The Voice of Leadership: Models and Performances of Automatic Analysis in Online Speeches

Felix Weninger, *Member*, *IEEE*, Jarek Krajewski, *Member*, *IEEE*, Anton Batliner, and Björn Schuller, *Member*, *IEEE*

Abstract—We introduce the automatic determination of leadership emergence by acoustic and linguistic features in online speeches. Full realism is provided by the varying and challenging acoustic conditions of the presented YouTube corpus of online available speeches labeled by 10 raters and by processing that includes Long Short-Term Memory-based robust voice activity detection (VAD) and automatic speech recognition (ASR) prior to feature extraction. We discuss cluster-preserving scaling of 10 original dimensions for discrete and continuous task modeling, ground truth establishment, and appropriate feature extraction for this novel speaker trait analysis paradigm. In extensive classification and regression runs, different temporal chunkings and optimal late fusion strategies (LFSs) of feature streams are presented. In the result, achievers, charismatic speakers, and teamplayers can be recognized significantly above chance level, reaching up to 72.5 percent accuracy on unseen test data.

Index Terms—Personality analysis, dimensional analysis, acoustic/linguistic fusion

1 Introduction

Leadership and followership belong to the foundations of human society and without doubt the ability to recognize leaders and followers can be considered to be a vital aspect of human social competence. In evolutionary history, leader-follower structures evolved as a coordinated solution to challenges which could only be solved through collective efforts. Nowadays, effective leadership is still considered essential for professionals and organizations to foster productivity, financial revenue, customer satisfaction, development of human resources, and innovations [1], [2].

For the purpose of this study, it is of particular interest to determine which individuals are perceived as leaders. It has been found that, generally, followers prefer leaders who are perceived as both competent (acquiring resources for the group) and benevolent (sharing resources with the group) [3], [4]. In detail, prototypical leaders are often described by characteristics such as decisiveness (making timely and well-founded decisions), self-confidence (being able to face adversity), vision (being inspiring and charismatic), integrity (being upright, modest, and prioritizing group interest over personal ambition), and diplomacy (solving conflicts and integrating individuals into a team).

Along with many others, these traits are facets of the individual personality. Research on personality has a long tradition, leading to the now well-established Five-Factor Model [5]. For analysis of personality traits in *speech*, which is the focus of this paper, linguistic information has been used widely because self-assessment and peer-assessment of personality have mostly been conducted with the help of lists of verbal descriptors. Subsequently, these have been combined and condensed into descriptions of higher level dimensions. Metalanguage prevailed, and object language, i.e., the use of linguistic, phonetic, verbal, and nonverbal markers in the speech of subjects, was less exploited. Scherer [6] gives an overview of personality markers in speech and pertinent literature; a more recent account of the state of the art, especially on the automatic recognition of personality with the help of speech and linguistic information, and experimental results can be found in [7]. To refer to some related studies: Laskowski et al. [8] characterize participants roles in multiparty conversations; Rosenberg and Hirschberg [9] deal with acoustic/prosodic and lexical correlates of charismatic speech. Gregory and Gallagher [10] demonstrate that US presidential election outcomes can be predicted on the basis of spectral information beneath 0.5 kHz. Nass and Lee [11] experiment with computer-synthesized speech expressing personality. In the field of personality assessment from *text*, Argamon et al. [12] find that the use of appraisal predicts neuroticism and that function words are indicative of extraversion; furthermore, Oberlander and Nowson [13] employ textual features for personality classification of weblogs. In contrast, Mohammadi et al. [14] propose purely acoustic features for personality assessment in radio broadcasts and Metze et al. [15] point out the opportunities for automatic personality analysis in human-machine interaction.

- *F. Weninger, A. Batliner, and B. Schuller are with the Institute for Human-Machine Communication, Technische Universität München, Arcisstraße 21, München 80333, Germany. E-mail: {weninger, schuller}@tum.de, anton.batliner@lrz.uni-muenchen.de.*
- *J. Krajewski is with the Department of Experimental Industrial Psychology, Schumpeter School of Business and Economics, Bergische Universität Wuppertal, Gaußstraße 20, Wuppertal 42119, Germany. E-mail: krajewsk@uni-wuppertal.de.*

Manuscript received 4 Oct. 2011; revised 13 Mar. 2012; accepted 21 May 2012; published online 29 May 2012.
Recommended for acceptance by S. Narayanan.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2011-10-0066.
Digital Object Identifier no. 10.1109/T-AFFC.2012.15.

Building on those previous results, we propose a speech-based system which can automatically determine leadership emergence—i.e., whether speech from an individual is perceived as leader-like—by means of linguistic evidence from automatic speech recognition (ASR) in combination with acoustic analysis. Such automatic systems could help to avoid cost intensive observer ratings [16] in the context of human resources, and enable automatic voice coaching. Furthermore, we expect that the perceived "social competence" of robots and other technical systems can be further advanced if we can make them understand which people are leaders and which are followers, and adapt their discourse and interaction strategies accordingly. Other promising applications of automatically detecting leadership qualities are found in the multimedia and entertainment sector, e.g., by enhancing archives of online speeches by "tags" indicating traits such as charisma, self-confidence, or integrity, or synthesizing leader-like voices for avatars in computer games.

In light of this broad application potential, we strive to evaluate speech-based leadership recognition in real-life settings. To this end, we collected the YouTube corpus, a large corpus of online speeches from YouTube that were annotated in 10 dimensions of leadership, such as charisma, self-confidence, and diplomacy, by expert annotators (Section 2). We comment on the issues of correlated dimensions and annotation reliability (Sections 2.2 and 2.3), then move forward to concepts for fully automatic analysis: In Section 2.4, we present our approach for robust segmentation by long short-term memory recurrent neural networks (LSTM-RNN) and the classification and regression tasks used for evaluation (Section 2.5). Methods for acoustic-linguistic analysis by late fusion are outlined in Section 3, briefly discussing the measures for their evaluation in Section 4. Parameterization on the development set and final evaluation on the test set of the YouTube corpus are fleshed out in Section 5 before concluding in Section 6.

2 YouTube Corpus: A Database of Online Speeches

2.1 Data Collection

The YouTube corpus consists of 409 recordings, each about one minute long, from 143 speeches available on YouTube (143 male executives within the age range of about 20-75 years, mean = 51.1, standard deviation = 12.1). For those nine speeches where either the exact date of the speech or the speaker's age could not be determined, age was estimated by 10 annotators and the mean value of all annotators was considered in further analyses. Moreover, the speakers' age was not related to any perceived leadership dimension and could, therefore, not be considered as a relevant confounding factor. While a minority of 22 speeches (15.4 percent) seems to be read from a script, the remaining speakers presented either without any notes or only based on presentation slides. In addition, no differences in perceived leadership characteristics were found when comparing scripted and nonscripted speech.

As the approach of this study is to assess perceived leadership dimensions based on voice characteristics, the samples of the YouTube corpus were collected to represent persons with significant leadership abilities. The functions of the speakers can be summarized as follows: The vast majority (89.5 percent, or 128 speeches) are taken from top executives of "global players" (mostly derived from the Forbes Global 2000 list in 2010). The remaining 15 speeches are composed of leaders of nonprofit organizations (6), entrepreneurs (5), university professors (3), and one football team captain. Most speeches were derived from public presentations such as introduction of new products (100 or 69.9 percent), outlining future prospects (19 or 13.3 percent), or summarizing recent developments (11 or 7.7 percent). The remaining 13 speeches include interviews and university lectures. Although in [17] it has been shown that apart from rather invariant speaker characteristics, different speech settings and authorships of speeches affect linguistic features, we did not explicitly control for such possible confounders since our study does not aim at assessing invariant personality traits, but rather at the subjective—and possibly time variant—impression of leadership, which should not be masked by these uncontrolled confounders.

In order to obtain a nearly equal amount of data per speaker, not more than three recordings per speech were extracted. These recordings will be subsequently referred to as *tracks*. The speech signal was recorded with different microphones and qualities, mainly with a 16 kHz sampling rate. The recordings took place in lecture-rooms under varying levels of noise and reverberation (microphone-to-mouth distance > 0.3 m). The corpus was annotated by 10 raters (PhD students of psychology, five males, five females) aged between 23 and 58 years (mean = 36.9). Gender effects were controlled by Krippendorff's α and significance testing, but no significant effects could be revealed. All raters had been formally trained to apply a Likert scale on a standardized set of judging criteria and are experienced both in leadership research and in rating of all the dimensions which were used. First trainings were conducted with the assistance of unambiguous samples (negative and positive ones). This is considered sufficient training since the ratings are supposed to represent intuitive perception to gain the best possible external validity.

