Gesellschaft für Informatik e.V. (GI) GI. (Hrsg.): INFORMATIK 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 11

SCALA-Speech: An Interactive System for Finding and Analyzing Speech Content in Audio Data

Alessia Cornaggia-Urrigshardt[1], Nikita Jarocky[1], Frank Kurth[1], Sebastian Urrigshardt[1], Kevin Wilkinghoff[1]

Abstract: Audio data does not contain as much static information as images and texts and thus analyses inherently require more time. Although in monitoring applications it is likely that large quantities of the captured audio files do not contain meaningful information, without prior knowledge investigators need to listen to all audio files in full length. In this work, a system for automatically finding and analyzing speech content in audio data is presented. The system provides different speech processing algorithms as well as a graphical interface (SCALA) for assisting investigators in audio analysis. The system consists of four components: speech detection, language recognition, speaker diarization/recognition and keyword spotting. SCALA-Speech structures audio data by recognizing speech regions, used languages and speaker changes, thus enabling investigators to listen to audio data more efficiently. Furthermore, specific speakers and keywords can be annotated and searched for. Usage of SCALA-Speech is demonstrated on audio tracks of videos linked in Twitter posts related to an exemplary topic.

Keywords: Audio Monitoring; Speech Detection; Language Recognition; Speaker Diarization; Keyword Spotting; Deep Learning

1 Introduction

The general goal of processing digital data in law enforcement is to gather information about potential crimes or suspects and collect potential evidence. However, analyzing large amounts of digital data requires substantial amounts of manpower and time although most data files or large parts of them do not contain any interesting information. Analyzing audio data such as speech is even more costly because it does not contain static information and thus listening to it requires time proportional to its length. Additionally, there are several difficulties specific to speech data: First, multiple languages can be encountered in which a particular investigator may lack proficiency. Furthermore, it is difficult for a human to remember and correctly recognize a larger set of unfamiliar speakers. Hence, an interactive system for assisting investigators in analyzing audio data is vital.
There are several use cases for processing audio data in law enforcement. One prominent example is wire-tapping telephone lines [Di21] and more generally fighting organized

[1] Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Fraunhoferstraße 20, 53343 Wachtberg, Germany, alessia.cornaggia-urrigshardt@fkie.fraunhofer.de,nikita.jarocky@fkie.fraunhofer.de,frank.kurth@fkie.fraunhofer.de,sebastian.urrigshardt@fkie.fraunhofer.de,kevin.wilkinghoff@fkie.fraunhofer.de

crime by identifying members and the connections between them to gather insights on the organization's structure as done in the *ROXANNE* project [RO]. To robustly recognize speakers, other components such as language, accent, gender and age identification can be utilized as done in the *Speaker Identification Integrated Project (SIIP)* [Kh18]. In both projects, field tests focusing on a user-centered design have been conducted with several law enforcement agencies through INTERPOL. Speaker diarization systems can also be used for post-processing police interviews [AF12]. Another example is the usage of surveillance systems in public places [Cr16]. These systems usually focus on acoustic events such as gun shots, breaking glass and screams and also try to localize sources.
Since annotating audio data is important for obtaining reliable training data in many applications, appropriate tools for such a task have been developed. Commercial tools, for instance, are provided by *appen* [Ap] or *Shaip* [Sh]. The appen audio annotation tool supports users by automatically segmenting audio files and allowing to add time stamps and transcriptions. Shaip offers a variety of label tools for text, image, video, and audio annotations. With SCALA-Speech, we propose a tool which allows for both annotating audio data as well as systematically analyzing its content by providing a variety of speech processing algorithms, all of which combined in a single speech analysis tool.
The contributions of this paper are the following: Several components needed for assisting an investigator in analyzing speech data are proposed and described. SCALA-Speech, a tool incorporating these components and realizing them with exemplary algorithms that can be selected depending on the application, is presented. As an example, using SCALA-Speech is demonstrated on audio extracted from videos contained in Twitter posts related to Belarus.

