A comparison study on patient-psychologist voice diarization

Rachid Riad * and **Hadrien Titeux*** and **Xuan Nga Cao** and **Emmanuel Dupoux**
CoML, ENS/CNRS/EHESS/INRIA/PSL Research University

Laurie Lemoine and **Justine Montillot** and **Agnes Sliwinski**
and **Jennifer Hamet Bagnou** and **Anne-Catherine Bachoud-Lévi**
NPI, ENS/INSERM/UPEC/HD CENTER/PSL Research University

Abstract

Conversations between a clinician and a patient, in natural conditions, are valuable sources of information for medical follow-up. The automatic analysis of these dialogues could help extract new language markers and speed up the clinicians' reports. Yet, it is not clear which model is the most efficient to detect and identify the speaker turns, especially for individuals with speech disorders. Here, we proposed a split of the data that allows conducting a comparative evaluation of different diarization methods. We designed and trained end-to-end neural network architectures to directly tackle this task from the raw signal and evaluate each approach under the same metric. We also studied the effect of fine-tuning models to find the best performance. Experimental results are reported on naturalistic clinical conversations between Psychologists and Interviewees, at different stages of Huntington's disease, displaying a large panel of speech disorders. We found out that our best end-to-end model achieved 19.5% IER on the test set, compared to 23.6% achieved by the finetuning of the X-vector architecture. Finally, we observed that we could extract clinical markers directly from the automatic systems, highlighting the clinical relevance of our methods.

1 Introduction

During the last decades, it became easier to collect large naturalistic corpora of speech data. It is now possible to obtain new realistic measurements of turn-takings and linguistic behaviours (Ash and Grossman, 2015). These measurements can be especially useful during clinical interviews as they augment the current clinical panel of assessments and unlock home-based assessments (Matton et al., 2019). The remote automatic measure of symptoms of patients with Neurodegenerative diseases could greatly improve the follow-up of patients and speed-up ongoing clinical trials.

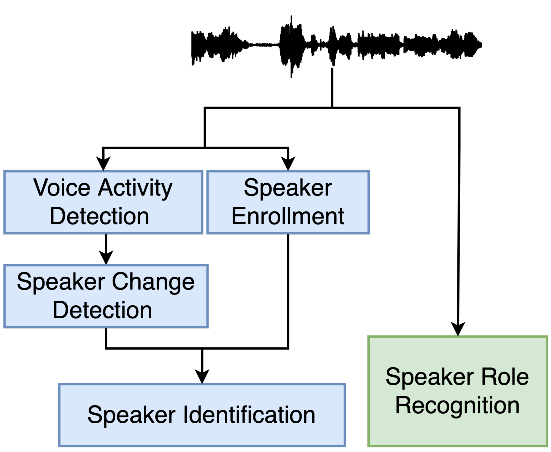
Yet, this methodology relies on the heavy burden of manual annotation to reach the necessary amount needed to draw significant conclusions. It is now indispensable to have robust speech processing pipelines to extract meaningful insights from these long naturalistic datasets (Lahiri et al., 2020). Huntington's Disease represents a unique opportunity to design and test these speech algorithms for *Neurodegenerative diseases*. Indeed, individuals with the Huntington's disease can exhibit a large spectrum of *speech and language* symptoms (Vogel et al., 2012) and it is possible to follow gene carriers even before the official clinical onset of the disease (Hinzen et al., 2018). The first unavoidable computational tasks to extract speech and linguistic information from medical interviews is the diarization: (1) the *detection* of speaker-homogeneous portions of voice activity (Graf et al., 2015) and (2) the *identification* of speaker (Bigot et al., 2010). Speaker turns are clinically informative for diagnostic in Huntington's Disease (Perez et al., 2018; Vogel et al., 2012).

