A Paralinguistic Approach To Speaker Diarisation

Using Age, Gender, Voice Likability and Personality Traits

Yue Zhang
Department of Computing
Imperial College London
London SW7 2AZ, U.K.
yue.zhang1@imperial.ac.uk

Felix Weninger
Nuance Communications
89077 Ulm, Germany
felix@weninger.de

Boqing Liu
Department of Computing
Imperial College London
London SW7 2AZ, U.K.
boqing.liu16@imperial.ac.uk

Maximilian Schmitt
Chair of Complex & Intelligent Systems, University of Passau
94032 Passau, Germany
maximilian.schmitt@uni-passau.de

Florian Eyben
audEERING GmbH
82205 Gilching, Germany
fe@audeering.com

Björn Schuller
Department of Computing
Imperial College London
London SW7 2AZ, U.K.
bjoern.schuller@imperial.ac.uk

ABSTRACT

In this work, we present a new view on automatic speaker diarisation, i. e., assessing "who speaks when", based on the recognition of speaker traits such as age, gender, voice likability, and personality. Traditionally, speaker diarisation is accomplished using low-level audio descriptors (e. g., cepstral or spectral features), neglecting the fact that speakers can be well discriminated by humans according to various perceived characteristics. Thus, we advocate a novel paralinguistic approach that combines speaker diarisation with speaker characterisation by automatically identifying the speakers according to their individual traits. In a three-tier processing flow, speaker segmentation by voice activity detection (VAD) is initially performed to detect speaker turns. Next, speaker attributes are predicted using pre-trained paralinguistic models. To tag the speakers, clustering algorithms are applied to the predicted traits. We evaluate our methods against state-of-the-art open source and commercial systems on a corpus of realistic, spontaneous dyadic conversations recorded in the wild from three different cultures (Chinese, English, German). Our results provide clear evidence that using paralinguistic features for speaker diarisation is a promising avenue of research.

KEYWORDS

Speaker diarisation; speaker identification; speaker characteristics; computational paralinguistics

1 INTRODUCTION

> The most intelligible factor in language is not the word itself, but the music behind the words, the passions behind the music, the person behind these passions: everything, in other words, that cannot be written.
>
> — Friedrich Nietzsche

Speaker diarisation is the task of determining "who speaks when" in an audio stream [2]. The research field has initially emerged within the realm of speech processing technology, where speaker diarisation serves as an upstream processing step for automatic speech recognition (ASR). Along with the grown maturity of this field over the last decade, the research topic has gained increasing attention due to its broad application spectrum including multimedia information retrieval (e. g., video tagging), human-machine interaction, and rich transcription (RT), as well as speaker recognition in phone call conversations, broadcast news, and meeting recordings [32, 41]. In standard systems, speaker diarisation is performed using clustering techniques based on hidden Markov models (HMMs), where each state, corresponding to a speaker, is represented by a Gaussian mixture model (GMM). Recent efforts to improve the diarisation performance head into the directions of using time-decay [3, 7], prosodic [11, 21], and multi-modal, audio-visual features [12, 27]. Moreover, many studies [4, 15, 18, 28, 43] have tackled the problem of overlapping speech, which needs to be assigned to multiple speakers, or else would considerably deteriorate diarisation performance.

Despite the manifold work done in this field, it has been disregarded in current research that humans are naturally able to discriminate their dialogue partners according to a variety of speaker attributes as carried over the voice. Intuitively, the voice timbre, conveying rich demographic information, naturally forms the voice identity of a speaker. Hence, we posit that speaker analysis is of crucial importance to identify the active speaker. Within the field of Computational Paralinguistics [35], research has been carried out on automatically recognising a plethora of speaker characteristics, including transient speaker states such as affect, health condition, and long-term speaker traits like personality and biological primitives, as well as speaking styles. Exemplary recognition tasks

MM '17, October 23–27, 2017, Mountain View, CA, USA.