2 SCALA-Speech

2.1 Single Channel Analysis and Labeling Application (SCALA)

SCALA – a Single Channel Analysis and Labeling Application – is a tool developed with the purpose of analyzing and annotating different types of audio data. The main aspects include the application of integrated speech analysis algorithms and the visualization of their results. In addition, different kinds of annotated data may be visualized temporally aligned with a time-frequency representation (i.e., a spectrogram) of the audio signal itself. Visualizations include, in addition to displaying different types of annotations from label files, also feature matrices or special data types such as F0 trajectories. SCALA-Speech provides tools for signal analysis and enhancement, e.g. by applying filters or adjusting the carrier frequency as in the case of processing amplitude modulated radio frequency (RF) communication signals. A typical example of the graphical user interface (GUI) is shown in Fig. 1. The layout of the GUI can be chosen flexibly from several pre-defined layouts, depending on the given application. Individual GUI layouts may be created as well, allowing a user to reduce the view to the essential components required for the given scenario.
SCALA-Speech offers a large set of built-in speech processing algorithms, as specified in Sec. 2.2 and partially explained in more detail in the following sections. A short overview

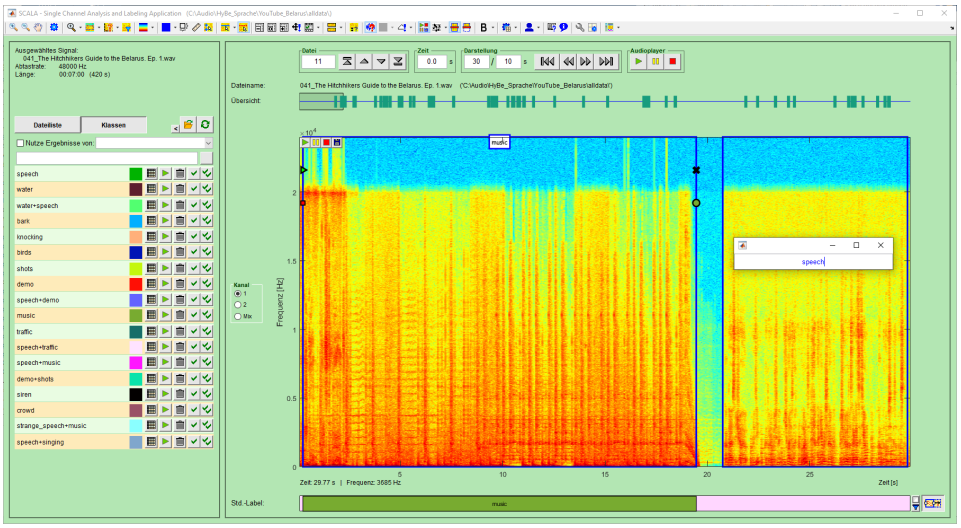

Fig. 1: SCALA-Speech: Layout in label mode generating a standard label file. The spectrogram contains two annotations, one in the process of being built, the left-hand side shows a class list.

about the covered speech analysis techniques is shown in Fig. 2 (left-hand side in blue). Next to the built-in methods, an excerpt of the basic tools is shown (right-hand side in green). Apart from analyzing signals by applying algorithms and visualizing results, one of the main tools provided is the *label mode*. Segments of signals may be marked by highlighting the corresponding regions of interest by simply specifying a rectangular region within the spectrogram. An example can be found in Fig. 1, depicting two rectangular regions in the spectrogram of the selected signal. The left-hand side shows a customized list of available classes. Annotations are stored in a user-defined format, e.g. in a standard label format including information about start time, end time, label, and – in case of modulated speech signals – carrier frequency, bandwidth, and sideband. In addition to machine-readable label files, the highlighted segments may also be saved as single wave files in class specific sub-folders thus allowing to train, for instance, keyword spotting or speaker recognition systems. For the latter case, the GUI offers a module for training speaker recognition models for several recognition methods. An additional feature allows for importing results of any algorithm applied to a signal as a starting point for manual annotations. Speech detection and segmentation, for instance, may provide relevant speech regions which can be labeled manually to generate speaker ground truth data. SCALA-Speech also includes a program for parameter optimization allowing to find optimal thresholds and parameter setups with respect to detection and false alarm rates. Finally, additional performance tracks may be shown.

