and their interaction as fixed effects and the clip ID as random effect. Given the question addressed in this experiment, we focus on the two corpora gathered with a LENA$^{\text{TM}}$ device. We declared ASE+ as the baseline for corpus (since it is closer to what the LENA$^{\text{TM}}$ system was developed on), and LENA$^{\text{TM}}$ as the baseline for system. Results are shown on Table 4. Effects of corpus will be discussed in the next subsection. Turning to the current key interest, LDC SAD led to a significantly higher FA and lower Miss rates than LENA$^{\text{TM}}$, whereas Noisemes led to a non-significantly lower FA and higher Miss rates. Given that the reduction in Miss is smaller than the gain in FA with the LDC SAD system, this appears like a competitive alternative to the LENA$^{\text{TM}}$ system, as does Noisemes which performed no better or no worse than LENA^{TM}. The results of the DER analyses, which compound errors over the SAD and TD phases, confirm these conclusions, as neither of our systems differed from the LENA$^{\text{TM}}$ significantly for ASE+, and there was only an interaction between LDC and corpus at t = 2.1.

Predictor	FA	M	DER
(Intercept)	17 (7)*	69 (4)*	138 (64)*
Tsi-lena	59 (7)*	-68 (4)*	-24 (61)
LDC	29 (8)*	-40 (5)*	125 (66)
Noisemes	-8 (8)	1 (5)	44 (61)
LDC*Tsi-lena	-42 (8)*	47 (5)*	-134 (64)*
Noisemes*Tsi-lena	-51 (8)*	33 (5)*	-118 (64)

Table 4: *Mixed model regressions predicting performance from the system, corpus, and their interaction. Each cell shows the estimate (and its standard error) corresponding to the crossing of the predictor and the dependent variable. An asterisk indicates an effect with t > 2.*

3.2. Experiment 2: How well do tools do with audio collected with other devices and untrained populations?

The main effect of corpus in Table 4 shows a significantly higher FA and significantly lower M for Tsi-lena than ASE+, due to the fact that there was a great deal more silence in the former files (in fact, nearly half of the Tsimane clips had no speech in them). This effect is caused by the Tsimane clips being randomly sampled throughout the day and night, whereas the ACLEW Starter set clips were selected because there was speech in them.

To provide a broader picture, we fit another mixed model predicting DER (which represents global performance for a given pipeline), this time with the 4 child corpora. As before, fixed effects were corpus, method, and their interaction, with baseline levels ASE+ and LENA$^{\text{TM}}$. Only two of the interactions (LDC*Tsi-LENA, and Noisemes*Tsi-LENA, indicating lower DERs in this corpus when our systems were used rather than the LENA system) had t>2, suggesting that all systems performed similarly to each other and across corpora. As for the impact of the hardware, the results of Tsi-LENA and Tsi-other (recorded with non-LENA devices) do not highlight better performances on Tables 1, 2, 3 when using the LENA hardware.

3.3. Experiment 3: Benchmarking against the DiHARD Challenge (data)

We had two goals by using the DiHARD Challenge data. First, the performance of the same tools across our child language acquisition data versus the DiHARD data indirectly speaks to how comparably difficult our datasets are. The DiHARD test data contains an heterogeneous mix of data, whereas all of the other datasets we tested here are children-centered, collected in a completely ecological fashion. We observe that DER is higher for the non-DiHARD datasets than the DiHARD Challenge dataset, regardless of the tool.

Second, we can compare the tools in DiViMe against the leaderboard of the Challenge on the DiHARD data so as to assess to what extent our tools are competitive. Our primary purpose was to offer a quick and easy access to speech processing tools to conduct research. Therefore, we did not expect the tools we introduced so far to outperform the state-of-the-art of SAD and TD. This expectation was confirmed: Our systems score at the bottom of the DiHARD chart for both tracks. This is not only the case due to our SAD being underperforming, as clear from the fact that TD with gold SAD still led to a very high error rate. However, we did not retrain our tools on our testing datasets to reflect an "out of the box" use of the VM. While we feel that DiViMe fits its function in terms of usability, we look forward to incorporating better-performing SAD and TD tools in the future.

4. Conclusions

We presented a Virtual Machine that almost anybody can use to detect speech segments using various advanced techniques.We outlined the VM's use, its internals, and provided pointers to currently available algorithms. Our benchmarks showed that ecological language acquisition data are particularly hard even when compared with the DiHARD Challenge data. We would look forward to integrating better-performing SAD and TD systems. In next steps, we will incorporate models that can be retrained inside the VM. In the meanwhile, for several tasks and dataset combinations, we remain competitive against the LENA$^{\text{TM}}$, which is the current go-to system in the language acquisition field, making DiViMe a competitive open-source solution for this audience. Additionally, the algorithms currently included are robust to variation in the recording hardware used and the population from which data are collected, which are crucial features for our target users. In sum, DiViMe is a promising tool that makes complex processing models accessible to non-technical users.