ISBN 978-1-4503-4906-2/17/10…$15.00

DOI: https://doi.org/10.1145/3123266.3123338

that have been featured in the INTERSPEECH Computational Paralinguistic ChallengE (ComParE) series [34] include emotion (2009), interest, age, and gender (2010), sleepiness and alcohol intoxication (2011), the OCEAN five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism), voice pathology, and likeability (2012), social signals, conflict, and autism (2013), physical and cognitive load (2014), the degree of nativeness, Parkinson's condition, and Eating condition (2015), deception, sincerity, and native language (2016), cold and addressee (2017).

All these paralinguistic characteristics – in particular those reveling speaker traits – can be considered highly relevant to identify the speakers according to their individual sound profiles. Attempts to make use of speaker information beyond basic acoustic features for speaker diarisation include the usage of *a priori* information on the speaker identity [25], gender specific background models [16], and speaker role n-gram models [42]. However, to the authors' best knowledge, speaker characterisation using a rich variety of paralinguistic trait information has never been applied to speaker diarisation before. In this work, we introduce a novel paralinguistic approach based on a three-tier system: speaker segmentation, characterisation, and clustering.

In the remainder of this work, we describe the speaker diarisation systems in Section 2. The performance evaluation on a corpus of spontaneous human-to-human conversationsdetailed in Section 3. Concluding remarks and impulses for future research are given in Section 4.

2 SPEAKER DIARISATION SYSTEMS

In general, prototypical speaker diarisation systems are constituted from several processing stages as depicted in Figure 2. At the front-end, voice activity detection (VAD) serves to split the audio stream into speech and non-speech segments, where the speech part can be further refined to include speaker homogeneous and overlapping speech [4]. Subsequently, cepstral features are extracted to perform clustering. To this end, hierarchical bottom-up (off-line only) or top-down (also suitable for on-line diarisation) approaches are applied to merge or divide the formed clusters based on the defined distance metrics and stopping criterion [22]. Typically, clustering is done by estimation of hidden Markov models (HMMs), of which each state corresponds to a speaker, whose feature distribution is, in turn, modelled by a Gaussian mixture model (GMM) [2].

Alternatively, segmentation and clustering are often performed in one step by employing Viterbi decoding between iterations [1, 14] and Bayesian adaption of a universal background model (UBM) [31]. Finally, on the output side, the (re-)segmented and clustered segments are assigned to relative speaker labels (e. g., speaker 1) or true identities (e. g., speaker name) [31, 40]. It is noted that speech overlap detection is a challenging task in itself as will be addressed in Section 2.2.

2.1 LIUM SpkDiarization

For comparison with standard open-source systems, we evaluated the LIUM framework (version 8.4.1) [23], which achieved the best performance for the task of speaker diarisation of broadcast news in the French Ester 2 evaluation campaign 2009 [13]. Although it is possible to adapt the system to process telephone data, the toolbox is primarily developed for radio or TV shows, thus one cannot expect the same level of performance on voice over IP phone conversation.

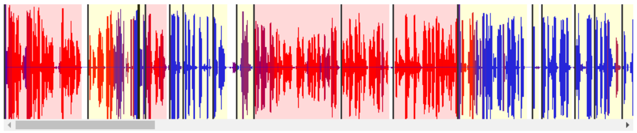

Figure 1: Audio diarisation by the sensAI system applied to an exemplary dyadic conversation. Waveform view (voice segments are highlighted – red waveform = female, blue waveform = male)

2.2 sensAI Voice Analytics Tools

The sensAI audio analysis engine is a commercial product provided by audEERING GmbH [1]. Its speaker diarisation unit makes use of a robust VAD technology based on Long Short-Term Memory Recurrent Neural Network (LSTM) [20] that locates voice signals even in highly noisy environments, such as background music or street noise (cf. Figure 1). Its algorithm is based on GMMs combined with UBMs and maximum a-posteriori (MAP) adaptation [16]. Starting with one background model each for male, female, and garbage, models for individual speakers are successively created based on the likelihood ratios. A major drawback of standard speaker diarisation systems is that speech segments are assigned to only one speaker, thus incurring missed speech errors and false alarms in regions where multiple speakers are active [4]. The problem of speech overlap detection was recently tackled in the work [15] with LSTM-RNN. Building upon this work, it has been shown in [18] that the performance of overlap detection can be largely improved by bidirectional LSTM (BLSTM) RNNs with four outputs (voice activity, speech overlap, male and female speaker recognition), trained with the CURRENNT library [44]. The sensAI software makes use of this algorithm for its VAD by segmenting the audio stream accordingly.