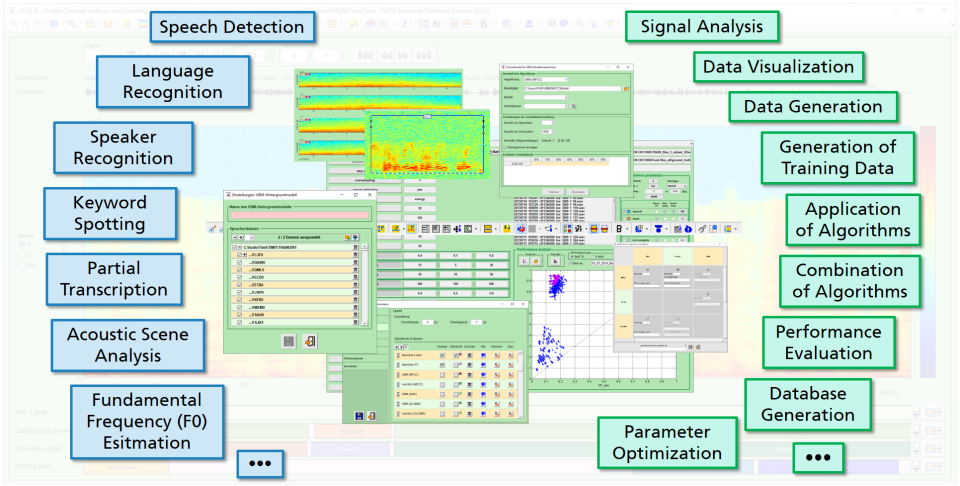

Fig. 2: SCALA-Speech: Overview about built-in components and tools. The system is modular and individual algorithms realizing these components can be selected depending on the application.

2.2 Speech Analysis Components

SCALA-Speech provides a user with a variety of audio processing modules, each of which realized by one or more algorithms particularly suited for specific speech analysis tasks. For e.g. speaker recognition, several different algorithms are available. An overview is given in Fig. 2 (left-hand side), including speech detection, audio segmentation, language recognition, speaker recognition, keyword spotting, as well as any combination of these. Note that these exemplary algorithms can be replaced with other algorithms depending on the particular application. Several included algorithms are described next.

2.2.1 Speech Detection

Speech detection (SD) or voice activity detection (VAD) is a mandatory pre-processing step for many speech analysis algorithms. In particular for heterogeneous data, it is important to extract those regions of audio signals containing speech, as opposed to other acoustic classes such as music, nature, or traffic noise. Depending on the given scenario, different SD algorithms may be used, each having different advantages depending on the given data. The so-called *Shift-F0* speech detection method extracts the fundamental frequency (F0) of speech by using an enhanced autocorrelation function (ACF) in the spectral domain [KCUU14, KCU14]. In contrast to the standard autocorrelation function, the shift-ACF applies a variable combination of product and minimum operations to extract multiply repeated components, while suppressing non-repeating structures [Ku13]. It is applied to the time dimension of a spectrogram, where the F0 appears as a horizontal energy-rich

component with additional energy at the respective harmonic frequencies. Those harmonics, being multiples of F0 (repeated structures in the spectral domain) can hence be detected by means of ACF-based methods. This approach is suited for RF communication scenarios that typically contain different and potentially strong noise types [KCU14, KCUU14].
VAD-MFCC is a voice activity detection (VAD) method based on Mel frequency cepstral coefficients (MFCC) in combination with spectral energy envelope extraction [vZ12]. This approach is a well-suited pre-processing step for further speech analysis algorithms such as speaker diarization or language recognition.It does not need any training material and is independent of speakers or language as opposed to statistical or learning-based approaches.
BBSD is a broadband speech detection algorithm developed for RF communication [UKK16]. It is composed of a combination of different methods that extract statistics of audio features such as spectral flux or spectral flatness. BBSD detects speech regions regardless of the given carrier frequency of the transmitted signals, distinguishing them from other RF signals.