First, a number of studies are trying to solve this problem directly from the audio signal and linguistic outputs, also referred to as *Speaker Role Recognition*. They are taking advantage of the specificities (ex: prosody, specific vocabulary, adapted language models) of each role in the different domains: Broadcast news programs (Bigot et al., 2010), Meetings (Sapru and Valente, 2012), Medical conversations (Flemotomos et al., 2018), Child-centered recordings (Lavechin et al., 2020;

* Equal contribution. We are very thankful to the patients that participated in our study. We thank Katia Youssov, Laurent Cleret de Langavant, Marvin Lavechin, and the speech pathologists for the multiple helpful discussions and the evaluations of the patients. This work is funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and Grants from Neuratris, from Facebook AI Research (Research Gift), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).

Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022), pages 30 - 36

Figure 1: Two approaches for the diarization of conversational clinical interviews. The steps for the Speaker Enrollment Protocol are in Blue, and Green for the Speaker Role Recognition.



Koluguri et al., 2020).

Another approach relies on *Speaker Enrollment* (Snyder et al., 2017; Heigold et al., 2016), it aims to check the identity of a given speech segment based on a enrolled speaker template. Our study differ from these studies as they are evaluating their pipelines with already segmented speaker-homogeneous speech segments. Another related approach is *Personal VAD* (Voice Activity Detection) model from (Ding et al., 2020) where they used enrolled speaker template to detect speech segments from each individual speaker.

None of these approaches have been compared under the same evaluation metric, despite prior works aiming at solving both these tasks (García et al., 2019) and their high degree of similarities.

Here in this paper, we aimed to *detect* automatically the portions of speech and to *identify* the speakers in medical conversation between Psychologists and Interviewees. These interviewees are either Healthy Controls (C), gene carriers without overt manifestation of Huntington's Disease (preHD) and manifest gene carriers of Huntington's Disease (HD). We introduced a novel way to split the datasets so that we are now capable to compare two different speech processing approaches to deal with these 2 problems (Figure 1): *Speaker Role Recognition* and *Speaker Enrollment Protocol*. We showed the clinical relevance of these pipelines with the extraction speech markers that have been found predictive in Huntington's Disease.

2 Data, evaluation splits, metrics

2.1 Dataset

Ninety four participants were included from two observational cohorts (NCT01412125 and NCT03119246) in this ancillary study at the Hospital Henri-Mondor Créteil, France): 72 people tested with a number of CAG repeats on the Huntingtin gene above 35 (CAG > 35), and 22 Healthy Controls (C). Mutant Huntington gene carriers were considered premanifest if they both score less than five at the Total Motor score (TMS) and their Total functional capacity (TFC) equals 13 (Tabrizi et al., 2009) using the Unified Huntington Disease Rating Scale (UHDRS). All participants signed an informed consent and conducted an interview with an expert psychologist. Therefore in the diarization setting, there are two roles in each interview: a *Psychologist* and an *Interviewee*. The speech data were annotated with Seshat (Titeux* et al., 2020) and Praat (Boersma et al., 2002) softwares. The dataset is composed of $K = 94$ interviews $\mathcal{I}_{1...K}$. We designed a new way to split of speech dataset to compare different diarization approaches: an end-to-end Speaker Role Recognition model and a Speaker Enrollment pipeline (See Figure 2). The dataset is split in three sets which we refer to *meta-train set* M_{train}, *meta-dev set* M_{dev} and *meta-test set* M_{test} with the ratio of 60%, 20%, and 20%, respectively. Interview $I \in \mathcal{I}_{1...K}$ is composed of N_I segments $I = \{U_0, U_2, \ldots, U_{N_I}\}$. Each segment U_i is pronounced by a speaker s_i. We summarized the corpus statistics in Table 1.

Each interview I in the *meta-dev* and *meta-test* is split in two sets which we refer *dev set* X_{dev} and *test set* X_{test}. X_{test} is always kept fixed through all experiments, and we study the influence of the size of the X_{dev} based on T_{dev} that filters the segments (cf Figure 2).

All the data from the *meta-train* set M_{train} is used to train or fine-tune the neural network models for voice activity detection, speaker change detection, speaker role recognition, and speaker enrollment. The dev set X_{dev} of the *meta-dev* set M_{dev} and the dev set X_{dev} of the *meta-test* set M_{test} are only used for the speaker enrollment experiments, to build the template representation of each speakers. The results on the test set X_{test} of the *meta-dev* set M_{dev} are used to select all the hyper-parameters and select the best model for each experiment. The final comparison is done with the test set X_{test} of the *meta-test* set M_{test}.