2.3 Paralinguistic Approach

The paralinguistic recogniser is integrated in the training module running parallel to the cepstral feature extractor in the standard processing chain (cf. Figure 2). The aim is to train discriminative classifiers to automatically determine speaker characteristics, thereby generating a paralinguistic trait vector as input for speaker clustering. Here, the paralinguistic informations bits are considered auxiliary and complementary to the acoustic features, and are not meant to replace these. Thus, we propose a hybrid approach by jointly using acoustic and paralinguistic features. In this section, the design and implementation of the paralinguistic diarisation framework are described in detail.

2.3.1 Voice Activity Detection and Segmentation. At first in a fundamental step, the VAD senses audio segments in which only one speaker is active. By that, accurate timestamps delineating the speaker change points can be provided for segmentation. For the purpose of this study, we adopt sensAI's output segments.

[1] http://audeering.com/technology/sensai/

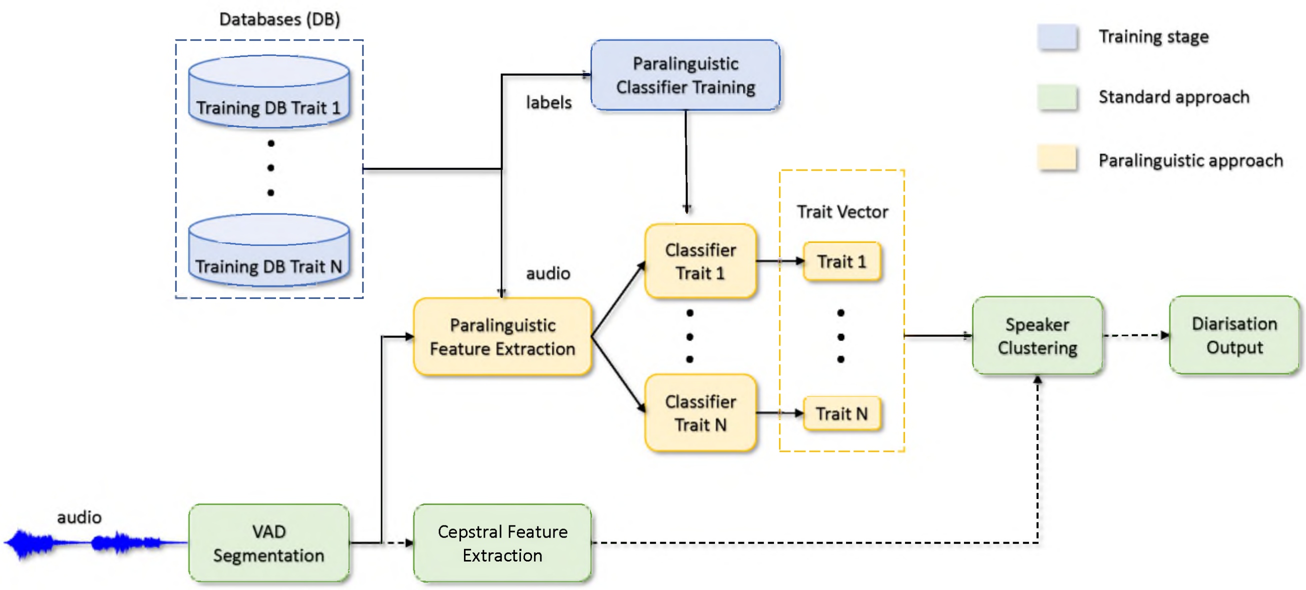

Figure 2: Process flow chart in paralinguistic (solid line) and standard (dashed line) speaker diarisation system.

