The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity and native language

Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. "The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity and native language." In *Interspeech 2016, 8-12 Sep 2016, San Francisco, CA, USA*, edited by Nelson Morgan, 2001–5. ISCA. https://doi.org/10.21437/interspeech.2016-129.

INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language

Björn Schuller[1,2], Stefan Steidl[3], Anton Batliner[2,3], Julia Hirschberg[4], Judee K. Burgoon[5], Alice Baird[4], Aaron Elkins[5], Yue Zhang[1], Eduardo Coutinho[1,6], Keelan Evanini[7]

[1]Department of Computing, Imperial College London, UK
[2]Chair of Complex & Intelligent Systems, University of Passau, Germany
[3]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[4]Department of Computer Science, Columbia University, New York, USA
[5]Center for the Management of Information, University of Arizona, Tucson, USA
[6]Department of Music, University of Liverpool, UK
[7]Educational Testing Service, USA
schuller@IEEE.org

Abstract

The INTERSPEECH 2016 Computational Paralinguistics Challenge addresses three different problems for the first time in research competition under well-defined conditions: classification of deceptive vs. non-deceptive speech, the estimation of the degree of sincerity, and the identification of the native language out of eleven L1 classes of English L2 speakers. In this paper, we describe these sub-challenges, their conditions, the baseline feature extraction and classifiers, and the resulting baselines, as provided to the participants.
Index Terms: Computational Paralinguistics, Challenge, Deception, Sincerity, Native Language Identification

1. Introduction

In this INTERSPEECH 2016 COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) – the eighth since 2009 [1], we address three new problems within the field of Computational Paralinguistics [2] in a challenge setting as will be outlined one by one in the following.

1.1. Deception

In the *Deception **(D)** Sub-Challenge*, deceptive speech has to be identified. Deception has been associated with manifestations of both fear and elation [3] and can in general be identified in verbal [4, 5, 6, 7] and non-verbal behaviour [8]. Furthermore, these cues benefit human judgement also in the presence of other information such as visual cues [9]. A few studies have included audio analysis [10]: Ekman et al. [11] found a significant increase in pitch for deceptive speech over truthful speech. Voice stress analysis procedures attempt to rely upon low level indicators of stress as indirect indicators of deception, and commercial systems promise to distinguish truth from lie with little independent evidence of success. [12] give ample evidence of the problems connected with 'Lie Detection from Voice'.

There has been little work on the automatic identification of deceptive speech from such acoustic, prosodic, and lexical cues. This is partially owed to the severity of the task [13]. Yet, distinguishing deceptive from non-deceptive speech automatically is of considerable practical interest, especially to law enforcement and other government agencies – to identify potential deception at border crossings and in military scenarios in the field and elsewhere, and to evaluate reports from informants at embassies and consulates throughout the world.

1.2. Sincerity

In the *Sincerity **(S)** Sub-Challenge*, the degree of perceived sincerity of speakers can be investigated empirically for the first time. The data collection for this task was motivated by the 2015 art installation 'A Sincere Apology' hosted in New York City, taking the form of a stylised and choreographed annotation experience initiated by Alice Baird. This experience inspired and backed up the hypothesis that sincerity will often be perceived less from the linguistic content of an utterance than from the way content was expressed [14]. However, audible sincerity appears to be subjective; the perception of the listener appears to be based in part on their native dialect. Moreover, it has been proposed in the literature that depth (lower pitch) in the voice indicates a sincere assertion [15]. Further, prosodic cues seem to be highly language dependent [16]. The body of literature on automatic sincerity evaluation is sparse and mostly deals with binary classification in the context of sarcasm recognition [17].

The data were collected in order to assess these hypotheses and to discover other spoken cues to sincerity.