Figure 2: Illustration of the data split with 4 interviews. Each line I_i represents an interview between the Interviewee and the Psychologist. The elevation of each row indicates 'who speaks when'. The segments can overlap.

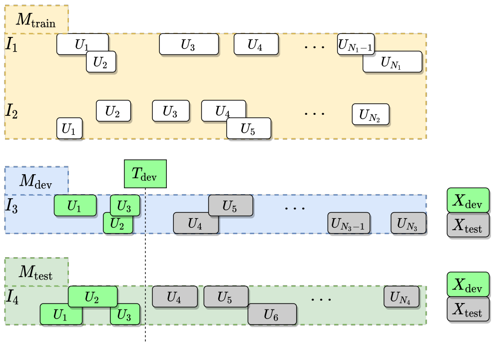


Table 1: Corpus statistics. P stands for Psychologist. IT stands for Interviewee. *Dur* stands for Duration and reported in hour. Durations are reported in hours.

	M_{train}	M_{dev}	M_{test}
#Interviews	57	18	19
#Segments IT	21400	7503	7788
#Segments P	4184	1381	1517
Dur Role IT	7.65	3.02	3.21
Dur Role P	3.54	1.14	1.15
Dur Overlap	1.10	0.50	0.45
C/preHD/HD	13/11/33	4/4/10	5/3/11

2.2 Metrics

To compare final performance of each approach, we use the Identification Error Rate (IER) taking into account both the segmentation and confusion errors. IER is obtained with `pyannote.metrics` (Bredin, 2017):

$$\text{IER} = \frac{T_{\text{false alarm}} + T_{\text{missed detection}} + T_{\text{confusion}}}{T_{\text{Total}}}$$

The $\frac{T_{\text{confusion}}}{T_{\text{Total}}}$ component in the IER is related to the Miss-classification Rate (MR%) used in Speaker Role Recognition study (Flemotomos et al., 2019), which is based on Frames and not duration of the turns. We compared the different approaches as a function of the size of the enrollment T_{dev} in Figure 3.

3 Methods

3.1 Speaker Role Recognition

We adapted the approach from (Lavechin et al., 2020) for the Speaker Role Recognition. We trained on M_{train} a unique model to detect each role (Psychologist,Interviewee), and selects the best epoch on M_{dev}. This is a multi-label multi-class segmentation problem. A threshold parameter for each role is optimized on the Meta-dev set M_{dev} for the two output units of the model. Therefore the two classes can be activated at the same time, i.e. we can also detect overlapped speech. To solve and model this task, we used SincNet filters (Ravanelli and Bengio, 2018) to obtain adapted speech features vectors from the audio signal. The SincNet output is fed to a stack of 2 bi-recurrent LSTM layers with hidden size of 128, then pass to a stack of 2 feed-forward layers of size 128 before a final decision layer. We used a binary cross-entropy loss and a cyclic scheduler as training procedure. The hyper-parameters to train our model can be found here [1].

3.2 Speaker enrollment protocol

The Speaker enrollment protocol can be decomposed into four tasks: (1) Voice Activity Detection (2) Speaker Change Detection, (3) Enrollment, (4) Identification. We extended the speech processing toolkit from (Bredin et al., 2020) `pyannote.audio` to run our experiments. Clinical laboratories can not all re-train in-domain speech processing models due to data scarcity or a lack of computational resources. Therefore, we evaluated pretrained models on open-source datasets and transfer models on our dataset to evaluate these out-of-domain performances with real clinical conversational conditions.

3.2.1 Voice Activity Detection

The first step is the Voice Activity Detection (VAD), i.e. obtain the speech segments in the audio signal. It can be modeled as an audio sequence labeling task. There are 2 classes (Speech or Non-Speech). The VAD labels for each interview I are the presence or not of a segment U_i at time t.