2.3.2 *Paralinguistic Speech Analysis.* For the paralinguistic recogniser, we consider the biological primitives (age, gender), the OCEAN personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism), and the voice likability, as featured in the INTERSPEECH 2010 Paralinguistics Challenge [36, 37] and the INTERSPEECH 2012 Speaker Trait Challenge [38, 39]. These traits were chosen due to their quasi time-invariance during the recording sessions, unlike the rapidly changing states, for instance, emotion. Altogether, they constitute an 8-dimensional paralinguistic trait vector, comprising seven binary class labels (OCEAN, likability and gender) and one numeric attribute (age). The rationale to treat age as a regression task is that the interlocutors are likely in the same age group (young, adult, old) in our scenario, however, for diarisation, it is important that their speech parts are associated with different prediction values.

Feature Extraction. The paralinguistic recogniser operates on suprasegmental acoustic features. These are obtained by using the openSMILE (Speech and Music Interpretation by Large audio-Space Extraction) toolkit [9, 10]. The *ComParE* set is a well-evolved feature set for automatic recognition of paralinguistic speech phenomena, serving as a standard reference in the speech community. It contains 6 373 static attributes resulting from the computation of various functionals over low-level descriptor (LLD) contours. The configuration files for openSMILE are provided with the openSMILE 2.1 public release. Important subgroups of the ComParE feature set comprise prosodic (*PROS*), Mel Frequency Cepstral Coefficients (*MFCC*), spectral (*SPEC*), and voice quality (*VQ*) features. Due to their relevance for speaker identification [26], the MFCC features are concatenated with the paralinguistic traits into one input vector for the clustering algorithm in the early fusion approach. A full description of the the acoustic features can be found in [8, 45].

Databases. The training of the trait predictors is carried out on the INTERSPEECH Challenge datasets. The German aGender corpus [5] contains 47 hours of telephone speech of 954 speakers. For the purpose of this study, the instances belonging to the 'Child' group (7-14 years old) were removed from the training and validation set. The age distribution of the remaining speakers ranges from 15 to 80 years, with a mean age of 43.6 and a standard deviation of 19.7. The labels of the test set are not provided. The Speaker Likability Database (SLD) [6] is a subset of the aGender database, including 800 speakers and one utterance each. Likability ratings on a seven point Likert scale were established by presenting the recordings to 32 participants (17 male, 15 female, aged 20-42, mean=28.6, standard deviation=5.4). To establish a consensus from the individual likability ratings, the evaluator weighted estimator (EWE) [17] was used. For the Challenge, the EWE value was discretised into the 'likable' (L) and 'non-likable' (NL) classes based on the median EWE rating of all stimuli in the SLD. The French Speaker Personality Corpus (SPC) [24] comprises 640 clips of 322 speakers in 1.7 hours of speech. The personality traits were assessed by 11 annotators according to the Big-Five personality dimensions [30]. The ratings were centered to zero mean on a per rater basis in order to eliminate individual biases. Each clip was labelled to be 'above average' (X) for a given trait X ∈ { O, C, E, A, N } if the majority of the judges assign a score higher than the arithmetic mean of their ratings for the specific trait; or else it is labelled as NX. Both the SLD and the SPC are age and gender balanced. Table 1 shows the partitioning of the databases into speaker-disjunct and stratified training, development, and test set.

Model Training. Model training is carried out on the training and development set, where each feature is standardised to zero mean and unit variance. Standardisation is done separately on the

Table 1: Partitioning into training, development, and test set for paralinguistic speech analysis. Binary classification: Speaker Likability Database by L: likable / NL: non-likable; Speaker Personality Corpus by X: high on trait X / NX: low on trait X, X ∈ { O, C, E, A, N }; aGender corpus (gender) by f: female and m: male. Regression: aGender corpus (age).