1.3. Native Language

Different from the *Degree of Nativeness **(DN)** Sub-Challenge* from last year's ComParE [18], where the proficiency of learners had to be assessed on a rating scale, in the *Native Language **(N)** Sub-Challenge*, the native language (L1) of non-native English L2 speakers from eleven L1 backgrounds has to be recognised. This task is similar in many ways to the tasks of language identification (in which a system distinguishes among a set of different spoken languages) and dialect or accent identification (cf. e. g., [19, 20, 21]) (in which a system distinguishes among a set of regional native-speaker dialects of a single spoken language, such as British vs. North American English), but has been much less widely studied. A few corpora with samples of non-native English speech from a variety of language backgrounds do exist and have been used for isolated native

http://dx.doi.org/10.21437/Interspeech.2016-129

language identification studies; these include the CSLU Foreign Accented English corpus [22][1], the CU-Accent corpus [23], and a corpus collected as part of the SUNSTAR project [24]. In contrast to these, the corpus that will be used for the Native Language Sub-Challenge – the ETS CORPUS OF NON-NATIVE SPOKEN ENGLISH – is larger and contains more speech for each native language. Thus, this corpus will be useful for testing algorithms that benefit from larger amounts of training data. In addition, the audio files in the corpus are sampled at 16 kHz, which represents an improvement over the next-largest previously available corpus (the CSLU Foreign Accented English corpus), which consists of telephone speech sampled at 8 kHz. It is envisioned that systems that perform native language identification will be increasingly in demand as spoken language applications become more frequent in global business and commerce; for example, a native language identification capability could enable a speech-based application that serves non-native speakers to use L1-specific ASR models, leading to better recognition accuracy, and to adapt to the user's profile to enable a more context-aware spoken dialogue.

For future studies, we can even relate the first task on deceptive speech with the recognition of L1, since it has been shown that a foreign accent with lower intelligibility can cause non-native speakers to sound less credible [25], and, in specific settings, perhaps even deceptive. Additionally, it is likely that phonetic and other traits of deceptive speech are not universal but, up to a certain extent, language-specific as well.

1.4. Overview

For all tasks, a target value/class has to be predicted for each speech file. Contributors can employ their own features and machine learning algorithms, however, a standard feature set is provided that may be used. Participants will have to use predefined training/development/test splits for each sub-challenge. They may report development results obtained from the training set (preferably with the supplied evaluation setups), but have only a limited number of five trials to upload their results on the test sets for the Sub-Challenges, whose labels are unknown to them.

Each participation must be accompanied by a paper presenting the results, which undergoes peer-review and has to be accepted for the conference in order to participate in the Challenge. The organisers preserve the right to re-evaluate the findings, but will not participate themselves in the Challenge.

As evaluation measures, for the **S** Sub-Challenge, we use Spearman's Correlation Coefficient (ρ) as the more 'conservative' and robust alternative to Pearson's correlation coefficient. For the **D** and **N** Sub-Challenge tasks, we employ Unweighted Average Recall (UAR) as used since the first Challenge held in 2009 [1], especially because it is more adequate for (more or less unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy).

In the next section (2) we describe the challenge corpora. Section 3 describes the baselines experiments and metrics for each sub-challenge. Then, we provide baseline results in Section 4 before concluding in Section 5.

[1] https://catalog.ldc.upenn.edu/LDC2007S08

Table 1: *Deceptive Speech Database (DSD): Number of instances per class in the train/devel/test splits used for the Challenge; D: deceptive; ND: non-deceptive.*

#	Train	Devel	Test	Σ
D	182	129	121	432
ND	390	357	376	1123
Σ	572	486	497	1555

2. Challenge Corpora

2.1. Deception (D)

In this sub-challenge, we introduce the DECEPTIVE SPEECH DATABASE (DSD) created at the University of Arizona. The DSD consists of the audio recordings obtained in an empirical study where university student participants were randomly assigned to two experimental conditions. In one condition, participants were asked to take the role of impostors with false identities and to retrieve ("steal") an exam key from a computer at the department's main office. In the other, participants played the role of innocent characters, maintaining their own identity and retrieving a leaflet from the same office. In the following phase, structured interviews were conducted by an Embodied Conversational Agent (ECA) with each participant, which provides a high degree of consistency in the interviews between subjects (something often difficult with human interviewers). Participants who stole the key (guilty/deceptive condition) were asked to lie about the theft during the interview phase. Participants in the innocent/truthful conditions were asked to tell the truth about their activities. The interviews consisted of a fixed set of short-answers and open-ended questions divided into two phases: i) ten background questions that served as a truthful baseline; and ii) specific questions about the theft (including direct accusations and questions testing recognition of the stolen items).