5. Acknowledgements

This work was supported by a Trans-Atlantic Platform "Digging into Data" collaboration grant (ACLEW: Analyzing Child Language Experiences Around The World), with the support of Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW; ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC) and the National Endowment for the Humanities (HJ-253479-17), as well as funding from the J. S. McDonnell Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). We also directly benefited from interactions at the 2017 Frederick Jelinek Memorial Summer Workshop, which was supported by Amazon, Apple, Facebook, Google, and Microsoft (in alphabetical order). We are grateful to our beta testers, particularly Okko Räsänen.

6. References

[1] N. Ryant, E. Bergelson, K. Church, A. Cristia, J. Du, S. Ganapathy, S. Khudanpur, D. Kowalski, M. Krishnamoorthy, R. Kulshreshta, M. Liberman, Y.-D. Lu, M. Maciejewski, F. Metze, J. Profant, L. Sun, Y. Tsao, and Z. Yu, "Enhancement and analysis of conversational speech: JSALT 2017," in *Proc. ICASSP*. Calgary, BC; Canada: IEEE, Apr. 2018, accepted.

[2] "The first DIHARD speech diarization challenge," https://coml.lscp.ens.fr/dihard/index.html, accessed: 2018-06-17.

[3] A. Plummer, E. Riebling, A. Kumar, F. Metze, E. Fosler-Lussier, and R. Bates, "The speech recognition virtual kitchen: Launch party," in *Proc. INTERSPEECH*. Singapore: ISCA, Sep. 2014, http://www.speechkitchen.org/.

[4] "The speech recognition virtual kitchen," https://github.com/srvk, accessed: 2018-06-17.

[5] "ACLEW - analyzing child language experiences around the world," https://sites.google.com/view/aclewdid/home, accessed: 2018-06-17.

[6] "Vagrant by hashicorp," https://www.vagrantup.com/, accessed: 2018-06-17.

[7] "Oracle vm virtualbox," https://www.virtualbox.org/, accessed: 2018-06-17.

[8] "Docker - build, ship and run any app, anywhere," https://www.docker.com/, accessed: 2018-06-17.

[9] "An object oriented description of speech for EARS," https://catalog.ldc.upenn.edu/docs/LDC2004T12/RTTM-format-v13.pdf, accessed: 2018-06-17.

[10] N. Ryant, "Ldc sad," https://github.com/Linguistic-Data-Consortium, accessed: 2018-06-17.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[12] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[13] Y. Wang and F. Metze, "A first attempt at polyphonic sound event detection using connectionist temporal classification," in *Proc. ICASSP*. New Orleans, LA; U.S.A.: IEEE, Mar. 2017.

[14] S. Burger, Q. Jin, P. F. Schulam, and F. Metze, "Noisemes: Manual annotation of environmental noise in audio streams," Carnegie Mellon University, Pittsburgh, PA; U.S.A., Tech. Rep. CMU-LTI-12-07, 2012.

[15] S. Strassel, A. Morris, J. G. Fiscus, C. Caruso, H. Lee, P. D. Over, J. Fiumara, B. L. Shaw, B. Antonishek, and M. Michel, "Creating havic: Heterogeneous audio visual internet collection," in *Proc. LREC*. Istanbul, Turkey: ELRA, May 2012.

[16] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[17] D. Vijayasenan and F. Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[18] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 250–255.

[19] N. Ryant, "Diarization evaluation," https://github.com/nryant/dscore, accessed: 2018-06-17.

[20] E. Bergelson, A. Warlaumont, A. Cristia, M. Casillas, C. Rosemberg, M. Soderstrom, F. Metze, E. Dupoux, O. Rasanen, C. Rowland, and S. Durrant, "Starter-aclew," 2017, accessed: 2018-06-17. [Online]. Available: http://databrary.org/volume/390

[21] C. Rowland, A. Bidgood, S. Durrant, M. Peter, and J. M. Pine, "The language 0-5 project corpus," 2016.

[22] E. Bergelson, "Bergelson seedlings homebank corpus," 2016.

[23] A. S. Warlaumont and G. M. Pretzer, "Warlaumont homebank corpus," 2016.

[24] K. McDivitt and M. Soderstrom, "Mcdivitt homebank corpus," 2016.

[25] C. Scaff, J. Stieglitz, and A. Cristia, "Daylong recordings from young children learning Tsimane in Bolivia," https://nyu.databrary.org/volume/445, accessed: 2018-03-03.

[26] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[27] D. Bates, M. Maechler, B. Bolker, S. Walker *et al.*, "lme4: Linear mixed-effects models using eigen and s4," *R package version*, vol. 1, no. 7, pp. 1–23, 2014.