All rating dimensions were derived from the culturally endorsed leadership (CLT) questionnaire [18], which is a highly validated, commonly applied rating instrument (e.g., [19], [20]) and widely considered as the leading cross-cultural leadership approach [21]. Each rater assigned an integer value from 1 ("not at all") to 5 ("very") to each of 10 dimensions, for which the following associated descriptors were given to the raters: *charismatic* (fascinating, captivating, winsome), *visionary* (stimulating, future oriented, far-sighted), *inspiring* (positive, dynamic, building confidence), *upright* (trustworthy, reliable, being of integrity), *team-integrating* (integrative, informing, team building), *nonmalicious* (benevolent, smart, anticipating), *diplomatic* (good negotiating, accomplishing best conditions, effective), *decisive* (fast decisions, determined, stringent argumentation), *performing* (improvement oriented, demanding excellence, active), and *self-confident* (professional, not nervous, not submissive). The order of tracks was randomized. In case the rater was unsure, he or she

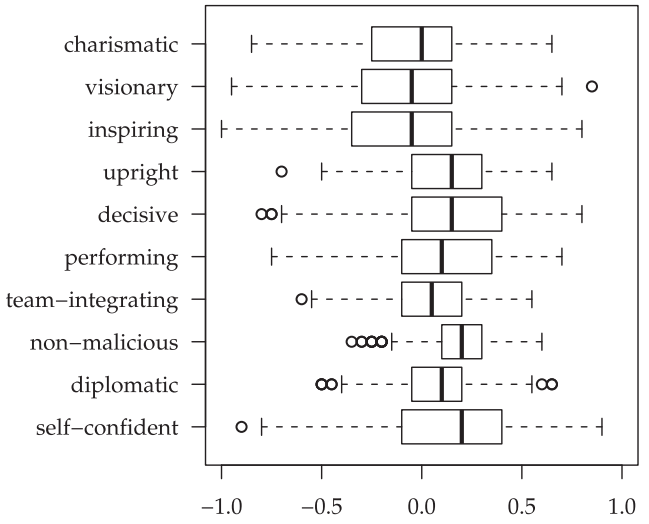


Fig. 1. Visualization of the distribution of the mean rating (for each dimension) across the instances as a box-and-whisker plot, with circles indicating outliers.

TABLE 1
Correlation Matrix between Inter-Rater Mean of the Original 10 Dimensions ($Deci = Decisive$)

	self-confident	*performing*	*team-integrating*	*upright*	*inspiring*	*charismatic*	*visionary*	*diplomatic*	*non-malicious*
deci	.90	.89	.65	.72	.83	.84	.86	.46	-.25
self		.81	.64	.73	.81	.81	.82	.53	-.19
perf			.59	.59	.70	.74	.76	.38	-.31
team				.76	.71	.74	.68	.60	.27
upri					.74	.77	.73	.68	.22
insp						.89	.94	.42	-.05
char							.91	.50	-.01
visi								.41	-.14
dipl									.31

assigned a symbol for a missing value ($\perp$). The validity of the applied dimensions and rating instruments have been proven in several studies [19], [22], [23], [24].

All ratings $\in \{1, \ldots, 5\}$ were mapped onto

$$\mathcal{C} := \{-1, -0.5, 0, 0.5, 1\} \tag{1}$$

by means of the linear mapping $c \mapsto (c-3)/2$, in order to ensure compatibility of our results with the standard representation of continuous dimensions, which mostly are conceptualized as coordinate systems with an origin.

We now formally introduce the terms related to the annotation. $c_{i,r}^{\text{dim}} \in \mathcal{C}$ shall denote the value for dimension dim that rater r gave to track i. For each track i and dimension dim, $R_i^{\text{dim}} \subseteq \{1, \ldots, 10\}$ specifies the raters that assigned a label to it, i.e., $c_{i,r}^{\text{dim}} \neq \perp$. Thus, for a dimension dim and a track i, the *mean rating* can be defined as follows:

$$\overline{c}_i^{\text{dim}} = \frac{1}{|R_i^{\text{dim}}|} \sum_{r \in R_i^{\text{dim}}} c_{i,r}^{\text{dim}} \in [-1; +1], \tag{2}$$

where $\overline{c}_i^{\text{dim}}$ corresponds to the maximum likelihood estimator of the true label of track i assuming that the rating of each rater is corrupted by additive Gaussian noise [25]. Calculating the mean rating from the ordinal ratings enables a quasi-continuous scale, taking into account that leadership traits—similarly to personality—are best represented in continuous dimensions, while observer ratings are typically performed on discrete valued ordinal scales.

The distributions of the mean ratings $\overline{c}_i^{\text{dim}}$ for each dimension dim are shown in Fig. 1 as a box-and-whisker plot [26]: Boxes range from the first to the third quartile; all instances i with $\overline{c}_i^{\text{dim}}$ exceeding that range by more than 1.5 times the width of the box are considered outliers, depicted by circles. While the mean rating distribution shows a somewhat strong tendency toward the scale center, a more in-depth analysis of the ratings yields that, on average, a range of 3.8 (on the original Likert scale from 1 to 5), or 95 percent of the available rating scale, have been used. In addition, at least 10 percent of all ratings are located at the extrema, and these extreme ratings are evenly distributed among the ratings of all raters. Finally, the low rate of missing values in the ratings (0.9 percent or 36 of 4,090 ratings) provides indirect evidence for high rater confidence. These missing values are not focused on certain samples or dimensions but appear to be random.

2.2 Cover Dimensions

Table 1, considering only those tracks that were later assigned to the training or development set (Section 2.4), shows that the 10 dimensions annotated are more or less correlated with each other, indicated by different gray values: the darker, the higher correlated. Thus it seems reasonable to assume some few, more basic dimensions. In order to obtain such dimensions in a data-driven and cluster-preserving way, we computed nonmetrical multidimensional scaling (NMDS) solutions [27], based on this correlation matrix. Using all 10 dimensions reveals that *nonmalicious* is rather isolated from all other dimensions, cf., the low correlation values in Table 1. We therefore computed another 2D NMDS solution using all dimensions but *nonmalicious*, yielding the representation depicted in Fig. 2 with very good quality (stress $= 0.07$, RSQ $= 0.98$).

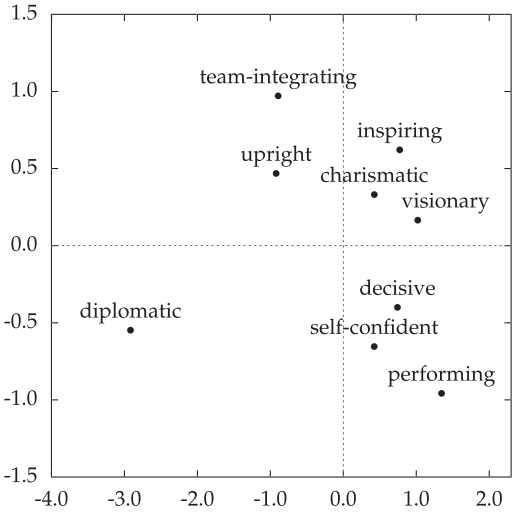


Fig. 2. Configuration of the derived stimulus (euclidean distance model) by nonmetrical multidimensional scaling on the interdimension correlation matrix.

TABLE 2
Average Rater Agreement with the Inter-Rater Mean, Measured by Correlation (ϱ), Kappa, and Alpha Reliability

Dimension	ϱ	κ	κ^1	κ^2	α	α^1	α^2
charismatic	.57	.17	.31	.46	.13	.28	.45
visionary	.58	.16	.31	.47	.11	.28	.46
inspiring	.57	.13	.29	.44	.08	.24	.41
upright	.51	.16	.26	.38	.12	.22	.36
decisive	.62	.19	.34	.51	.16	.32	.50
performing	.59	.18	.32	.47	.15	.29	.45
team-integrating	.46	.12	.20	.31	.06	.15	.28
non-malicious	.41	.17	.22	.29	.14	.19	.27
diplomatic	.47	.11	.20	.30	.04	.14	.28
self-confident	.62	.18	.35	.52	.16	.33	.51
ACHIEVER	.64	.23	.37	.53	.21	.36	.52
CHARISMATIC	.59	.17	.32	.48	.13	.30	.47
TEAMPLAYER	.51	.17	.26	.37	.14	.23	.34

κ^1, κ^2, α^1, and α^2 are weighted versions of Kappa and Alpha (absolute/squared label difference).

The four quadrants can be interpreted as representing archetypal personalities of leaders: *charismatics* (first quadrant), *achievers* (second quadrant), *diplomats* (third quadrant), and *teamplayers* (fourth quadrant). Clustered dimensions (ACHIEVER, CHARISMATIC, TEAMPLAYER) were obtained by averaging the ratings for the dimension in each cluster. The clustered dimensions can be recognized within the prototypical leadership characteristics mentioned in the introduction and can be roughly assigned to the classical dimensions of task- versus people-oriented leadership [28]. We will subsequently refer to the dimensions ACHIEVER, CHARISMATIC, TEAMPLAYER, NONMALICIOUS, and DIPLOMATIC as "cover dimensions," denoted by SMALL CAPS.

2.3 Assessment of Annotation Reliability

To measure the reliability of the annotation in the YouTube corpus, we considered the average agreement of a single rater with the mean rating. Taking into account the presence of missing values, multiple raters, and the ordinal scale of the rating, we decided for (weighted) α statistics [29] and the correlation coefficient (CC) ϱ. For reference, we also provide unweighted (Cohen's) and weighted κ. Weighted α and κ (α^1, α^2, κ^1, and κ^2) use the absolute value of disagreement, $|c_{r_1}^{\dim} - c_{r_2}^{\dim}|$, or its square as metrics to better reflect ordinal dependencies. Note that all of these measures are independent of the scaling of the ratings.