The full set of recordings includes approximately 162 minutes of speech from 72 speakers, leading to a total of 1 556 instances (see Table 1). All audio files are monophonic and encoded in 16-bit signed integer PCM WAV format, sampled at 16 kHz, and normalised to -3 dB (using SoX).

The golden standard for the challenge is the truthfulness of each response, as defined in the experimental scenario. In the first part of the interview, both guilty and innocent participants replied truthfully to the background questions (**ND** condition). In the second part, participants assigned to the guilty condition were expected to lie to all questions (**D** condition), whereas innocent subjects were asked to reply truthfully (**ND** condition). Some participants that failed to reply in accordance to the experimental condition were removed from the dataset.

2.2. Sincerity (S)

For this sub-challenge, the SINCERITY SPEECH CORPUS (SSC) is provided by Columbia University. This database was created in the context of the above mentioned art exhibition focusing on the communication of sincerity – particularly in the context of an apology. A number of individuals were asked to read six different sentences, with each sentence read in four different prosodic styles: monotonic, non-monotonic, slow and fast. The content of each sentence is a form of apology: 1) "Sorry." 2) "I am sorry for everything I have done to you." 3) "I cannot tell you how sorry I am for everything I did." 4) "Please allow me to apologise for everything I did to you, I was inappro-

Table 2: ETS CORPUS OF NON-NATIVE SPOKEN ENGLISH*: Number of instances per class in the train/devel/test split used for the Challenge.*

#	Train	Devel	Test	Σ
ARA	300	86	80	466
CHI	300	84	74	458
FRE	300	80	78	458
GER	300	85	75	460
HIN	300	83	82	465
ITA	300	94	68	462
JAP	300	85	75	460
KOR	300	90	80	470
SPA	300	100	77	477
TEL	300	83	88	471
TUR	300	95	90	485
Σ	3 300	965	867	5 132

priate and lacked respect." 5) "It was never my intention to offend you, for this I am very sorry." 6) "I am sorry but I am going to have to decline your generous offer. Thank you for considering me." The full set of recordings includes approximately 72 minutes of speech by 32 speakers (15 m, 17 f, age between 20 and 65) and a total of 911 instances. All audio files are stereophonic and encoded in 16-bit signed integer PCM WAV format, sampled at 16 kHz, and normalised to -3 dB (using SoX).

Each instance was rated in terms of perceived sincerity using an ordinal rating scale ranging from 0 (not sincere at all) to 4 (extremely sincere) by at least 13 annotators (up to a maximum of 19). The ratings were standardised to zero mean and unit standard deviation on a per subject basis in order to eliminate individual biases in the use of the rating scale. The golden standard consists of the average sincerity standardised ratings across all annotators. The dataset was divided into speaker disjoint, stratified partitions, containing 22 subjects for the training set (655 instances) and 10 subjects for the test set (256 instances).

2.3. Native Language (N)

Educational Testing Service (ETS) provides the ETS CORPUS OF NON-NATIVE SPOKEN ENGLISH for this sub-challenge.[2] This corpus includes more than 64 hours of speech from 5,132 non-native speakers of English, with eleven different L1 backgrounds (Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR)). Each language is represented by recordings ranging from 458 to 485 different speakers representing a range of English speaking proficiencies. Each speech recording is 45 seconds long and was obtained in the context of the TOEFL iBT® assessment, which is designed to measure a non-native speaker's ability to use and understand English at the university level.