The model can be used already *Pretrained* or *Retrained* on the meta-train set M_{train} of our dataset. We choose the DIHARD dataset (Ryant et al., 2019) as a potential pretrained dataset as it contains multiple source domain data (clinical interviews among them). When trained from scratch, the training is done for 200 pyannote epochs and the model is selected on the Meta-dev M_{dev}. The model is also composed of SincNet filters with 2 bi-recurrent LSTM layers and 2 feed-forward layers. The full

[1] `https://tinyurl.com/etfrky3w`

specifications can be found here[2].

3.2.2 Speaker Change Detection

The second step is the Speaker Change Detection (SCD), i.e. obtain the moment when one of a speaker starts or stops talking. It can aslo be modeled as an audio sequence labeling task. There are 2 classes (Change or No-Change). The SCD labels for each interview I are the start or end of a segment U_i at time t. We also compared *Pretrained* on DIHARD and *Retrained* models. We used the same model as for the Voice Activity Detection. The full specifications can be found here.

Based on VAD and SCD outputs, for each Interview I we obtain a set of N'_I candidates speaker-homogeneous segments $\{\hat{U}_1, \dots \hat{U}_{N'_I}\}$.

3.2.3 Enrollment

In the enrollment stage, we need to get a Speaker Embedding function f_θ for our specific task. We combined SincNet filters and the X-vector architecture (Snyder et al., 2017) as in (Bredin et al., 2020). For finetuning, we froze all layers and finetuned the last layer. We used the VoxCeleb2 dataset (Nagrani et al., 2017) as a pretraining dataset as it contains a diverse distribution of speakers and recording conditions.

Then, we used the set of segments from the dev set X_{dev} of the *meta-dev* and *meta-test* to build a template vector m_j for each speaker j in the interview I. X_{dev} contain a set of segments $U_{\text{enrollment speaker } j}$ from each speaker j. The start of each segment $U_{\text{enrollment speaker } j}$ needs to be smaller than T_{dev}. We computed the average of the representations for each speaker j:

$$m_j = \frac{1}{|U_{\text{enrollment speaker } j}|} \sum_{U \in U_{\text{enrollment speaker } j}} f_\theta(U) \tag{1}$$

In principle, the more data you have to build template of each speaker, the easier it is to distinguish them. Thus, we studied the effect of the size of the enrollment based on the parameter $T_{dev} \in (90s, 100s, \dots, 180s)$ to build the template m_j (Larcher et al., 2014).

3.2.4 Identification

For the identification stage, we use the function f_θ and the different representation m_j of the speakers from the enrollment stage. We used the following cosine distance D to build a scoring function and compare each segment $\hat{U} \in \{\hat{U}_1, \dots \hat{U}_{N'_I}\}$ to each template m_j:

$$D(\hat{U}, m_j) = \frac{1}{2}\left(1 - \frac{f_\theta(\hat{U})^\top m_j}{\left[\|f_\theta(\hat{U})\| \, \|m_j\|\right]}\right) \tag{2}$$

$$\text{argmin}_j \, D(\hat{U}, m_j) : \text{Selects Speaker } j \tag{3}$$

In addition, we analysed topline performance of the speaker embedding models when the Ground Truth Segmentation is provided. Finally, we computed a chance baseline based on speaker Enrollment by randomly permutating all the cosine distances. Spearman correlation is computed to compare clinical markers extracted from our best system to ground truth extractions (Figures 4 and 5).