Trait	Class	Train	Devel	Test	Σ
Likability	L	189	92	119	400
	NL	205	86	109	400
Openness	O	97	70	80	247
	NO	159	113	121	393
Conscientious.	C	110	81	99	290
	NC	146	102	102	350
Extraversion	E	121	92	107	320
	NE	135	91	94	320
Agreeableness	A	139	79	105	323
	NA	117	104	96	317
Neuroticism	N	140	88	90	318
	NN	116	95	111	322
Gender	f	14 135	9 644	–	23 779
	m	13 985	8 508	–	22 493
Age	numeric	28 120	18 152	–	46 272

database to which diarisation is applied, in order to to alleviate cross-corpus effects. To foster reproducible research, we employ the open-source data mining toolkit WEKA (version 3.8.1) [19]. In particular, we use Support Vector Machines (SVM) with linear kernels for the classification tasks, and for age estimation Support Vector Regression (SVR; also with linear kernels) with epsilon-insensitive loss, which are generally robust to over-fitting in high dimensional feature spaces. To train the classifiers, the Sequential Minimal Optimisation (SMO; [29]) is applied, for which the complexity parameter $C \in \{10^{-5}, 10^{-4}, \cdots, 10^{-2}\}$ is optimised on the development set. For consistency, we use the same evaluation measures as in the Challenges, i. e., Unweighted Average Recall (UAR) for classification and Spearman's Correlation Coefficient (ρ) for regression. The obtained performances are 60.6 % (likability; $C = 10^{-2}$), 64.0 % (openness, $C = 10^{-5}$), 73.6 % (conscientiousness, $C = 10^{-2}$), 83.6 % (extraversion, $C = 10^{-4}$), 65.8 % (agreeableness, $C = 10^{-5}$), 70.2 % (neuroticism, $C = 10^{-4}$), 95.3 % (gender, $C = 10^{-4}$), and .482 (age, $C = 10^{-3}$). Deviations from the Challenge baseline can be explained by the usage of the ComParE feature set and the latest openSMILE version.

2.3.3 Speaker Clustering. Whereas the segmentation step aims at separating adjacent windows which belong to different speakers, clustering operates globally on the audio stream, trying to identify and group together segments for each speaker [2]. Ideally, there should be one cluster for each speaker. As in most standard diarisation systems, the unsupervised agglomerative (bottom-up) clustering mechanism is applied in the proposed framework to identify and grouping together the segments of the same speaker. To this end, a suitable distance metric and a stopping criterion need to be defined to form the optimal number of clusters. Given the fact that a majority of the paralinguistic traits contains binary labels, we decide on the Manhattan distance and the Elbow method. As implementation, we choose the hierarchical clustering package from the open-source software SciPy (version 0.13.3).

Table 2: Statistics (gender and age distribution and duration) of the SEWA sessions used in the evaluation.

Culture	# Subjects			Duration
	Female	Male	Age (mean ± stddev)	[min]
British	33	33	33.4 ± 14.4	100.7
Chinese	34	36	30.4 ± 9.8	89.1
German	25	39	31.6 ± 12.6	91.4

3 PERFORMANCE EVALUATION

In this section, we investigate the performance level of the proposed approach in comparison with the state-of-the-art diarisation systems as described in Section 2. For this purpose, the evaluation procedure is described, followed by a discussion of results.

3.1 Evaluation Measure

The diarisation output is processed in the RTTM (Rich Transcription Time Marked) format. The standard evaluation measure is the Diarisation Error Rate (DER) as used by the National Institute of Standards and Technology (NIST) in the RT evaluations[2]. The DER corresponds to the ratio of incorrectly detected speaker time to total speaker time, where the system output speaker segment sets are mapped to reference speaker segment sets so as to minimise the total error. It is defined as the sum of the miss (speaker in reference but not in hypothesis), false alarm (speaker in hypothesis but not in reference), and speaker error (correctly detected speech but not assigned to the right speaker) rates, divided by the total speech time in reference (time-weighted).