In this sub-challenge, the task consists of determining the speakers' native language from these recordings. The original audio files (encoded in Speex format), were converted to monophonic 16-bit signed integer PCM WAV format, sampled at 16 kHz, and normalised to -3dB (using SoX). The dataset was divided into stratified partitions: 3,300 instances (64%, approximately 41.3 hours) were selected as training data, 965 instances (19%, approximately 12.1 hours) for the development set, and 867 responses (17%, approximately 10.8 hours) will be used as test data (see also Table 2).

[2]The corpus is freely available for non-commercial research purposes. Please contact `kevanini@ets.org` for licensing information.

Table 3: *Results for the three sub-challenges. The* **official baselines** *for test which correspond to the best results obtained for development are highlighted (bold and greyscale). C: Complexity parameter of SVM/SVR. ρ: Spearman's correlation coefficient.* ***UAR****: Unweighted Average Recall.* **D***: Deception;* **S***: Sincerity;* **N***: Native Language.*

	D (UAR [%])		**S** (ρ)		**N (UAR** [%])	
C	Devel	Test	Devel	Test	Devel	Test
10^{-5}	57.6	62.3	.250	.533	31.4	30.9
10^{-4}	61.9	**68.3**	.474	**.602**	42.8	44.1
10^{-3}	58.9	67.4	.431	.524	44.9	47.7
10^{-2}	58.8	67.3	.427	.509	45.1	**47.5**
10^{-1}	58.8	67.3	.427	.509	45.1	47.5
1	58.8	67.3	.427	.509	45.1	47.5

3. Experiments

3.1. ComParE Acoustic Feature Set

The official baseline feature set is the same as has been used in the three previous editions of the INTERSPEECH ComParE challenges [26, 27, 18]. The COMPARE feature set contains 6 373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours. The configuration file is the IS13_ComParE.conf, which is included in the 2.1 public release of openSMILE [28, 29]. A full description of the feature set can be found in [30].

3.2. Basics for the Challenge Baselines

The primary evaluation measure for the **D** and **N** sub-challenges (both classification tasks) is Unweighted Average Recall (UAR). The motivation to consider *unweighted* rather than weighted average recall ('conventional' accuracy) is that it is also meaningful for highly unbalanced distributions of instances among classes (as is the case for the **D** sub-challenge). For the **S** sub-challenge, the official metric is Spearman's rank correlation coefficient ρ.

For the sake of transparency and reproducibility of the baseline computation, we use open-source implementations from the data mining algorithms (WEKA 3, revision 3.7.13; [31]); in line with previous years, the machine learning paradigm chosen is Support Vector Machines (SVM). In particular, we use WEKA's SVM implementation with linear kernels for the classification tasks, and Support Vector Regression (SVR; also with linear kernels) with epsilon-insensitive loss (which are known to be robust against overfitting; a fixed ϵ of 1.0 is used). In all tasks the Sequential Minimal Optimisation (SMO; [32]) as implemented in WEKA was used as training algorithm.

Features were scaled to zero mean and unit standard deviation (option `-N 1` for Weka's SMO/SMOreg), using the parameters from the training set (when multiple folds where used for development, the parameters were calculated on the training set of each fold). For all tasks, the complexity parameter C was optimised during the development phase.

Each sub-challenge package includes scripts that allows participants to reproduce the baselines and perform the testing in a reproducible and automatic way (including pre-processing,

Table 4: ETS CORPUS OF NON-NATIVE SPOKEN ENGLISH*: Confusion matrix in percent for the devolpment set, $C = 10^{-2}$, rounded to integer.*

		Hypothesis										
	[%]	GER	FRE	ITA	SPA	ARA	TUR	HIN	TEL	JAP	KOR	CHI
Reference	GER	65	6	8	6	6	0	1	1	1	2	4
	FRE	10	36	5	14	14	8	0	0	4	1	9
	ITA	6	10	49	11	6	4	6	1	0	4	2
	SPA	6	15	4	32	6	5	2	1	9	9	11
	ARA	8	6	6	8	34	8	6	7	7	7	3
	TUR	5	5	6	5	6	48	2	0	7	8	5
	HIN	0	1	2	2	5	1	57	25	2	2	1
	TEL	2	2	2	2	2	2	29	52	2	2	0
	JAP	2	5	1	12	5	1	2	1	42	13	15
	KOR	2	1	3	6	4	6	2	3	16	36	21
	CHI	5	6	2	7	5	1	6	5	6	12	45

model training, model evaluation, and scoring by the competition and further measures). In what follows, we will briefly summarise the baselines for each sub-challenge; development and test results are also summarised in Table 3.