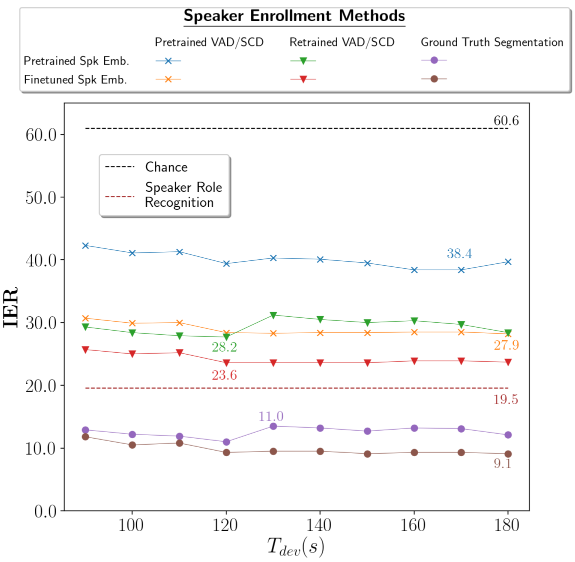


Figure 3: Identification Error Rates for the different combination of approaches on the test set X_{test} of the meta-test set M_{test} as a function of the size of the enrollment parameter T_{dev}. *Spk Emb.*, *VAD,SCD* stand for Speaker Embedding, Voice Activity Detection and Speaker Change Detection. Best performance of each approach is displayed at the best T_{dev}.

4 Results and discussions

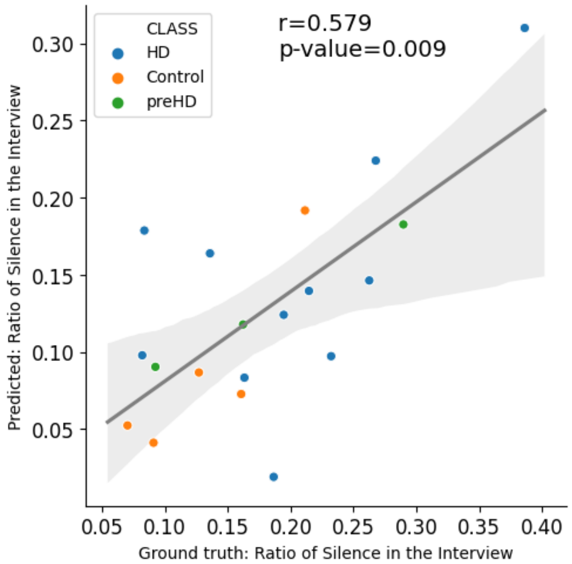
Figure 3 shows results in term of IER for the different approaches. Both approaches greatly improved over chance. If we consider pipelines solving both segmentation and identification, our best performance is obtained using the Speaker Role Recognition approach with IER=19.5% while the Speaker

[2] https://tinyurl.com/44677f7c

Table 2: Speaker Role Recognition Ablation study: Identification Error Rates on the test set X_{test} of the meta-test set M_{test} as a function of the percentage of interview in the meta-train set M_{train}. MD stands for Missed detection, FA for False Alarm and Conf. for Confusion

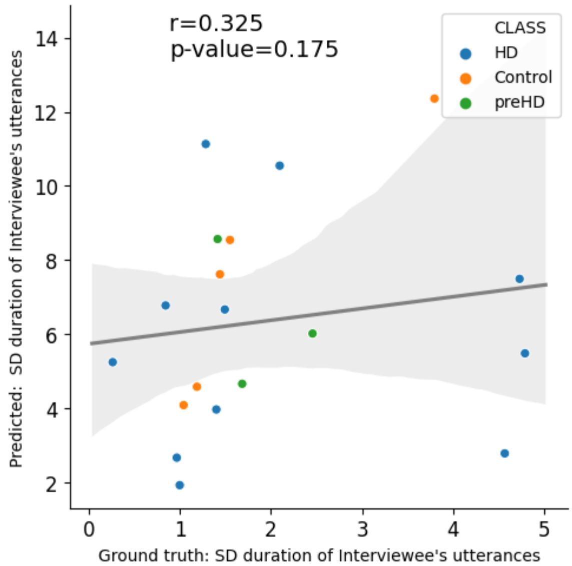
% of M_{train}	MD	FA	Conf.	IER
10%	8.0	14.5	3.9	26.5
20%	7.8	12.4	3.8	24.0
50%	7.5	10.4	2.5	20.7
100%	7.1	10.2	2.3	19.5

Figure 4: Ratio of Silence from the Ground truth segmentation and from the best Speaker role recognition pipeline.