3.2 Database

In order to compare the diarisation systems in a real-life scenario, the Sentiment Analysis in the Wild (SEWA) database[3] is used. It contains video calls between subjects from six different cultures with a broad distribution in age (from 18 to over 60). For the purpose of this study, we use a subset of the database comprising all video chat recordings of the three cultures *British*, *Chinese*, and *German*. In the data collection, the subjects were first asked to watch a 90 seconds long commercial presented in their native language. Then, in each recording session, two subjects who were acquainted with each other exchanged their views on the commercial in a three minutes long video chat via a online recording platform. The subjects' demographics and the total duration of the recordings are shown in Table 2. Manual transcription of all conversations were completed by a native speaker of the respective language. In detail, the transcript marks every utterance's start and end time (in seconds), the speaker's subject ID, and the verbal content. Some non-verbal utterances, such as laughter, coughing and breathing sound, were annotated as well.

[2] https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

[3] http://db.sewaproject.eu/

Table 3: Comparison of the speaker diarisation systems (LIUM, sensAI, paralinguistic approach) in terms of miss, false alarm, speaker error (spkr err), and diarisation error rate (DER) [%].

Culture	Miss	False Alarm	Spkr Err	DER
LIUM				
CHN	6.48	10.00	35.05	51.52
ENG	4.56	14.89	35.87	55.32
GER	4.08	3.51	39.81	47.40
sensAI				
CHN	7.20	1.89	6.48	15.58
ENG	6.71	5.03	11.01	22.76
GER	11.43	0.32	7.14	18.89
Paralinguistic				
CHN	7.20	1.89	38.16	47.26
ENG	6.71	5.03	37.05	48.79
GER	11.44	0.32	35.57	47.32
MFCC				
CHN	7.20	1.89	37.30	46.40
ENG	6.71	5.03	35.90	47.65
GER	11.44	0.32	35.37	47.13
Paralinguistic + MFCC				
CHN	7.20	1.89	37.27	46.36
ENG	6.71	5.03	35.89	47.63
GER	11.44	0.32	35.35	47.11

3.3 Results

Table 3 depicts a quantitative comparison of the diarisation systems (cf. Section 2). The LIUM tool achieves the lowest miss rate (6.48 %) at the expense of higher false alarm rate up to 15 %, while the sensAI system is more prone to missed speech. Conceivably, tuning of the VAD operating points could alleviate these differences. Furthermore, the LIUM and the proposed paralinguistic system provide similar performance levels regarding the speaker error rate, whereas the sensAI engine outpaces them in this metric. In the end result, sensAI yields the lowest DER, followed by the paralinguistic approach, and – slightly behind – the LIUM system. No significant difference is observed regarding the different cultural backgrounds.

To further investigate the performance difference between the proposed paralinguistic approach and the classical cepstral feature based systems, we exchange the paralinguistic trait features by the cepstral feature subset of the ComParE feature set (1 400 suprasegmental features), while keeping the simple clustering algorithm. The results (denoted as 'MFCC' in Table 3) indicate that the proposed approach is competitive despite using two orders of magnitude less features, suggesting that the paralinguistic features yield a particularly compact representation of the voice timbres. Finally, we also considered early fusion of the paralinguistic trait features with the suprasegmental MFCC features. However, the results do not improve over either of the single feature sets. This is somewhat surprising as we would expect these feature sets to carry complementary information. We speculate that a late fusion approach (system combination) could be more beneficial than early fusion due to the vastly different size of the feature sets.

Generally, it needs to be taken into account that the real-world data recorded "in the wild" comprise background and environmental noise as well as transmission characteristics, which would explain the general high level of error rates. It is therefore all the more important to improve the segmentation correctness in order to provide accurate and clean data for the downstream steps. Above all, it is highly notable that our purely paralinguisic approach is competitive against the baseline approaches which exploit cepstral features for speaker clustering – the latter can be considered highly effective on the studied data set of video calls where the transfer function differs between the interlocutors.

4 CONCLUSION AND OUTLOOK

In this work, we proposed a novel paralinguistic approach to speaker diarisation based on speaker characterisation. Taking humans as an example, our system is able to automatically assess various speaker traits for each speech segment using pre-trained models. In this way, a multi-dimensional trait vector containing the predicted age, gender, voice likability, and personality label is obtained to describe the speaker in each segment. Using these paralinguistic features, the segments are then clustered to identify unique speakers. Our results show the potential of our approach in comparison with benchmark open-source and commercial systems.