For the purpose of calculating Kappa and Alpha statistics of a rater versus the mean rating, the mean rating (2) is mapped onto the nearest value in $\mathcal{C}$ (1) through a function $\mathbb{R} \to \mathcal{C}$ given by

$$c \mapsto \frac{1}{2}\left\lfloor 2c + \frac{1}{2} \right\rfloor.$$

The average agreement of the individual raters with the inter-rater mean is shown in Table 2. Overall, despite the high correlation of the dimensions, their reliability considerably differs. Furthermore, for every single kind of measure, the ACHIEVER cover dimension displays the highest average rater agreement with the inter-rater mean. On the other side of the scale, there is low agreement on the *nonmalicious*, *diplomatic*, and *team-integrating* dimensions.

2.4 Automatic Segmentation: From Track to Chunk Level

In this paper, we constrain ourselves to static classification and regression using segment-wise features, which we will describe in detail in Sections 3.1 and 3.2. As a result, there is a need for adequate segmentation: Segment-wise analysis can be applied to entire tracks, thus taking into account long-range context, but arguably also losing information about short-time variations of the features. Besides, a smaller unit of analysis, termed *chunk level* in the following discussion, can also be motivated from a machine learning point of view since using chunk level features arguably enables more stable classifier training due to the increased number of instances and more meaningful functionals. Then, a track level prediction—which is most important considering possible applications—can be established by fusing chunk level predictions, as discussed in Section 3.4.

Tracks were split into chunks through automatic voice activity detection (VAD). Relying on VAD disposes of the need for a ground truth transcription, which is required for more elaborate schemes based on syntactic and prosodic criteria [30], but is not available for a real-life system that is applied "in the wild" to unknown data. We found in a preliminary study on the training set that due to the challenging acoustic conditions in the database, particularly varying levels of reverberation and noise, a simple energy threshold was not appropriate for segmentation due to inaccurate recognition of speech pauses. Thus, we implemented a VAD using the output of a Long Short-Term Memory recurrent neural network [31]. LSTM-RNNs are able to take into account arbitrary amounts of context from earlier feature and prediction vectors; as a consequence, they are able to adapt to instationary background noise, as could be demonstrated, e.g., in [32].

We trained an LSTM-RNN on a modified version of the TIMIT database: The recordings of the TIMIT training set were split speaker-independently into training (3,326 utterances) and validation set (370 utterances) and were overlaid with noise (babble and street noise from the Aurora database [33]) at signal-to-noise ratios from 0 to 30 dB after adding silence of random length (0 to 2 seconds) at the beginning and end. Twelve perceptual linear prediction (PLP) features, along with first and second order regression coefficients, were extracted using our open-source feature extractor openSMILE [34]. The LSTM-RNN had one input layer of size 36 (input feature vector size), one hidden layer with 200 LSTM cells, and one output layer with one output indicating the posterior voicing probability. From the manually phone aligned transcripts, we generated a binary voicing ground truth to be used as the target for network training by mapping all phones to 1 and silence to 0. Training was performed by gradient descent with early stopping once the error on the validation set had not decreased for more than 40 iterations. The correlation coefficient between the voicing ground truth and the posterior voicing probability output by the LSTM-RNN is 0.868, at a mean linear error (MLE) of 0.123, on the noisy data created from the TIMIT test set, indicating robust segmentation in challenging conditions.

The trained network was applied to the audio tracks in the YouTube corpus, and tracks were cut at pauses which were

TABLE 3
YouTube Corpus: Train, Develop(ment), and Test Set, and Corresponding Numbers of Instances (#) on Track (tr.) and Chunk (ch.) Level as well as Number (#) of Chunks per Track; Mean ± Standard Deviation

set	# tr.	length [s]	# ch.	length [s]	# ch. / tr.
Train	167	60.0 ± 1.6	1 740	5.2 ± 6.1	10.4 ± 5.2
Develop	125	60.7 ± 0.4	1 281	5.3 ± 6.8	10.2 ± 4.9
Test	117	63.0 ± 4.1	1 272	5.4 ± 6.8	10.9 ± 6.3
Σ	409		4 293		

indicated by the output of the neural network as staying below a threshold of 0.2 for longer than 500 ms. Finally, we divided both tracks and chunks into a training (TR), development (DE), and test set (TE). The partitioning was chosen to strictly enforce speaker independence, as needed in most real-life applications. For easy reproducibility, the subdivision was performed by ordering the speaker IDs in ascending numeric order and assigning the first 57 (≈ 40 percent) of the 143 speakers to the training, the next 43 (≈ 30 percent) to the development, and the remaining 43 (≈ 30 percent) to the test set. We found that this partitioning also provides for stratification by speaker age. The resulting number of tracks and chunks is shown in Table 3. Note that we do not preselect "friendly" instances, such as instances with a high rater agreement, for evaluation. Rather, in line with recent studies in paralinguistic information retrieval (e.g., [35]), our goal is to design a system that robustly classifies all available data as needed for a system operating "in the wild."

2.5 Task Definition: Regression versus Classification

As mentioned above, the mean rating per track, $\overline{c}_i^{\dim}$, provides a natural target for regression in the feature space. On the other hand, it can be argued that in practical applications, for example, automatic tagging of audio archives, an exact assessment is not required; rather, a binary decision such as *charismatic/noncharismatic* is adequate. Furthermore, the latest series of INTERSPEECH Challenges dealing with recognition of speaker states and traits from speech in real-life conditions have shown that such tasks become robustly tractable when reduced to a reasonably limited number of classes [36], [37], [38], [39].

Thus, we additionally created binary classification tasks for each dimension to discriminate between high- and low-rated instances by binarizing the quasicontinuous mean ratings in accordance with recent evaluation campaigns in the field [39], [40]. Each instance i was assigned a "positive" label (1) for dimension $\dim$ whenever $\overline{c}_i^{\dim}$ was ~~below~~ the sample median of mean ratings $\tilde{\overline{c}}^{\dim}$ in the union of training and development set, or a "negative" (0) label in case that $\overline{c}_i^{\dim} < \tilde{\overline{c}}^{\dim}$. The choice of the sample median of means as threshold binarizes the quasi-continuous rating given by the mean ratings in a natural way, enforcing balanced training with the union of training and development set—this disposes of the need for upsampling or other techniques that are often applied to prevent a classifier bias toward the majority class. It is left to assign all instances i with $\overline{c}_i^{\dim} = \tilde{\overline{c}}^{\dim}$ to either a positive or negative label; we simply chose the option that minimizes class imbalance among the union of training and development set. Note that our definition of the classification problem does not guarantee a balanced development or test set: In particular, the imbalance of the test set—measured as the ratio of majority class over minority class instances—is highest for NONMALICIOUS (1.85), followed by DIPLOMATIC (1.29); in contrast, it is low (< 1.2) for the three cover dimensions.

TABLE 4
The Official 1,582D Acoustic Feature Set of the INTERSPEECH 2010 Paralinguistic Challenge: 38 Low-Level Descriptors with Regression Coefficients, 21 Functionals. Abbreviations: DDP: Difference of Difference of Periods, LSP: Line Spectral Pairs, Q/A: Quadratic, Absolute

Descriptors	Functionals
PCM loudness	max. / min. (position)
MFCC [0–14]	arith. mean, std. deviation
log Mel Freq. Band [0–7]	skewness, kurtosis
LSP [0–7]	lin. regression slope, offset
F0 by Sub-Harmonic Sum	lin. regression error Q/A
F0 Envelope	quartile 1 / 2 / 3
Voicing Probability	quartile range 2–1 / 3–2 / 3–1
Jitter local	percentile 1 / 99 (≈ min. / max.)
Jitter DDP	percentile range 99–1
Shimmer local	up-level time 75 / 90

3 Methods for Automatic Analysis

Having defined concrete tasks for automatic analysis based on the characteristics of the YouTube corpus, we now proceed to describe how these can actually be solved. To this end, we describe baseline methods for acoustic and linguistic analysis, and propose an effective method to combine these modalities by means of late fusion.

3.1 Acoustic Analysis: Relevant Low-Level Descriptors (LLDs) and Functionals

Our approach to acoustic feature extraction includes features reported in the literature as relevant in leadership-related contexts, and at the same time relies on a publicly available feature set for reproducibility. Thus, for all experiments, the full, 1,582D feature set given for the INTERSPEECH 2010 Paralinguistic Challenge [37] was extracted. Features are obtained by extracting low-level descriptors at 100 frames per second using window sizes from 25 to 60 ms, then applying track- or chunk-wise functionals (cf., Table 4) intended to capture time variation in a single feature vector that is independent of the length of the speech signal. Low-level descriptors include spectral features (which were associated in [10] with election outcomes), cepstral features (describing timbre of the voice, which is relevant for likability [41]), prosodic features including loudness and fundamental frequency (F0) that are known to be related to extroversion [6] and charisma [9], and, finally, voice quality features, including jitter and shimmer, to characterize the "roughness" of the voice. The LLDs are smoothed by moving average low-pass filtering with a window length of three frames, and their first order regression coefficients [42] are added.