Enrollment Protocol obtained at best IER=23.6% at $T_{dev} = 120s$, with Retrained VAD/SCD models and Finetuned Speaker Embedding. Even though, the Speaker Enrollment protocol has per-speaker templates, it is not surpassing the Speaker Role Recognition approach. The topline with Ground Truth Segmentation (IER=9.1%) indicated that Speaker Enrollment could benefit greatly from a better detection of speaker-homogeneous turns. Errors of Speaker Enrollment are accumulated through the steps and can not be recovered, while Speaker Role Recognition takes advantage of solving all steps together in an end-to-end approach. Increasing the size of the Template Enrollment m_j for each speaker with T_{dev} lead to slight improvements to all Speaker Enrollment methods. The finetuning of the X-vector speaker embedding model with in-domain is especially crucial (ex: Based on retrained VAD/SCD the IER decreases from 28.2%

Figure 5: Standard Deviations (SD) of the Duration of Utterances of Interviewees from the Ground truth segmentation and the best Speaker role recognition system.



to 23.6%). We ran an additional ablation experiment (Table 2) for the Speaker Role Recognition to measure the amount of data necessary. This ablation study informed us on the necessary amount of data to reach certain level of performance. Even though models are better than Chance, we found out that at least 50% of our dataset (28 Interviews) is necessary to outperform the Speaker Enrollment Protocol pipeline (IER of 20.7% vs 23.6%). The analysis of the pattern of errors showed that the most important component is the False Alarm (FA), and a tenfold increase in dataset size allows to gain 4 points of FA. Therefore, most of the errors come from the voice activity detection part of the system. One of our hypothesis is that the system is confused by too much ambient noises from the hospital environment and thus potentially trigger too much positive presence of speech.

In previous studies in Huntington's Disease (Vogel et al., 2012; Perez et al., 2018), the Ratio of Silence and Statistics on utterances were informative to distinguish between classes of Individuals. These speech markers can be extracted directly from the predictions of the Speaker Role Recognition outputs. We computed the Ratio of Silence and the Standard Deviation of Duration of Utterances on the test set of the Meta-test set M_{test}. This computation was done both from the Ground Truth Segmentation and the segmentation provided by the Speaker role recognition system (Figures 4, 5. We observed that the automatic system outputs behaved differently as a function of clinical marker. The Ratio of Silence was better predicted (significant spearman correlation of $r = 0.579, p = 0.009$) than the SD of Duration of Utterances (non significant spearman correlation of $r = 0.325, p = 0.175$). One potential interpretation of our results is that the difference between the ratio and the standard deviation reveals that our pipeline is great overall to obtain summary statistics of the interview, but its precision at the turn-taking level is not sufficient to obtain turn statistics. Some bias of the predictive system might not hurt the IER metric but hurt the reliability of some clinical measures.

5 Conclusion and future work

Detection and Identification of speaker turns are fundamental problems in speech processing, especially in healthcare applications. While works studying these problems in isolation has provided valuable insights, in this work, we showed that Speaker Role Recognition was the most suitable approach for Interviewees at different stages of Huntington's Disease. For future work, we plan to investigate the use of these methods to derive robust biomarkers automatically and compare them to more classic approaches (Riad et al., 2020; Perez et al., 2018; Romana et al., 2020).

References

Sharon Ash and Murray Grossman. 2015. Why study connected speech production. *Cognitive neuroscience of natural language use*, pages 29–58.

Benjamin Bigot, Julien Pinquier, Isabelle Ferrané, and Régine André-Obrecht. 2010. Looking for relevant features for speaker role recognition. In *Eleventh Annual Conference of the International Speech Communication Association*.

Paul Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5.

Hervé Bredin. 2017. pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Interspeech*, Stockholm, Sweden.

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*, pages 7124–7128. IEEE.

Shaojin Ding, Quan Wang, Shuo-Yiin Chang, Li Wan, and Ignacio-Lopez Moreno. 2020. Personal vad: Speaker-conditioned voice activity detection. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pages 433–439.