For future research, we aim to include more speaker states, traits, and speaking styles, but also linguistic features such as Bag-of-Words [33] into the paralinguistic diarisation framework. Due to the modular composition, the system components can be easily adapted and exchanged. In particular, we can also consider paralinguistic trait regression instead of binary classification, which is expected to help clustering. Furthermore, we can combine paralinguistic trait features with the GMM clustering approach as used by the best performing sensAI system. The crux is to find an appropriate fusion strategy of frame-level MFCCs with suprasegmental trait predictions. Finally, prediction uncertainty of paralinguistic trait models can be exploited to improve speech overlap detection.

5 ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Framework Programme HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA).

REFERENCES

[1] Jitendra Ajmera and Chuck Wooters. 2003. A robust speaker clustering algorithm. In *Proc. of Workshop on Automatic Speech Recognition and Understanding*. IEEE, St. Thomas, Virgin Islands, 411–416.

[2] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 2 (2012), 356–370.

[3] Xavier Anguera, Chuck Wooters, and Javier Hernando. 2007. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 7 (2007), 2011–2022.

[4] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland. 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *Proc. of ICASSP*. IEEE, Las Vegas, NV, 4353–4356.

[5] Felix Burkhardt, Martin Eckert, Wiebke Johannsen, and Joachim Stegmann. 2010. A Database of Age and Gender Annotated Telephone Speech. In *Proc. of LREC*. ELRA, Valletta, Malta.

[6] Felix Burkhardt, Björn Schuller, Benjamin Weiss, and Felix Weninger. 2011. 'Would You Buy A Car From Me?' – On the Likability of Telephone Voices. In *Proc. of INTERSPEECH*. ISCA, Florence, Italy, 1557–1560.

[7] Nicholas WD Evans, Corinne Fredouille, and Jean-François Bonastre. 2009. Speaker diarization using unsupervised discriminant analysis of inter-channel delay features. In *Proc. of ICASSP*. IEEE, Kyoto, Japan, 4061–4064.

[8] Florian Eyben. 2015. *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer International Publishing, Switzerland.

[9] Florian Eyben and Björn Schuller. 2014. openSMILE:) The Munich Open-Source Large-Scale Multimedia Feature Extractor. *ACM SIGMM Records* 6, 4 (2014).

[10] Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. of ACM Multimedia*. ACM, Barcelona, Spain, 835–838.

[11] Gerald Friedland, Oriol Vinyals, Yan Huang, and Christian A Müller. 2009. Prosodic and other Long-Term Features for Speaker Diarization. *IEEE Transactions on Audio, Speech & Language Processing* 17, 5 (2009), 985–993.

[12] Gerald Friedland, Chuohao Yeo, and Hayley Hung. 2009. Visual speaker localization aided by acoustic models. In *Proc. of ACM Multimedia*. ACM, New York, NY, 195–202.

[13] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proc. of INTERSPEECH*, Vol. 9. ISCA, Portland, OR, 2583–2586.

[14] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. 1998. Partitioning and transcription of broadcast news data.. In *Proc. of ICSLP*, Vol. 98. Sydney, Australia, 1335–1338.

[15] Jürgen T. Geiger, Florian Eyben, Björn Schuller, and Gerhard Rigoll. 2013. Detecting Overlapping Speech with Long Short-Term Memory Recurrent Neural Networks. In *Proc. of INTERSPEECH*. ISCA, Lyon, France, 1668–1672.

[16] Jürgen T Geiger, Frank Wallhoff, and Gerhard Rigoll. 2010. GMM-UBM based open-set online speaker diarization. In *Proc. of INTERSPEECH*. ISCA, Makuhari, Japan, 2330–2333.

[17] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*. Cancun, Mexico, 381–385.

[18] Gerhard Hagerer, Vedhas Pandit, Florian Eyben, and Björn Schuller. 2017. Enhancing LSTM RNN-based Speech Overlap Detection by Artificially Mixed Data. In *Proc. AES International Conference on Semantic Audio*. AES, Audio Engineering Society, Erlangen, Germany, 1–8. to appear.