2.2.2 Language Recognition

After detecting regions that contain speech, the next step is to identify *which* language is being spoken. For this purpose, we utilize a Speechbrain [Ra21] model based on the ESCAPA-TDNN achitecture [DTD20] to extract embeddings suitable for discriminating among languages. More concretely, the model is pre-trained on the Voxlingua107 dataset [VA21], which contains 6628 hours of speech from Youtube videos in 107 different languages. For specific use cases e.g. with fewer languages, another classifier can be trained on top by using the pre-trained embeddings belonging to only these languages as an input. Moreover, embeddings can be extracted for music datasets such as Orchset [BMG16] and environmental sounds such as the DCASE 2019 acoustic scene classification dataset [MHV19] and utilized as additional classes to obtain more robust representations in case the speech detection algorithms yield some false positive results.

2.2.3 Speaker Diarization and Recognition

Another important component of SCALA-Speech is the possibility to search for specific speakers (speaker recognition) or detect changes of speakers (speaker diarization [Pa22]). As in language recognition, the state-of-the-art in speaker recognition and diarization is to utilize discriminative embeddings. We use a pre-trained x-vector model [Sn18] from Speechbrain [Ra21] trained on VoxCeleb2 [CNZ18], which consists of over 1 million utterances from Youtube videos of approximately 6000 celebrities. It is also possible to train a model for extracting embeddings from scratch but this requires a reasonable amount of labeled data. For speaker diarization, the procedure presented in [Ar20] based on spectral clustering [NJW01] is used. More classical methods such as Universal Background Model (UBM) [RQD00] and i-vector [De11] are also available. For all methods, multiple complementary speech features can be utilized and their results can be combined [Wi18].

2.2.4 Keyword Spotting

HFCC-KWS is a keyword spotting (KWS) system based on MFCC-like audio features, the so-called *HFCC-ENS* features (human factor cepstral coefficients using energy normalized statistics) [KvZ10, vZKM10, ZU13]. It is a query-by-example approach using dynamic time warping (DTW) for matching. This method is independent of the spoken language or of the speakers, though the quality of the detection results depends much on the similarity of query and data. The best results are obtained when using queries of similar acoustic characteristics as the target signal. As these are not always available, this is a trade-off between performance and the amount of required training data.
A variation of the above method uses a pre-trained wav2vec 2.0 model [Ba20] for cross-lingual representation learning [Co21] from Speechbrain [Ra21] to extract representations as feature vectors. The actual KWS method is the same as the one used with HFCC-KWS.
DTW-KENS-KWS combines traditional methods with deep learning approaches by using discriminative *embeddings* extracted with a neural network as features and DTW for extracting time positions of matches [WCUG21]. Instead of extracting paths, DTW is only used to calculate score curves. The resulting score matrix is then post-processed and enhanced. Even though deep learning approaches typically require large amounts of training data, this method requires only a few examples of each keyword and thus can be considered a low-resource approach. The performance of the proposed system has been efficiently demonstrated on a dataset containing spoken coordinates. While HFCC-KWS is better suited for longer phrases (cf. [KvZ10]), this method can also deal with short queries.

3 Use case

One possible application of SCALA-Speech is speech monitoring in a big data context, i.e., in application scenarios where large amounts of data have to be analyzed and thus automatic procedures for supporting a user are desirable. Some of the algorithms described above have been combined and integrated in a system to detect *hybrid threats* [GSS21]. Hybrid threats or attacks in general denote the coordinated application of a mixture of measures from different domains in order to exploit vulnerabilities of an opponent. In [GSS21], hybrid threats are considered involving measures in the social media and cyber domains as well as potential attacks in the electromagnetic spectrum (EMS). The authors describe a system consisting of a variety of different sensors including both so-called local sensors such as an RF-scanner (radio frequency) or a passive radar, as well as virtual sensors for, e.g. cyber threat analysis or social media analysis, all of which gathering the retrieved information in a shared visualization tool. The speech analysis algorithms proposed in this paper are included into that approach by systematically analyzing the audio content retrieved from videos or blogs linked in twitter messages, while the textual parts of the messages are scanned by a social media analysis sensor. It is worth to mention that the latter sensor can also be used for detecting *fake news* [PS21]. Integrated in such a system, a speech analysis tool can provide additional information to the purely text-based social media analysis.