This "brute-force" combination of LLDs and functionals yields 16 zero information features which are discarded, e.g., minimum F0 (always zero). Finally, two single features, the number of F0 onsets and turn duration, are added. These indicate the number of voiced segments and the duration

TABLE 5
Feature Relevance: Selected Track-Wise Functionals of LLDs by Correlation Coefficient with the Mean Rating and t Statistic against the Binary Labels for ACHIEVER, CHARISMATIC, and *Nonmalicious*

LLD	Functional	CC	t
Δ MFCC 4	Quartile range 1–3	.389	7.09
Δ LSP 3	Quartile range 1–2	.388	7.07
Voicing prob.	percentile 99	.347	6.96
PCM loudness	Skewness	-.335	-4.15
PCM loudness	Quartile range 1–3	.334	6.01
Δ F0 Envelope	Quartile range 1–2	.323	4.88
Δ PCM loudness	Kurtosis	-.322	-3.34
log. MFB 2	up-level time 90 %	.249	4.64
F0 by SHS	Quartile 2 (median)	.268	5.59
PCM loudness	Std. dev.	.236	4.64
PCM loudness	Arith. mean	.217	4.07

(a) ACHIEVER

LLD	Functional	CC	t
Voicing prob.	Percentile 99	.369	6.96
Δ LSP 3	Quartile range 1–2	.324	5.23
Δ MFCC 4	Quartile range 1–3	.308	5.87
Δ F0 Envelope	Quartile range 1–3	.259	4.97
F0 by SHS	Quartile 1	.251	4.46
log. MFB 2	up-level time 90 %	.250	3.22
Δ PCM loudness	Kurtosis	-.221	-2.06
PCM loudness	Arith. mean	.141	1.45

(b) CHARISMATIC

LLD	Functional	CC	t
F0 by SHS	Linear regr. offset	-.329	-4.39
PCM loudness	Quartile 2 (median)	-.232	-4.24
Voicing prob.	Quartile 1	-.208	-3.14
Jitter local	Linear regr. offset	-.191	-2.64

(c) *non-malicious*

between speech pauses. For straightforward reproducibility, we use our open-source feature extractor openSMILE [34] that also provided the features for the Challenge [37].

We verified the relevance of the extracted acoustic features on the proposed evaluation database. The correlation coefficient of selected features (LLDs and track-wise functionals) with the mean rating across all instances as well as their t-test score against the binary labels is shown in Table 5. For achievers (Table 5a), we observe a higher variation in speech in general, as indicated by the importance of quartile range functionals. More specifically, the change (deltas) in MFCC and LSP features, somewhat corresponding to phonetic content, has wider range, which can be interpreted as achievers varying their articulation stronger. PCM loudness seems to be strongly associated with achievers; interestingly, among the functionals of PCM loudness, skewness shows strong negative correlation with the mean achiever rating, probably indicating that achievers make more targeted pauses (negative skewness of the energy envelope). In contrast, rough voices (low maximum voicing probability) characterize nonachievers somewhat, as well as high kurtosis of loudness changes (i.e., presence of sharp changes). Further, we observe that "standard" functionals of PCM loudness such as standard deviation and arithmetic mean show considerably lower correlation with the ACHIEVER rating than the quartile range, which can be due to the increased noise robustness of the latter. Besides, while achievers generally seem to have higher median F0 as expected, the dynamic range of their F0 (quartile range of deltas) is even more strongly associated with the achiever ratings in our data. Finally, our findings corroborate [10] since the 90 percent up level time of the Mel frequency band 2, ranging from 166 to 341 Hz, is correlated with the notion of achievers.

Second, regarding charismatics (Table 5b), we see on the one hand a considerable overlap of relevant features with the achievers, but correlations are lower, except for maximum voicing probability. While the correlation of F0 with charisma is corroborated by Rosenberg and Hirschberg [9], the lower correlations than for ACHIEVER generally indicate that charismatic speech is a more multifaceted phenomenon than "achiever" speech. In particular, we see that the mean PCM loudness is much weaker correlated with charismatics than with achievers.

Third, for the *nonmalicious* dimension (Table 5c), we observe that only a few features are significantly correlated with the ratings. However, these seem to characterize well the nature of this dimension: Apparently, the presence of "shrieking," i.e., high F0, PCM loudness, voicing probability, and jitter, is judged by the raters as a sign of maliciousness. Finally, we see that some features are best suited to coarse classification of the extremes, exhibiting high t-values but lower CC values (e.g., median F0 for ACHIEVER), and conversely, others enable fine-grained assessment, e.g., skewness of PCM loudness for ACHIEVER.

All correlations reported in Table 5 are significant at the 0.001 level. In summary, we conclude that the INTERSPEECH 2010 Paralinguistic Challenge feature set is very well suited to capturing the features relevant for leadership traits.

3.2 Automatic Linguistic Analysis

In addition to the acoustic features, we considered linguistic features in the shape of bag-of-words (BoW) vectors. To strictly enforce realism, we obtained these features by ASR. An ASR engine was built on top of HDecode [42], using $3 \cdot (12 + 1)$ PLP features along with short-time energy and first and second order regression coefficients in a hidden Markov model (HMM) framework. Thirty-nine monophones and silence were represented in three-state left-to-right HMMs with 16 Gaussian mixtures (32 for silence). The initial monophone models with a single Gaussian mixture were trained using four iterations of embedded Baum-Welch re-estimation. After that, the monophones were mapped to tied-state cross-word triphone models with shared state transition probabilities. Two Baum-Welch iterations were performed for reestimation of the triphone models. Finally, the number of mixture components of the triphone models was increased in successive rounds of mixture doubling and reestimation (four iterations in every round).

Since the investigated speeches unite characteristics of both read and spontaneous speech, the training data for the acoustic models consisted of the union of the *Wall Street Journal* (*WSJ*) and Buckeye [43] corpora, using the segmentation described in [44] for the latter. Finally, a back-off trigram language model was built from all the 778 transcripts of public speeches available at the TED talks website[1] as of

1. www.ted.com/talks.

December 2010 (2.0 million words), in order to ensure good adaptation of the language model to the target domain, resulting in a vocabulary size of 30.6 K. To ensure consistency between chunk and track level linguistic features, decoding was first performed on the chunk level, and transcriptions were concatenated to form a transcription on track level.

Since, in a fully realistic setting, ground truth transcripts of online speeches are generally not available, we primarily evaluate the performance of the resulting BoW features rather than measuring ASR performance directly in terms of word accuracy. In fact, it has been shown that, e.g., text classification is robust against ASR errors [45]. Still, to complement our task-based evaluation of ASR, we obtained a rough ASR performance estimate by manually transcribing a randomly selected subset of 30 speeches (spread across training, development, and test set). The obtained word accuracy in these speeches ranges from -20 to 52 percent, reflecting the challenge of the ASR task, which we attribute mainly to the varying reverberant and partly noisy acoustic conditions which lead to a considerable amount of word insertion errors, and the spontaneous, nonscripted speech. To mitigate the effect of the erroneous ASR, we replaced all words whose confidence measure was below a threshold of 10 percent of the average confidence on the training and development set by a marker word (LC for "low confidence").

Finally, BoW vectors were generated from the words that occurred in the ASR transcript of the training and development set with a minimum term frequency of 3, resulting in a BoW size of 859. Note that the BoW vectors also include the frequencies of the LC word—this feature could help in determining the "intelligibility" of a speech, as good speakers are more likely to produce high acoustic (and language model) likelihoods due to clarity in articulation and syntax, aside from recording artifacts.

3.3 Training of Low-Level Classifiers

In line with the choice of acoustic and linguistic features which have been thoroughly explored in paralinguistics research, we relied on well-proven classifiers as well: We opted for the setups used for the baselines of the INTERSPEECH 2009 Emotion Challenge (2-class task) and INTERSPEECH 2010 Paralinguistic Challenge (Affect Subchallenge). A key part of our study will be to combine the low-level classifiers by late fusion, as laid out in the next section.

In particular, the binary low-level classifiers are support vector machines (SVMs) with a linear kernel, trained using the sequential minimal optimization (SMO) algorithm [46] on normalized features. SVMs have been selected for classification in this study as they are well suited to the large acoustic and linguistic feature sets due to their robustness against overfitting—their complexity does not depend on the number of features; furthermore, linear SVM are known to be well suited to classification by linguistic features [47]. The complexity constant for the SMO training algorithm was set to 1.0. For regression tasks, we used unpruned REPTrees with 25 cycles in Random-Sub-Space metalearning [48] (500 iterations, subspace size 5 percent). Since this regression algorithm builds a large number of regression trees, each on a small feature subspace, it is suitable for high-dimensional feature spaces, and furthermore outperformed Support Vector Regression in a preliminary experiment. To enforce transparency of results, all experiments were based on the classifier implementations found in the WEKA toolkit [49].