[19] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] David Imseng and Gerald Friedland. 2010. Tuning-robust initialization methods for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 8 (2010), 2028–2037.

[22] Konstantin Markov. 2009. Advanced approaches to speaker diarization of audio documents. In *Proc. of Joint Conferences on Pervasive Computing*. IEEE, Taipei, Taiwan, 179–184.

[23] Sylvain Meignier and Teva Merlin. 2010. LIUM SpkDiarization: an open source toolkit for diarization. In *Proc. of CMU SPUD Workshop*. Dallas, TX.

[24] G. Mohammadi, A. Vinciarelli, and M. Mortillaro. 2010. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proc. of International Workshop on Social Signal Processing*. ACM, Florence, Italy, 17–20.

[25] Daniel Moraru, Laurent Besacier, and Eric Castelli. 2004. Using a priori information for speaker diarization. In *ODYSSEY Speaker and Language Recognition Workshop*. ISCA, Toledo, Spain, 355–362.

[26] Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka. 2012. Speaker identification and verification by combining MFCC and phase information. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 4 (2012), 1085–1095.

[27] Athanasios Noulas, Gwenn Englebienne, and Ben JA Krose. 2012. Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 1 (2012), 79–93.

[28] Scott Otterson and Mari Ostendorf. 2007. Efficient use of overlap information in speaker diarization. In *Proc. of Automatic Speech Recognition and Understanding*. IEEE, Kyoto, Japan, 683–686.

[29] John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[30] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212.

[31] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing* 10, 1 (2000), 19–41.

[32] Douglas A Reynolds and P Torres-Carrasquillo. 2005. Approaches and applications of audio diarization. In *Proc. of ICASSP*, Vol. 5. IEEE, Philadelphia, PA, 953–956.

[33] Maximilian Schmitt and Björn Schuller. 2016. openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit. *arxiv.org* 1605.06778 (2016). 9 pages.

[34] Björn Schuller. 2012. The Computational Paralinguistics Challenge. *IEEE Signal Processing Magazine* 29, 4 (2012), 97–101.

[35] Björn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley.

[36] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. of INTERSPEECH*. ISCA, Makuhari, Japan, 2794–2797.

[37] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2013. Paralinguistics in Speech and Language – State-of-the-Art and the Challenge. *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language* 27, 1 (2013), 4–39.

[38] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. 2012. The INTERSPEECH 2012 Speaker Trait Challenge. In *Proc. of INTERSPEECH*. ISCA, Portland, OR, 254–257.

[39] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. 2015. A Survey on Perceived Speaker Traits: Personality, Likability, Pathology, and the First Challenge. *Computer Speech and Language, Special Issue on Next Generation Computational Paralinguistics* 29, 1 (2015), 100–131.

[40] Sue E Tranter. 2006. Who really spoke when? Finding speaker turns and identities in broadcast news audio. In *Proc. of ICASSP*, Vol. 1. IEEE, Toulouse, France, 1013–1016.

[41] Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 5 (2006), 1557–1565.

[42] Fabio Valente, Deepu Vijayasenan, and Petr Motlicek. 2011. Speaker diarization of meetings based on speaker role n-gram models. In *Proc. of ICASSP*. IEEE, Prague, Czech Republic, 4416–4419.

[43] Ravichander Vipperla, Jürgen Geiger, Simon Bozonnet, Dong Wang, Nicholas Evans, Björn Schuller, and Gerhard Rigoll. 2012. Speech Overlap Detection and Attribution Using Convolutive Non-Negative Sparse Coding. In *Proc. of ICASSP*. IEEE, Kyoto, Japan, 4181–4184.

[44] Felix Weninger, Johannes Bergmann, and Björn Schuller. 2015. Introducing CURRENNT: the Munich Open-Source CUDA RecurREnt Neural Network Toolkit. *Journal of Machine Learning Research* 16 (2015), 547–551.

[45] Felix Weninger, Florian Eyben, Björn Schuller, Marcello Mortillaro, and Klaus R. Scherer. 2013. On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common. *Frontiers in Emotion Science* 4, 292 (2013), 1–12.