Regarding further applications of the toolset made available in SCALA-Speech, for instance keyword spotting or speaker identification can be used in contexts where little training or reference data is available. This can be the case when analyzing particular languages or dialects, where no or only very few training data is given, e.g. in extremely non-standard acoustic conditions with particular or strong background noise.
The proposed speech analysis tool allows a user to gather information of different types and analyze audio data in different use cases. In use cases requiring training data, e.g., when using AI-based algorithms, SCALA-Speech provides a possibility to generate such material. One example is one of the built-in speaker identification methods. The algorithm is trained on a large number of different speakers and needs to be adapted when looking for particular speakers. In such a case, a small amount of reference audio data is needed, which can be created with the corresponding labeling tool. The same holds for KWS scenarios.

3.1 Dataset

Several of the built-in speech analysis algorithms of SCALA-Speech have been applied to a particular dataset developed and annotated in the context of the hybrid threat analysis tool proposed in [GSS21]. It consists of a variety of different audio files extracted from video files referenced in twitter messages. In addition to personal blogs, these include mainly YouTube videos, all related to the topic *Belarus*. Since the considered files are linked to the given topic only by means of predefined keywords, in this case *Belarus*, the data to be analyzed may comprise any topic, also, for instance, videos related to the Eurovision song contest. Other files may simply contain music, e.g., if someone has linked a music video in a twitter post. Heterogeneity of the considered data implies particularly parameterized analysis methods. In a first step, the relevant data needs to be extracted by means of SD and VAD methods to separate speech content from the rest. A language recognition algorithm provides background information on the keywords and speakers to search for. The dataset contains 139 audio files with a total length of 39 hours, 92 of which contain speech. The files vary from 1 minute to 6.7 hours of duration. This dataset is only used for demonstration purposes and a systematic evaluation cannot be presented since the work is still in progress.

3.2 Demonstration

Fig. 3 shows a typical use case-specific GUI layout. In addition to the file list and the spectrogram of the selected signal, this layout includes data tracks for speech detection, language recognition, speaker diarization, and keyword spotting, which are all accompanied by a separate label track providing manual annotations. SCALA-Speech allows a user to navigate through a given set of files.To get an overview about the relevant information, the list of files can be filtered according to available results of a selected algorithm, e.g. speech detection. The GUI can be configured in order to allow a performance comparison of different speech processing methods applied to the same task. Fig. 3 depicts an example

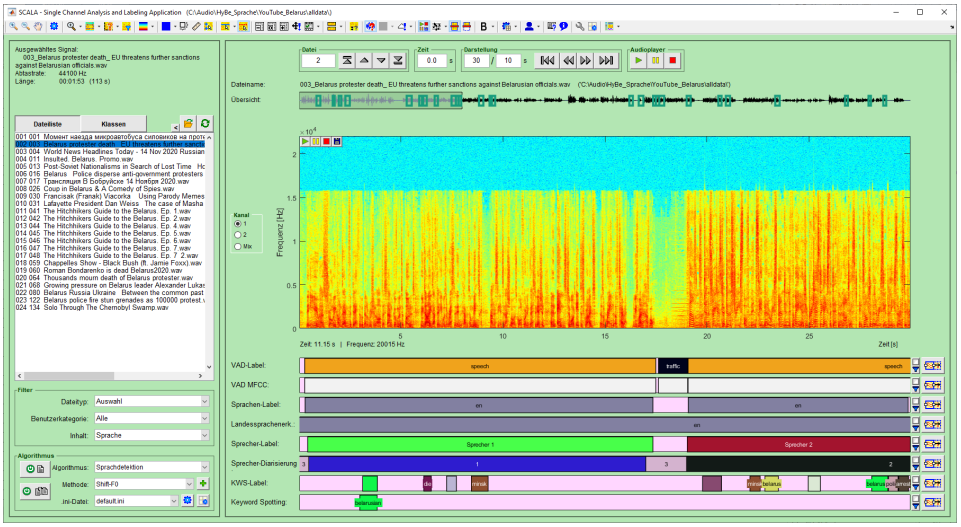

Fig. 3: SCALA-Speech algorithms: synchronous visualization of annotations and algorithm results.