4. Baselines

4.1. Deception

For this sub-challenge, a train-development-test schema was used to establish the baseline (see Table 1). The features in all sets were standardised to the mean and standard deviation of the training set. The optimal complexity determined in the development phase was $1.0E-4$, which resulted in a UAR of 61.1% (chance level: 50%). The test set performance was obtained from a model trained on the concatenated train and development sets, using the optimal complexity value determined in the development phase.

The baseline for this task is UAR $= 68.3\%$; see Table 3.

4.2. Sincerity

For this sub-challenge, given the relatively small size of the database, we used a leave-one-speaker-out cross-validation (LOSO-CV) schema during development. In each fold, features were standardised to the mean and standard deviation of the respective training set. The optimal complexity determined in the development phase was 10^{-4}, which resulted in $\rho = .474$. The test set performance was obtained from a model trained on the full development set, using the the optimal complexity value determined in the development phase.

The baseline for this task is $\rho = .602$; see Table 3.

4.3. Native Language

Similarly to the Deception task, we use a train-development-test schema for this sub-challenge (see Table 2 for the details on the folds). The features in all sets were standardised to the mean and standard deviation of the training set. The optimal complexity determined in the development phase was 10^{-2}, which resulted in a UAR of 45.1% (with a chance level of 9.1%). The test set performance was obtained from a model trained on the concatenated train and development sets, and using the optimal complexity value determined in the development phase.

The baseline for this task is UAR $= 47.5\%$; see Table 3. Table 4 displays the confusion matrix in a sort of 'geographical' order from west to east. We can see some patterns of higher confusion which might inspire future work. Note, however, that the number of cases per cell – apart from the diagonal – is rather low so we have to expect some random effects as well.

5. Conclusion

The two new tasks in the **D** and **S** sub-challenges represent two sides of one 'coin': to deceive or not to deceive, to be honest or not to be honest – arguably one of the most relevant aspects in communication, both in human-human and human-machine communication.

The **N** sub-challenge takes up another L1-L2-topic after last years' task dealing with the degree of nativeness: to be able to recognise a learner's L1 just by looking at his L2 – again, a highly relevant information in human-human communication not yet broadly considered in human-machine communication, yet often being of high social and contextual relevance.

All three tasks are highly relevant for future spoken dialogue and more general human-machine communication systems, and in many other applications such as in those related to security and forensics.

Due to the novelty and complexity of the tasks, the results reported for the baselines are not especially high. Yet, feature sets and learning procedures are standard - competitive but not optimised and kept generic for all tasks by intention to provide transparent and easily re-doable baselines.

We expect participants to obtain considerably better performance measures by employing novel (combinations of) procedures and features including such tailored to the particular tasks.

Beyond these tasks and past tasks featured in this challenge series, there remains a broad variety of further information that is conveyed in the acoustics of speech and the spoken words themselves that have not been dealt with either at all or in a well-defined competition framework. Many of these bear, however, great application potential, and remain to be investigated more closely.

6. Acknowledgements

The research leading to these results has received funding from the European Union's Framework Programme for Research and Innovation HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA) and the European Union's Seventh Framework Programme under the ERC Starting Grant No. 338164 (iHEARu). The authors would further like to thank the sponsors of the Challenge: audEERING GmbH, the Association for the Advancement of Affective Computing (AAAC), and Educational Testing Service. The responsibility lies with the authors.

7. References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first Challenge," *Speech Communication, Special Issue on Sensing Emotion and Affect – Facing Realism in Speech Processing*, vol. 53, pp. 1062–1087, 2011.

[2] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[3] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. WW Norton & Company, 2009.