Nikolaos Flemotomos, Panayiotis Georgiou, David C Atkins, and Shrikanth Narayanan. 2019. Role specific lattice rescoring for speaker role recognition from speech recognition outputs. In *ICASSP*, pages 7330–7334. IEEE.

Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, and Shrikanth Narayanan. 2018. Combined speaker clustering and role recognition in conversational speech. *Proc. Interspeech 2018*, pages 1378–1382.

Paola García, Jesus Villalba, Hervé Bredin, Jun Du, Diego Castan, Alejandrina Cristia, Latane Bullock, Ling Guo, Koji Okabe, Phani Sankar Nidadavolu, et al. 2019. Speaker detection in the wild: Lessons learned from jsalt 2019.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. 2015. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(1):91.

Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. 2016. End-to-end text-dependent speaker verification. In *ICASSP*, pages 5115–5119. IEEE.

Wolfram Hinzen, Joana Rosselló, Cati Morey, Estela Camara, Clara Garcia-Gorro, Raymond Salvador, and Ruth de Diego-Balaguer. 2018. A systematic linguistic profile of spontaneous narrative speech in presymptomatic and early stage huntington's disease. *Cortex*, 100:71–83.

Nithin Rao Koluguri, Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan. 2020. Meta-learning for robust child-adult classification from speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8094–8098. IEEE.

Rimita Lahiri, Manoj Kumar, Somer Bishop, and Shrikanth Narayanan. 2020. Learning domain invariant representations for child-adult classification from speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6749–6753. IEEE.

Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li. 2014. Text-dependent speaker verification: Classifiers, databases and rsr2015. *Speech Communication*, 60:56–77.

Marvin Lavechin, Ruben Bousbib, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2020. An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*.

Katie Matton, Melvin G McInnis, and Emily Mower Provost. 2019. Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. *Proc. Interspeech*, pages 1438–1442.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *Telephony*, 3:33–039.

Matthew Perez, Wenyu Jin, Duc Le, Noelle Carlozzi, Praveen Dayalu, Angela Roberts, and Emily Mower Provost. 2018. Classification of huntington disease using acoustic and lexical features. In *INTERSPEECH*, volume 2018, pages 1898–1902.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE.

Rachid Riad, Hadrien Titeux, Laurie Lemoine, Justine Montillot, Jennifer Hamet Bagnou, Xuan Nga Cao, Emmanuel Dupoux, and Anne-Catherine Bachoud-Lévi. 2020. Vocal markers from sustained phonation in huntington's disease. *arXiv preprint arXiv:2006.05365*.

A Romana, J Bandon, N Carlozzi, A Roberts, and EM Provost. 2020. Classification of manifest huntington disease using vowel distortion measures. In *Interspeech*, volume 2020, pages 4966–4970.

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. *Proc. Interspeech*, pages 978–982.

Ashtosh Sapru and Fabio Valente. 2012. Automatic speaker role labeling in ami meetings: recognition of formal and social roles. In *ICASSP*, pages 5057–5060. IEEE.

David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. 2017. Deep neural network embeddings for text-independent speaker verification. *Proc. Interspeech*, pages 999–1003.

Sarah J Tabrizi, Douglas R Langbehn, Blair R Leavitt, Raymund AC Roos, Alexandra Durr, David Craufurd, Christopher Kennard, Stephen L Hicks, Nick C Fox, Rachael I Scahill, et al. 2009. Biological and clinical manifestations of huntington's disease in the longitudinal track-hd study: cross-sectional analysis of baseline data. *The Lancet Neurology*, 8(9):791–801.

Hadrien Titeux*, Rachid Riad*, Xuan-Nga Cao, Nicolas Hamilakis, Kris Madden, Alejandrina Cristia, Anne-Catherine Bachoud-Lévi, and Emmanuel Dupoux. 2020. Seshat: A tool for managing and verifying annotation campaigns of audio data. In *LREC*, Marseille. * Equal contribution.

Adam P Vogel, Christopher Shirbin, Andrew J Churchyard, and Julie C Stout. 2012. Speech acoustic markers of early stage and prodromal huntington's disease: a marker of disease onset? *Neuropsychologia*, 50(14):3273–3278.