In the following, the decision of a low-level classifier trained on acoustic features will be denoted by $d^{\dim}_{\mathrm{ac,ch}}(i)$ and $d^{\dim}_{\mathrm{ac,tr}}(i) \in \{0,1\}$ for chunk and track level, respectively. For classifiers trained on linguistic features, the notations $d^{\dim}_{\mathrm{lng,ch}}(i)$ and $d^{\dim}_{\mathrm{lng,tr}}(i)$ will be used.

3.4 Late Asynchronous Acoustic/Linguistic Fusion

Based on the above low-level classifier setup, late fusion strategies (LFSs) were designed, taking into account the following: First, in the targeted application scenario, a prediction for each track has to be deduced, which is also the level that annotation was performed on; thus, it is necessary to combine the chunk-level decisions onto track level. There is a straightforward strategy for this: In case of regression, we take the mean regressor output, while for the classification tasks we perform a majority vote.

Second, and more interestingly, it is desired to integrate acoustic and linguistic information. We opted for a late (decision-level) fusion as this allows us to integrate the fusion of chunk level decisions with the fusion of acoustic and linguistic information. As it is not clear which unit of analysis provides the best tradeoff between the predictive power of features and providing enough data for the classifier, we can let both the acoustic and linguistic information be processed, each on chunk or track level, independently or *asynchronously*, and derive for each possible combination a decision function as shown below. Thus, our fusion methods go beyond merging the outcomes of two classifiers, each operating on the same data. We will evaluate each type of strategy on the development as well as the test set in Section 5.3.

For the sake of clarity, we constrain the following discussion to classification, as regression on chunk level delivered unsatisfactory performance in our first experiments on the development set (Section 5.1, Table 6); yet, the methodology can be easily extended to regression or classification with confidences. First, the decision function for track i when fusing acoustic information on chunk level and linguistic information on track level is defined by

$$d^{\dim}_{1}(i) = \frac{1}{1+\lambda_1}\left[\lambda_1 d^{\dim}_{\mathrm{lng,tr}}(i) + \frac{1}{|\mathrm{CH}(i)|}\sum_{j\in \mathrm{CH}(i)} d^{\dim}_{\mathrm{ac,ch}}(j)\right], \quad (3)$$

where $\mathrm{CH}(i)$ is the set of chunks that track i consists of. The fused class decision is then 1 if and only if $d^{\dim}_{1}(i) > 0.5$. Note that the parameter λ_1 roughly resembles the weight factor that is commonly used in ASR to premultiply language model likelihoods when total acoustic and linguistic likelihoods are calculated. Precisely, λ_1 is the weight that the linguistic classifier decision is given with respect to the majority vote among acoustic, chunk-level classifiers. If $\lambda_1 = 0$, the class decision is equal to decision by majority vote among chunk level acoustic classifiers. Conversely to (3) above, acoustic information on track level and linguistic information on chunk level is fused by

$$d_2^{\dim}(i) = \frac{1}{1+\lambda_2}\left[d_{\text{ac,tr}}^{\dim}(i) + \frac{\lambda_2}{|\text{CH}(i)|}\sum_{j\in\text{CH}(i)} d_{\text{lng,ch}}^{\dim}(j)\right], \quad (4)$$

where λ_2 is the weight of the linguistic majority vote with respect to the acoustic classifier decision on track level. Finally, considering both types of information on chunk level results in

$$d_3^{\dim}(i) = \frac{1}{1+\lambda_3}\left[\frac{1}{|\text{CH}(i)|}\sum_{j\in\text{CH}(i)}\left(d_{\text{ac,ch}}^{\dim}(j) + \lambda_3 d_{\text{lng,ch}}^{\dim}(i)\right)\right]. \quad (5)$$

The decision resulting from $d_3^{\dim}$ corresponds to a weighted majority vote on all chunks, where the weight of any linguistic decision with respect to any acoustic decision is given by λ_3.

Note that if one assumes hard class decisions (0 or 1) for $d_{\text{lng,tr}}^{\dim}$, λ_1 should be chosen ≤ 1 since otherwise the outcome $d_1^{\dim}(i)$ (3) will be equal to the decision of the linguistic classifier; conversely, for LFS2 (4), $\lambda_2 \geq 1$ if $d_{\text{ac,tr}}^{\dim}$ is a hard class decision.

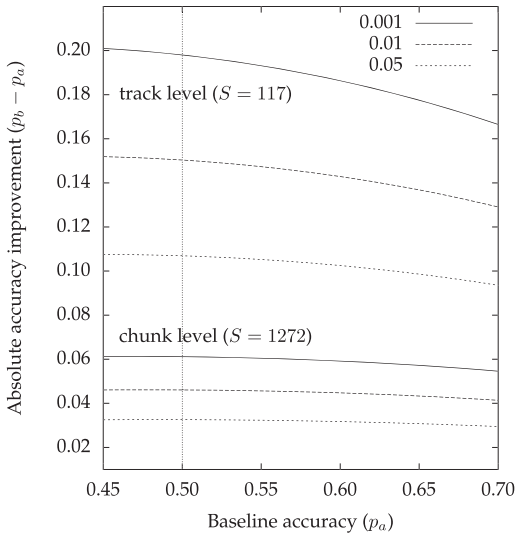


Fig. 3. Lines of significant absolute accuracy improvements for different levels of significance (0.001, 0.01, or 0.05), for experiments on the test set of the YouTube corpus, on track level (sample size $S = 117$) or chunk level ($S = 1{,}272$). Vertical line: chance level accuracy (0.5).

4 Evaluation Measures and Significances

4.1 Performance of Classification and Regression

Before discussing the performance of automatic analysis in detail, let us first clarify employed performance measures. Our primary evaluation measure for classification is unweighted average recall (UAR) which is tailored to imbalanced problems—remember that the test set is imbalanced for some dimensions and optimizing on accuracy may introduce a bias towards picking the majority class. For the two-class problems considered in this study, this measure simply reads

$$\text{UAR} = \frac{\text{Recall of Class '0'} + \text{Recall of Class '1'}}{2}.$$

UAR has been the competition measure of the INTERSPEECH 2009-2012 Challenges dealing with paralinguistic phenomena [36], [37], [38], [39]. We additionally consider conventional accuracy for reference.

For evaluation of regression tasks, we rely on correlation coefficient between the outputs of the regression function and the corresponding target values and mean linear error, which is the expected absolute difference. The CC is a scale-independent measure that quantifies whether a high rating results in high prediction and vice versa, while the MLE is the expected absolute deviation of the prediction from the mean rating, thus being scale-dependent and penalizing overshooting as well as underestimation. These are standard evaluation measures in recognition of paralinguistic information from speech (cf., the INTERSPEECH 2010 Paralinguistic Challenge's Affect Subchallenge [37]) and machine learning in general [50]. While our goal is to recognize how well-automated predictions reflect the best possible consensus of annotators derived from their ratings, the evaluation by CC with the mean rating can also be interpreted from a different perspective: The CC of the prediction and the mean rating is equivalent to the expected CC of the prediction and any individual rater, assuming that the mean rating has the same standard deviation as any individual rating—the latter could be trivially established by scaling.

4.2 Significance Testing for Classification

Statistical significance testing is especially desirable in our case: It is necessary to evaluate whether automatic classification of leadership traits from speech, arguably a challenging task, performs significantly better than chance. Thus, we evaluate significance of performance differences for binary classification in terms of weighted accuracy using a correlated proportions test [51]. It is based on the assumption that the accuracy difference between a classifier A and a baseline B with accuracies p_a and p_b is a normally distributed random variable with mean $p_a - p_b$ and variance $2p(1-p)/S$, where $p = (p_a + p_b)/2$ and S is the number of instances of the test set. We use a one-tailed test, i.e., the null hypothesis (H_0) is that $p_a \leq p_b$ or, informally, A is not better than the baseline B. To model comparison with chance level accuracy, p_b is set to 0.5. In Fig. 3, we show how large the accuracy improvement of A with respect to B must be on the test set on track level (117 instances) or chunk level (1,272 instances) to reject H_0 at either the 0.05, 0.01, or 0.001 level: The required accuracy improvement is given by the intersection of the vertical line corresponding to the baseline accuracy and the curve corresponding to the level of significance.