taken from the described dataset. As can be seen, with the given data, most of the available speech analysis methods need to be adjusted, e.g. by additional training like in the case of keyword spotting or by parameter optimization as in the case of speech detection, and the language recognition approach needs to be combined with a VAD method. SCALA-Speech provides an overview on the quality of the data to be analyzed and represents an annotation tool for generating the necessary training or reference data.

4 Conclusions and Future Work

In this work, SCALA-Speech, a tool for assisting investigators in annotating and analyzing audio signals, has been presented. SCALA-Speech offers various modules for speech signal analysis such as voice activity detection, language recognition, speaker diarization/recognition and keyword spotting. The usage of SCALA-Speech has been demonstrated on audio tracks of videos obtained from Twitter posts related to Belarus.
For future work, systematic experimental evaluations with a focus on user experience for specific use cases are planned. To this end, we collaborate with partners representing different application domains. Some of the components have already been discussed with users in the domain of security-related applications. Based on the obtained results, individual signal analysis modules of SCALA-Speech will be adapted in order to meet the requirements of the particular use cases. The system will furthermore serve as a framework for integrating and testing audio and speech signal analysis algorithms developed in the future. Concerning mass data analysis, the audio analysis algorithms evaluated in SCALA-Speech can subsequently

be integrated in automatic signal production systems such as those we are developing in parallel in the context of RF signal analysis.

Bibliography

[AF12] Alexander, Anil; Forth, Oscar: Blind Speaker Clustering Using Phonetic and Spectral Features in Simulated and Realistic Police Interviews. In: IAFPA. 2012.

[Ap] Appen, `https://appen.com/data-types/#data-audio`, visited on 2022-05-12.

[Ar20] Aronowitz, Hagai et al.: New Advances in Speaker Diarization. In: INTERSPEECH. ISCA, pp. 279–283, 2020.

[Ba20] Baevski, Alexei et al.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: NeurIPS. 2020.

[BMG16] Bosch, Juan J.; Marxer, Ricard; Gómez, Emilia: Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. Journal of New Music Research, 45(2):101–117, 2016.

[CNZ18] Chung, Joon Son; Nagrani, Arsha; Zisserman, Andrew: VoxCeleb2: Deep Speaker Recognition. In: INTERSPEECH. ISCA, pp. 1086–1090, 2018.

[Co21] Conneau, Alexis et al.: Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In: INTERSPEECH. ISCA, pp. 2426–2430, 2021.

[Cr16] Crocco, Marco et al.: Audio surveillance: A systematic review. CSUR, 48(4):1–46, 2016.

[De11] Dehak, Najim et al.: Front-End Factor Analysis for Speaker Verification. IEEE Trans. Speech Audio Process., 19(4):788–798, 2011.

[Di21] Dikici, Erinç et al.: ROXSD: a Simulated Dataset of Communication in Organized Crime. In: SPSC. ISCA, pp. 32–36, 2021.

[DTD20] Desplanques, Brecht; Thienpondt, Jenthe; Demuynck, Kris: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In: INTERSPEECH. ISCA, pp. 3830–3834, 2020.

[GSS21] Gerz, Michael; Schwarze, Arne; Stuch, Hans Peter: Connecting the Dots–Enhancing the Information Processing Chain for the Detection of Hybrid Threats for Host Nation Support and Territorial Operations. In: STO, NATO-OTAN. 2021.

[KCU14] Kurth, Frank; Cornaggia-Urrigshardt, Alessia: Detection of audio events with repetitive structure using generalized autocorrelations. In: ITG Speech. VDE, 2014.