[4] D. P. Twitchell, J. F. Nunamaker Jr, and J. K. Burgoon, "Using speech act profiling for deception detection," in *Intelligence and Security Informatics*. Springer, 2004, pp. 403–410.

[5] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Singapore, Singapore: Association for Computational Linguistics, 2009, pp. 309–312.

[6] J. Arciuli, D. Mallard, and G. Villar, ""Um, I can tell you're lying": Linguistic markers of deception versus truth-telling in speech," *Applied Psycholinguistics*, vol. 31, no. 03, pp. 397–411, 2010.

[7] J. C. Bachenko and M. J. Schonwetter, "Method and system for the automatic recognition of deceptive language," Dec. 14, 2010, US Patent 7,853,445.

[8] A. Vrij, P. A. Granhag, and S. Porter, "Pitfalls and opportunities in nonverbal and verbal lie detection," *Psychological Science in the Public Interest*, vol. 11, no. 3, pp. 89–121, 2010.

[9] A. S. Manstead, H. L. Wagner, and C. J. MacDonald, "Face, body, and speech as channels of communication in the detection of deception," *Basic and Applied Social Psychology*, vol. 5, no. 4, pp. 317–332, 1984.

[10] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *Proceedings of ICASSP*, vol. 1. Toulouse, France: IEEE, 2006.

[11] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Invited article: Face, voice, and body in detecting deceit," *Journal of nonverbal behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[12] J. Kreiman and D. Sidtis, *Foundations of Voice Studies – An Interdisciplinary Approach to Voice Production and Perception*. Wiley & Sons, 2011.

[13] J. D. Harnsberger, H. Hollien, C. A. Martin, and K. A. Hollien, "Stress and deception in speech: evaluating layered voice analysis," *Journal of forensic sciences*, vol. 54, no. 3, pp. 642–650, 2009.

[14] J. Eriksson, "Self-expression, Expressiveness, and Sincerity," *Acta Analytica*, vol. 25, no. 1, pp. 71–79, 2010.

[15] E. S. Hinchman, "Assertion, Sincerity, and Knowledge," *Noûs*, vol. 47, no. 4, pp. 613–646, 2013.

[16] H. S. Cheang and M. D. Pell, "Recognizing sarcasm without language: A cross-linguistic study of english and cantonese," *Pragmatics & Cognition*, vol. 19, no. 2, pp. 203–223, 2011.

[17] J. Tepperman, D. R. Traum, and S. Narayanan, ""Yeah right": sarcasm recognition for spoken dialogue systems." in *Proceedings of INTERSPEECH*, 2006.

[18] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition," in *Proceedings of INTERSPEECH 2015*, Dresden, Germany, 2015, pp. 478–482.

[19] L. M. Arslan and J. H. Hansen, "Language accent classification in American English," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[20] L. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *Proceedings of ICASSP*, vol. 1. Phoenix, Arizona: IEEE, 1999, pp. 221–224.

[21] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Columbia University, 2011.

[22] G. Choueiter, G. Zweig, and P. Nguyen, "An empirical study of automatic accent classification," in *Proceedings of ICASSP*, Las Vegas, Nevada, 2008, pp. 4265–4268.

[23] J. H. Hansen, S. S. Gray, and W. Kim, "Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification," *Speech Communication*, vol. 52, pp. 777–789, 2010.

[24] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *Fourth International Conference on Spoken Language Processing (ICSLP)*, vol. 3, 1996, pp. 1784–1787.

[25] S. Lev-Ari and B. Keysar, "Why don't we believe non-native speakers? The influence of accent on credibility," *Journal of Experimental Social Psychology*, vol. 46, pp. 1093–1096, 2010.

[26] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proceedings of INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[27] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & physical load," in *Proceedings of INTERSPEECH*, Singapore, 2014, pp. 427–431.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of ACM Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[29] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of ACM MM*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[30] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Emotion Science*, vol. 4, no. 292, pp. 1–12, May 2013.

[31] 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.

[32] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999, pp. 61–74.