This test allows us to easily assess the significance of any difference in accuracy encountered throughout analysis; yet, results of this test should only be interpreted as a heuristic measure since the estimates of p_a and p_b on the test set are not independent [51]. Furthermore, it is not straightforward to measure significance of differences in *unweighted* accuracy. Caution must be exercised when applying the above-mentioned significance tests to results on chunk level as the required assumption of statistical independence among samples is not necessarily given; while in ASR one often assumes that results of a recognition algorithm on parts separated by speech pauses are independent of each other [52], it is not clear that this is also the case for recognition of speech traits that arguably

TABLE 6
Results on the Development Set by Regression/Classification Using Either Acoustic or Linguistic Features

	Tracks		Chunks		Chunk mean	
Dimension	CC	MLE	CC	MLE	CC	MLE
Non-malicious	.183	.151	.031	.152	.004	.154
Diplomatic	.097	.188	-.025	.197	-.031	.190
Achiever	.435	.225 ●	.232	.255 ●●	.253	.240 ○
Charismatic	.463	.217 ○	.239	.255	.305	.231
Teamplayer	.218	.181	.069	.189	.087	.184

(a) Regression, acoustic features

	Tracks		Chunks		Chunk mean	
Dimension	CC	MLE	CC	MLE	CC	MLE
Non-malicious	.041	.154	.015	.152 ○	-.040	.152 ○
Diplomatic	.164	.185 ○	.084	.195 ●●	.119	.187
Achiever	.234	.250 ●	.208	.261 ●●	.198	.246 ○
Charismatic	.298	.236	.204	.254 ●●	.246	.236
Teamplayer	.219	.180 ○	.126	.183 ●	.152	.182

(b) Regression, linguistic features

[%]	Tracks		Chunks		Maj.vote	
Dimension	UAR	Acc.	UAR	Acc.	UAR	Acc.
Non-malicious	55.7	57.6	54.8	58.0 ●●	58.0	61.6 ○
Diplomatic	57.2	57.6	54.6	54.6 ●	57.5	57.6
Achiever	62.4	61.6 ○	59.0	57.0 ●●	57.7	56.0
Charismatic	59.7	59.2	53.5	52.1	55.7	54.4
Teamplayer	54.4	52.8	57.2	55.7 ●	59.9	58.4

(c) Binary classification, acoustic features

	Tracks		Chunks		Maj.vote	
Dimension	UAR	Acc.	UAR	Acc.	UAR	Acc.
Non-malicious	48.3	49.6	49.7	54.9 ●	49.0	56.8
Diplomatic	52.9	52.8	48.5	48.6	51.2	52.0
Achiever	53.5	53.6	54.6	56.5 ●●	59.1	62.4 ○
Charismatic	53.5	52.0	54.5	57.1 ●●	51.9	55.2
Teamplayer	58.2	57.6	52.9	54.3 ○	51.1	52.8

(d) Binary classification, linguistic features

The significance of MLE w.r.t. dummy prediction; significance of accuracy (Acc.) w.r.t. chance (○: $p < 0.05$, ●: $p < 0.01$, ●●: $p < 0.001$).

evolve slowly over time. Note that we do not correct for repeated measurements, as it was suggested already in [53] to use significance not in the inferential meaning but as a sort of descriptive device—a more objective measure of differences worth being discussed.

4.3 Significance-Based Evaluation Measure for Regression

Given the distribution of the regression targets (mean rating $\overline{c}_i^{\dim}$) as shown in Fig. 1, it must be taken into account that the MLE itself is insufficient when comparing the results achieved for different dimensions as the range of possible target values for the test data varies considerably. Furthermore, given the cumulation of annotated targets around the mean, it can be assumed that a regression function will be biased toward predicting the mean across the training data. Thus, we decided to use a paired t-test for comparing the trained regression function to a "dummy" reference, that is, a constant function that always predicts the arithmetic mean of the dimension computed from training and development data. In that case, a one-tailed t-test is used, with the null hypothesis (H_0) assuming that the mean difference between the linear errors of the regression function and the dummy is greater or equal to zero. Thus, the error probability for rejecting H_0 is a judgment of whether the achieved MLE is significantly lower than the one resulting from "always predicting the mean."

5 Experimental Results

5.1 Acoustic and Linguistic Analysis: Degrees of Freedom and Performances

In order to design the system for automatic analysis of speeches that serves for the final evaluation on the test set, we first performed an extensive evaluation on the development set: The degrees of freedom comprise the kind of target variable (continuous or nominal), the features (acoustic or linguistic), and the unit of analysis (track level, chunk level, aggregation of chunk level results).

Results for acoustic features and regression on the mean rating are shown in Table 6a. It can be seen that on track level, a notable CC of 0.435 and 0.463 is achieved for the Achiever and Charismatic dimensions, respectively; this is on the order of magnitude of the best results obtained for interest detection from speech in the Interspeech 2010 Paralinguistic Challenge [37]. Yet, the CC for the other dimensions (Non-malicious, Diplomatic, and Teamplayer) is considerably lower; for Diplomatic, there is no significant correlation (0.097, $p > 0.05$). Regarding the MLE, it is again only for the Achiever and Charismatic dimensions that the regressor significantly ($p < 0.01, p < 0.05$) outperforms the dummy prediction. Comparing the MLE between dimensions, it is interesting that the MLE assumes the two of its absolute highest values (0.225, 0.217) for the Achiever and Charismatic dimensions; however, the results of the significance test suggest that this phenomenon can be entirely attributed to the high variability of the ratings on those dimensions as opposed to the low variability for the other dimensions (see also Fig. 1). In comparison, regression on chunk level seemingly delivers lower CC and higher MLE; still, for the Achiever dimension, the MLE is significantly better than the one of the dummy. Furthermore, it seems that the mean of chunk level results does not deliver better predictions than the track level regression in terms of CC and MLE, for all dimensions, which can probably be attributed to the generally unsatisfactory performance on chunk level.

Next, binary classification on acoustic features is evaluated in Table 6c. The results mirror the ones for regression to some extent: For instance, better-than-chance accuracy is achieved on the Achiever dimension on track level, and performance is lower (yet not significantly) on chunk level. Interestingly, for the Non-malicious dimension, results are now significantly above chance level (61.6 percent accuracy, $p < 0.05$) when performing a majority vote among chunk level classification results.

For linguistic features, we shortly summarize the results as follows: Regression (Table 6b) overall delivers lower CC than for acoustic features on track level; still, the MLE is

TABLE 7
Evaluation of Track Level Analysis on the Development and Test Set by Late Acoustic + Linguistic Fusion

[%]	LFS1			LFS2			LFS3		
	λ_1	UAR	Acc.	λ_2	UAR	Acc.	λ_3	UAR	Acc.
NON	0.3	**61.3**	66.4 •	1.0	55.7	57.6	0.5	59.0	64.8 •
DIP	0.6	**58.5**	58.4	1.0	57.2	57.6	0.1	56.7	56.8
ACH	0.5	**64.6**	62.4 ○	2.5	62.9	64.0 ○	1.7	63.8	64.0 ○
CHR	0.4	58.2	56.0	5.0	61.2	63.2 ○	1.3	**63.2**	64.0 ○
TPL	0.2	**64.0**	62.4 ○	2.0	57.6	56.8	1.0	62.2	61.6 ○

(a) on development set

LFS1		LFS2		LFS3	
UAR	Acc.	UAR	Acc.	UAR	Acc.
56.7	54.7	55.4	63.2	**59.6**	60.7
50.5	50.4	**56.3**	59.0	56.0	56.4
72.5	72.6 ••	61.8	62.4 ○	64.4	65.0 ○
67.3	66.7 •	52.7	51.3	60.3	59.0
62.5	62.4 ○	59.7	59.0	59.0	58.1

(b) on test set

LFS1: Chunk level acoustic, track level linguistic features. LFS2: Track level acoustic, chunk level linguistic features. LFS3: Chunk level acoustic, chunk level linguistic features. $\lambda_1, \lambda_2, \lambda_3$*: Linguistic weights according to (3) through (5) optimized on the development set. Significance of accuracy (Acc.) w.r.t. chance (*○*:* $p < 0.05$, •*:* $p < 0.01$, ••*:* $p < 0.001$*).*

significantly better than the dummy for three of five dimensions (DIPLOMATIC, ACHIEVER, TEAMPLAYER). On chunk level, the MLE even passes the significance test for all five dimensions; still, for the reasons mentioned in Section 4.2, it is disputable whether this cannot simply be attributed to artifacts of statistical dependence: Indeed, the track level MLE of averaged chunk level results remains insignificant for three of five dimensions. Concerning classification by linguistic features (Table 6d), the only dimension where performance significantly exceeds chance level is—again—ACHIEVER.

Overall, it is hard to derive a concise conclusion from the results on the development set as they are very mixed: In particular, it is not possible to rule out a certain choice of features (acoustic or linguistic) or unit of analysis (track or chunk level). If any, a noticeable tendency is that the recognition of achievers is promising, delivering better-than-chance accuracy or correlation in 8 of the 12 scenarios considered so far. Since we aim at a general strategy for combined acoustic and linguistic analysis that is largely independent of the type of (low-level) classifier and its parameters and because the optimal choice of unit of analysis remains unclear, we now move on to discussing late fusion of acoustic and linguistic features using all three strategies presented in Section 3.4 and their optimization on the development set.

5.2 Development of Late Fusion Strategies

In particular, we optimize the linguistic stream weight for binary classification on track level by LFS 1-3 on the unweighted accuracy on the development set. The range of possible parameters was determined by the considerations in Section 3.4: For LFS1, λ_1 was chosen from $\{\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}, 1\}$; since LFS2 is the "inverse" of LFS1, λ_2 for LFS2 was chosen from $\{1, \frac{10}{9}, \frac{10}{8}, \ldots, \frac{10}{2}, 10\}$; finally, the union of these parameter sets was the set of possible values for λ_3.

The most noticeable improvement over the acoustic and linguistic baselines is achieved when considering late fusion of chunk level decisions (LFS3): In case of the *charismatic* dimension, all parameter values improve the UAR over the baselines (54.5 and 53.5 percent UAR, respectively), with a maximum of 62.9 percent UAR (64.0 percent accuracy) obtained at $\lambda_3 = \frac{10}{8}$. Thus, in that case, the fused result is observed significantly above chance level accuracy ($p < 0.05$).