[KCUU14] Kurth, Frank; Cornaggia-Urrigshardt, Alessia; Urrigshardt, Sebastian: Robust F0 estimation in noisy speech signals using shift autocorrelation. In: ICASSP. IEEE, pp. 1468–1472, 2014.

[Kh18] Khelif, Khaled et al.: SIIP: An innovative speaker identification approach for law enforcement agencies. In: STO, NATO-OTAN. pp. 1–14, 2018.

[Ku13] Kurth, Frank: The shift-ACF: Detecting multiply repeated signal components. In: WASPAA. IEEE, pp. 329–332, 2013.

[KvZ10] Kurth, Frank; von Zeddelmann, Dirk: An Analysis of MFCC-like Parametric Audio Features for Keyphrase Spotting Applications. In: ITG Speech. VDE, 2010.

[MHV19] Mesaros, Annamaria; Heittola, Toni; Virtanen, Tuomas: Acoustic Scene Classification in DCASE 2019 Challenge: Closed and Open Set Classification and Data Mismatch Setups. In: DCASE. pp. 164–168, 2019.

[NJW01] Ng, Andrew Y.; Jordan, Michael I.; Weiss, Yair: On Spectral Clustering: Analysis and an algorithm. In: NIPS]. MIT Press, pp. 849–856, 2001.

[Pa22] Park, Tae Jin et al.: A review of speaker diarization: Recent advances with deep learning. Comput. Speech Lang., 72:101317, 2022.

[PS21] Pritzkau, Albert; Schade, Ulrich: Vorsicht: mögliche „Fake News “–ein technischer Ansatz zur frühen Erkennung. In: Wahrheit und Fake im postfaktisch-digitalen Zeitalter, pp. 115–131. Springer, 2021.

[Ra21] Ravanelli, Mirco et al.: SpeechBrain: A General-Purpose Speech Toolkit. CoRR, abs/2106.04624, 2021.

[RO] ROXANNE Real time netwOrk, teXt and speaker ANalytics for combating orgaNized crimE, https://roxanne-euproject.org/, visited on 2022-05-16.

[RQD00] Reynolds, Douglas A.; Quatieri, Thomas F.; Dunn, Robert B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Process., 10(1-3):19–41, 2000.

[Sh] Shaip, https://www.shaip.com/offerings/audio-annotation, visited on 2022-05-12.

[Sn18] Snyder, David et al.: Spoken Language Recognition using X-vectors. In: Odyssey. ISCA, pp. 105–111, 2018.

[UKK16] Urrigshardt, Sebastian; Kreuzer, Sebastian; Kurth, Frank: General detection of speech signals in the time-frequency plane. In: ITG Speech. VDE, 2016.

[VA21] Valk, Jörgen; Alumäe, Tanel: VOXLINGUA107: A Dataset for Spoken Language Recognition. In: SLT. IEEE, pp. 652–658, 2021.

[vZ12] von Zeddelmann, Dirk: A feature-based approach to noise robust speech detection. In: ITG Speech. VDE, 2012.

[vZKM10] von Zeddelmann, Dirk; Kurth, Frank; Müller, Meinard: Vergleich von Matching-Techniken für die Detektion gesprochener Phrasen. DAGA, pp. 257–260, 2010.

[WCUG21] Wilkinghoff, Kevin; Cornaggia-Urrigshardt, Alessia; Gökgöz, Fahrettin: Two-Dimensional Embeddings for Low-Resource Keyword Spotting Based on Dynamic Time Warping. In: ITG Speech. VDE, pp. 9–13, 2021.

[Wi18] Wilkinghoff, Kevin; Baggenstoss, Paul M.; Cornaggia-Urrigshardt, Alessia; Kurth, Frank: Robust Speaker Identification by Fusing Classification Scores with a Neural Network. In: ITG Speech. VDE, pp. 261–265, 2018.

[ZU13] Zeddelmann, Dirk von; Urrigshardt, Sebastian: Ein Demonstrator zum Keyword-Spotting basierend auf gehörangepassten Audiomerkmalen. In: GI-Jahrestagung. Gesellschaft für Informatik e.V., pp. 3026–3028, 2013.