The overall performance of the late fusion binary classifiers on the development set is shown in Table 7a; all in all, it is very motivating: In short, for four of the five dimensions considered for evaluation, i.e., all except DIPLOMATIC, the line of significant accuracy (60.3 percent accuracy) is crossed by at least one of the fusion strategies. In terms of average UAR across the five dimensions, LFS2 (58.9 percent) falls slightly behind LFS1 (61.3 percent) and LFS3 (61.0 percent); still, neither of the corresponding accuracy differences is statistically significant, which is why we proceed to evaluate each LFS on the test set.

5.3 Evaluation on Test Set

For the final evaluation of our fusion strategies on the test set, we retrained the low-level classifiers on the union of the training and development sets and used the fusion weights determined on the development set for fusing the predictions on the test set. Our results are shown in Table 7b: It is striking that LFS1 performs especially well on the test set, boosting the accuracy to over 72 percent UAR and accuracy for the ACHIEVER dimension, also exhibiting remarkable performance for CHARISMATIC (67.3 percent UAR/66.7 percent accuracy), and above-chance accuracy for TEAMPLAYER (62.4 percent UAR/62.5 percent accuracy). On the other hand, LFS2 and LFS3 fall considerably behind LFS1 on the test set—for LFS2, even significantly in some cases—which deserves some further investigation. An important difference of LFS1 with respect to LFS2 and LFS3 is that the linguistic stream weight is smaller than one for LFS1, while it is greater than or equal to one for LFS2 and LFS3. Thus, an explanation could be that linguistic features perform worse on the test set: This hypothesis, however, can be rejected by comparing the performance of the majority vote among linguistic, chunk-level classifiers on the development as opposed to the test set. There, only statistically insignificant differences in the order of 1 percent UAR could be found. Thus, we hypothesize that the reason for the high performance of LFS1 is based on increased predictive ability of the acoustic features; indeed, we can provide evidence for this by the results of binary classification on purely acoustic features, which are shown in Table 8. The

TABLE 8
Binary Classification on the Test Set by Acoustic Features: Track and Chunk Level, and Track Level by Maj(ority) Vote

[%]	TRACKS		CHUNKS		MAJ.VOTE	
	UAR	Acc.	UAR	Acc.	UAR	Acc.
NON	56.1	64.1 •	56.1	55.9 •	**61.0**	58.1
DIP	54.1	57.3	53.8	54.3 ○	**56.5**	56.4
ACH	68.2	68.4 •	65.6	66.4 ••	**71.7**	71.8 ••
CHR	59.3	59.8	61.8	61.9 ••	**63.0**	62.4
TPL	55.5	55.6	62.0	62.1 ••	**65.2**	65.0

Significance of accuracy (Acc.) w.r.t. chance (○*:* $p < 0.05$, •*:* $p < 0.01$, ••*:* $p < 0.001$*).*

majority vote among chunk level acoustic classifiers is comparable in performance to LFS1—thus, on the test set, the benefit of adding linguistic information seems to be smaller.

The remarkable performance of acoustic features on the test set led us to another experiment, to investigate whether the test set is easier to classify by acoustic features or the benefit stems from the additional training instances of the development set. To this end, we evaluated the performance of the very same low-level acoustic classifiers that were used for classifying the development set on the test set. It turned out that their performance was clearly below the one of classifiers trained on training and development set (Table 8): For majority vote among chunks and on average across the five dimensions, the UAR was 60.1 percent, as opposed to 63.5 percent when training with both the training and development sets. For the ACHIEVER dimension, the difference in accuracy (61.5 percent versus 71.8 percent) is even significant with $p < 0.05$. This provides evidence that the acoustic features in development and test set are "more compatible" than in the training and test set.

5.4 Discussion and Outlook

In summary, we have demonstrated that fused acoustic and linguistic information delivers remarkable accuracy in recognizing different facets of leadership in a real-life audio archive. Particularly promising results have been accomplished for the ACHIEVER dimension: Here, the binary decision by late fusion achieved over 72 percent unweighted accuracy on the test set and was always significantly above chance level for both the development and test set and all late fusion strategies. Still, it is notable that the performance on the test set does not always increase by taking into account linguistic features—while this can be attributed to challenging conditions for ASR, it is somewhat surprising, as previous findings suggest that text classification based on ASR is robust even against high word error rates [45]. Finally, it is important to point out that since we deal with signal level speech analysis, not text classification, the performance of our linguistic features does not allow definite conclusions as to whether linguistics are an important factor in determination of leadership emergence. On a related note, we believe that the proposed YouTube corpus will be an interesting testbed for evaluation of adaptive, robust ASR technologies in future research as the speech is corrupted by essentially unknown noise and reverberation.

A strong focus has been laid in this study on fusion strategies that are independent of the architecture of the low-level classifier: We even proposed a high-level classification paradigm that allows asynchronous decisions to be fused to be able to take into account classifier decisions from different units of analysis (track or chunk level). In fact, it has been exactly a strategy for late fusion of track level linguistic and chunk level acoustic classifiers that prevailed on both, the development and the test set. Naturally, this still leaves room for improvement on the classifier level: In fact, it would be a natural extension to investigate other classifiers besides SVM, especially those that provide a meaningful confidence measure for late decision fusion. Finally, directions of future research can also be accounted for on the feature level: For instance, detection of nonlinguistic vocalizations (such as filled pauses) could be integrated into the BoW vectors or considered as a separate chunk or track level stream. The same holds for the voice activity curve as output by the LSTM-VAD, whose shape could be an interesting track level feature that indicates speaking style.

6 Conclusions

We have introduced the challenging task of automatic determination of leadership emergence in online speeches. Furthermore, we have proposed a system that allows robust automatic recognition of achievers, charismatic speakers, and teamplayers in full realism, that is, using automatic voice activity detection and speech recognition prior to feature extraction. By using the real-life YouTube corpus for evaluation, we have demonstrated that our approach generalizes over a large variety of acoustic conditions. Among the dimensions of leadership that were considered, highest accuracy on unseen test data (72.5 percent) is reached in recognizing achievers. This result is somehow expected since ACHIEVER is 1) among the dimensions that are best correlated with the acoustic features considered, and 2) the leadership trait with the highest agreement among professional human assessors, which naturally creates more reliable training and test labels for machine learning, but also indicates that this leadership trait is particularly evident in speech.

In future work, we will integrate strategies for unsupervised learning on unlabeled speech data collected from online sources in order to iteratively improve automatic speech recognition as well as acoustic-linguistic recognition of leadership traits.

Acknowledgments

The authors would like to thank Florian Eyben for helpful discussions.

References

[1] T.A. Judge and R.F. Piccolo, "Transformational and Transactional Leadership: A Meta-Analytic Test of Their Relative Validity," *J. Applied Psychology,* vol. 89, pp. 755-768, 2004.

[2] J. Kuoppala, A. Lamminpää, J. Liira, and H. Vainio, "Leadership, Job Well-Being, and Health Effects—A Systematic Review and a Meta-Analysis," *J. Occupational and Environmental Medicine,* vol. 50, no. 8, pp. 904-915, 2008.

[3] M. Van Vugt, R. Hogan, and R.B. Kaiser, "Leadership, Followership, and Evolution—Some Lessons from the Past," *Am. Psychologist,* vol. 63, pp. 182-196, 2008.

[4] R. Hogan and R. Kaiser, "What We Know about Leadership," *Rev. General Psychology,* vol. 9, pp. 901-910, 2005.

[5] R.R. McCrae and O.P. John, "An Introduction to the Five-Factor Model and Its Applications," *J. Personality,* vol. 60, pp. 175-215, 1992.

[6] K.R. Scherer, "Personality Markers in Speech," *Social Markers in Speech,* K.R. Scherer and H. Giles, eds., pp. 147-209, Cambridge Univ. Press, 1979.

[7] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *J. Artificial Intelligence Research,* vol. 30, pp. 457-500, 2007.

[8] K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling Vocal Interaction for Text-Independent Participant Characterization in Multi-Party Conversation," *Proc. Ninth SIGdial Workshop Discourse and Dialogue,* pp. 148-155, 2008.

[9] A. Rosenberg and J. Hirschberg, "Acoustic/Prosodic and Lexical Correlates of Charismatic Speech," *Proc. Interspeech,* pp. 513-516, 2005.

[10] S.W. Gregory and T.J. Gallagher, "Spectral Analysis of Candidates' Nonverbal Vocal Communication: Predicting U.S. Presidential Election Outcomes," *Social Psychology Quarterly,* vol. 65, pp. 298-308, 2002.

[11] C. Nass and K.M. Lee, "Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction," *J. Experimental Psychology: Applied,* vol. 7, pp. 171-181, 2001.

[12] S. Argamon, S. Dawle, M. Koppel, and J. Pennebaker, "Lexical Predictors of Personality Type," *Proc. Joint Ann. Meeting of the Interface and the Classification Soc. of North Am.,* 2005.

[13] J. Oberlander and S. Nowson, "Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text," *Proc. COLING/ACL Main Conf. Poster Sessions,* pp. 627-634, 2006.

[14] G. Mohammadi, M. Mortillaro, and A. Vinciarelli, "The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions," *Proc. Int'l Workshop. Social Signal Processing,* pp. 17-20, 2010.

[15] F. Metze, A. Black, and T. Polzehl, "A Review of Personality in Voice-Based Man Machine Interaction," *Proc. 14th Int'l Conf. Human-Computer Interaction: Interaction Techniques and Environments,* pp. 358-367, 2011.

[16] J.E. Bono and T.A. Judge, "Personality and Transformational and Transactional Leadership: A Meta-Analysis," *J. Applied Psychology,* vol. 89, pp. 901-910, 2004.

[17] J.W. Pennebaker and T.C. Lay, "Language Use and Personality during Crises: Analyses of Mayor Rudolph Giuliani's Press Conferences," *J. Research in Personality,* vol. 36, no. 3, pp. 271-282, 2002.

[18] R.J. House, P.J. Hanges, M. Javidan, P.W. Dorfman, and V. Gupta, *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies.* Sage Publications, 2004.

[19] M.V. Grachev and M.A. Bobina, "Russian Organizational Leadership—Lessons from the GLOBE Study," *Int'l J. Leadership Studies,* vol. 1, no. 2, pp. 67-79, 2006.

[20] J.A. Irving, "Educating Global Leaders: Exploring Intercultural Competence in Leadership Education," *J. Int'l Business and Cultural Studies,* vol. 3, no. 1, pp. 30-49, 2010.

[21] M.W. Dickson, D.N. Den Hartog, and J.K. Mitchelson, "Research on Leadership in a Cross-Cultural Context: Making Progress, and Raising New Questions," *The Leadership Quarterly,* vol. 14, pp. 729-768, 2003.

[22] A.M. Bertsch, "Validating GLOBE Scales: A Test in the U.S.A," *Proc. Cambridge Business and Economics Conf.,* p. 31, 2011.

[23] S. Joy and D.A. Kolb, "Are There Cultural Differences in Learning Style?" *Int'l J. Intercultural Relations,* vol. 33, no. 1, pp. 69-85, 2009.

[24] J. Mansour, R.J. House, P.W. Dorfman, P.J. Hanges, and M.S. de Luque, "Conceptualizing and Measuring Cultures and Their Consequences: A Comparative Review of GLOBE's and Hofstede's Approaches," *J. Int'l Business Studies,* vol. 37, pp. 897-914, 2006.

[25] M. Grimm and K. Kroschel, "Evaluation of Natural Emotions Using Self Assessment Manikins," *Proc. IEEE Automatic Speech Recognition and Understanding Conf.,* pp. 381-385, 2005.

[26] J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey, *Graphical Methods for Data Analysis.* Wadsworth & Brooks/Cole, 1983.

[27] J. Kruskal and M. Wish, *Multidimensional Scaling.* Sage Univ., 1978.

[28] G. Yukl, *Leadership in Organizations,* sixth ed. Pearson-Prentice Hall, 2006.

[29] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology.* Sage Publications, 2004.

[30] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M=Syntax+Prosody: A Syntactic-Prosodic Labelling Scheme for Large Spontaneous Speech Databases," *Speech Comm.,* vol. 25, no. 4, pp. 193-222, Sept. 1998.

[31] A. Graves, "Supervised Sequence Labelling with Recurrent Neural Networks," PhD dissertation, Technische Universität München, 2008.

[32] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," *Proc. CHiME Workshop,* pp. 24-29, 2011.

[33] D. Pearce and H.-G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," *Proc. Automatic Speech Recognition,* pp. 181-188, 2000.

[34] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile—The Munich Versatile and Fast Open-Source Audio Feature Extractor," *Proc. ACM Multimedia,* pp. 1459-1462, Oct. 2010.

[35] S. Steidl, B. Schuller, A. Batliner, and D. Seppi, "The Hinterland of Emotions: Facing the Open-Microphone Challenge," *Proc. Third Int'l Conf. Affective Computing and Intelligent Interaction and Workshops,* pp. 690-697, Sept. 2009.

[36] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learned from the First Challenge," *Speech Comm.,* vol. 53, nos. 9/10, pp. 1062-1087, 2011.

[37] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," *Proc. 13th Int'l Speech Comm. Assoc.,* pp. 2794-2797, Sept. 2010.

[38] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," *Proc. Int'l Speech Comm. Assoc.,* pp. 3201-3204, 2011.

[39] B. Schuller, S. Steidl, A. Batliner, E. Nth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," *Proc. 13th Ann. Conf. Int'l Speech Comm. Assoc.,* Sept. 2012.

[40] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2012—The Continuous Audio/Visual Emotion Challenge," *Proc. Second Int'l Audio/Visual Emotion Challenge and Workshop, Grand Challenge and Satellite of ACM ICMI '12,* Oct. 2012.

[41] B. Weiss and F. Burkhardt, "Voice Attributes Affecting Likability Perception," *Proc. Int'l Speech Comm. Assoc.,* pp. 2014-2017, 2010.

[42] S.J. Young, G. Evermann, M.J.F. Gales, D. Kershaw, G. Moore, J.J. Odell, D.G. Ollason, D. Povey, V. Valtchev, and P.C. Woodland, *The HTK Book,* version 3.4, Cambridge Univ. Eng. Dept., 2006.

[43] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (Second Release),* Dept. of Psychology, Ohio State Univ. (distributor), www.buckeyecorpus.osu.edu, 2007.

[44] F. Weninger, B. Schuller, M. Wöllmer, and G. Rigoll, "Localization of Non-Linguistic Events in Spontaneous Speech by Non-Negative Matrix Factorization and Long Short-Term Memory," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 5840-5843, 2011.

[45] S. Agarwal, S. Godbole, D. Punjani, and S. Roy, "How Much Noise Is Too Much: A Study in Automatic Text Classification," *Proc. IEEE Seventh Int'l Conf. Data Mining,* pp. 3-12, 2007.

[46] J.C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," *Advances in Kernel Methods: Support Vector Learning,* pp. 185-208, MIT Press, 1999.

[47] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proc. 10th European Conf. Machine Learning,* pp. 137-142, 1998.

[48] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 832-844, Aug. 1998.

[49] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter,* vol. 11, no. 1, pp. 10-18, 2009.

[50] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques,* second ed. Morgan Kaufman, 2005.

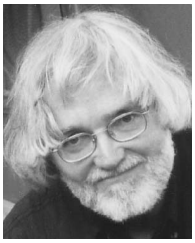
[51] T.G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation,* vol. 10, pp. 1895-1923, 1998.

[52] L. Gillick and S.J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 532-535, 1989.

[53] H. Eysenck, "The Concept of Statistical Significance and the Controversy about One-Tailed Tests," *Psychological Rev.,* vol. 67, pp. 269-271, 1960.

Felix Weninger received the master's degree in computer science from the Technische Universität München (TUM), one of the first three German Excellence Universities, in 2009. Currently, he is working toward the PhD degree as a researcher in the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication. He has authored or coauthored more than 30 publications in peer-reviewed books, journals, and conference proceedings. His research focuses on robust techniques for real-life speech and audio recognition tasks, especially the retrieval of paralinguistic information. He is a member of the IEEE.

Jarek Krajewski received the diploma in 2004 and the doctoral degree for his study on acoustic sleepiness detection in 2008, both in psychology and signal processing from the University of Wuppertal and RWTH Aachen. He has been an assistant professor in experimental industrial psychology since 2009 and is vice director of the Center of Interdisciplinary Speech Science at the University of Wuppertal. He has (co)authored more than 50 publications in peer reviewed books, journals, and conference proceedings in the field of sleepiness detection and signal processing. He is a member of the IEEE, ISCA, Human Factors and Ergonomics Society, German Society of Psychology (Section Industrial Psychology, Section Traffic Psychology).



Anton Batliner received the MA degree in Scandinavian languages and the Dr phil degree in phonetics in 1978, both from LMU Munich. He has been a member of the research staff of the Institute for Pattern Recognition at FAU since 1997. He is coeditor of one book and author/coauthor of more than 200 technical articles, with a current H-index of 29 and more than 3,000 citations. His research interests include the modeling and automatic recognition of emotional user states, all aspects of prosody and paralinguistics in speech processing, unimodel and multimodal focus of attention, pronunciation assessment, and spontaneous speech phenomena such as disfluencies, irregular phonation, etc.

Björn Schuller received the diploma in 1999 and the doctoral degree for his study on automatic speech and emotion recognition in 2006, both in electrical engineering and information technology from the Technische Universität München (TUM). He has been tenured as a senior researcher and lecturer in pattern recognition and speech processing heading the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication since 2006. His research interests include advancing audiovisual processing and affective computing. He has (co)authored three books and more than 270 publications in peer reviewed books (21), journals (38), and conference proceedings in the field, leading to more than 2,700 citations—his current H-index equals 27. He is a member of the IEEE, the IEEE Computer Society, the ACM, the HUMAINE Association, and the ISCA.