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

Prediction of Deception and Sincerity from Speech using Automatic Phone Recognition-based Features

Robert Herms

Chair Media Informatics, Technische Universität Chemnitz, Germany
robert.herms@cs.tu-chemnitz.de

Abstract

As part of the Interspeech 2016 COMPARE challenge, the two different sub-challenges Deception and Sincerity are addressed. The former refers to the identification of deceptive speech whereas the degree of perceived sincerity of speakers has to be estimated in the latter. In this paper, we investigate the potential of automatic phone recognition-based features for these use case scenarios. The speech transcriptions were used to process the appearing tokens (phoneme, silent pause, filled pause) and the corresponding durations. We designed a high-level feature set including the four groups: vowels, phones, pseudo syllables, and pauses. Additionally, we selected suitable predefined acoustic feature sets and fused them with our introduced features showing a positive effect on the prediction. Moreover, the performance is further boosted by refining these fused features using the ReliefF feature selection method. Experiments show that the final systems outperform the baseline results of both sub-challenges.

Index Terms: computational paralinguistics, deception, sincerity, phoneme and pause duration, high-level features, challenge

1. Introduction

Speech can often be very complex especially when the content uttered by a speaker is not meant literally, e.g., in the case of irony. In other circumstances, the intended literal meaning does not reflect the beliefs and feelings of a speaker. The intended goal of deceptive speech is to conceal the meanings and attitudes using a literal message. In contrast, sincere speech can be described as the speaker's true beliefs, emotions or attitudes that match the uttered content. When the true feelings diverge from what is literally said it can be described as insincere or in certain cases as a white lie. [1]

The automatic recognition of deception from speech is of practical interest, especially in the field of law enforcement and other government agencies. Questioning and reporting could consequently uncover cases of fraud. Besides approaches to distinguish between truth and lies expressed in written text (e.g, [2] and [3]), some previous works investigated the effects of deception on speech. Ekman et al. [4] found a significant increase of pitch measures in deceptive speech. Similar results are reported in the work [5], with a higher pitch when lying than when telling the truth. The work of [6] shows that Teager energy-related features and formant variations indicate the possibility of discriminating between truthful and deceptive utterances. The automatic detection of deceptive speech using the combination of acoustic, prosodic, and lexical features is presented in [7].

The detection of sincerity from speech can be advantageous in the field of human-computer interaction. For instance, the intention of a user could be better considered by a spoken dialogue system. Basically, to speak sincerely means that the expressed utterance reflects a state of mind that one has (see [8]). The work of [9] studied that sincerity can be linked to benevolence and that a slow speaking rate sounds cold. Speech with a varying and lower pitch was evaluated as more benevolent. In connection with sincerity, some research has been done for sarcasm detection and how humans recognize and understand sarcastic speech (e.g., [10], [11], and [12]). Cheang & Pell [13] found overall reductions in mean pitch, decreases in pitch variation, and changes in harmonics-to-noise ratio to be indicative for sarcastic speech. Moreover, the results of [11] demonstrate the importance of spectral and contextual features.

Our approach in this work is focused on the potential of automatic phone recognition-based features for the prediction of deception and sincerity from speech. A motivation for this kind of features is the hypothesis that there are detectable changes of speech disfluency respectively speaking rate and rhythms when the degree of deception as well as sincerity of a speaker changes. Previous works demonstrated the importance of high-level features based on automatic speech recognition (ASR) systems for other use case scenarios (e.g., [14], [15], and [16]). We derive units of transcripts obtained from a speech recognizer including phonemes, vowels, pseudo syllables, silent pauses, and filled pauses with the corresponding durations. Furthermore, based on these units we extract 29 static features associated with durations, speaking rates, and ratios. As previous studies showed further indicators for the both tasks – deception and sincerity – we fuse our features with carefully selected acoustic feature sets. Additionally, we demonstrate how to further boost the performance of the predictions by refining these fused features using feature selection.

This Paper is organized as follows: In the next section we describe the corpora used for the Deception and Sincerity sub-challenges. In Section 3 we present our set of phone recognition-based features and state the procedure of feature extraction. Experiments are reported in section 4 including the assessment of diverse strategies and the results on the independent test sets. Finally, we conclude this paper in Section 5 and give some future directions.

2. Databases

For the Deception sub-challenge, the DECEPTIVE SPEECH DATABASE (DSD) was introduced which was created at the University of Arizona. The database includes recordings of approximately 162 minutes of speech from 72 speakers collected in a study where student participants were randomly assigned to two conditions: 1) Participants take the role of impostors with false identities and were asked to retrieve (steal) an exam key from the department's main office; 2) Participants with the role

http://dx.doi.org/10.21437/Interspeech.2016-971

of innocent characters and their own identity, retrieving a leaflet from the same office. Next, structured interviews were conducted with each participant. Participants of 1) should lie about the theft whereas the ones of 2) should tell the truth about their activities. The interviews consists of ten background questions for the truthful baseline and specific questions about the theft. The labels for the dataset correspond to the two conditions: deception and non-deception.

Concerning the Sincerity sub-challenge, the SINCERITY SPEECH CORPUS (SSC) is provided by the Columbia University. The database includes recordings of approximately 72 minutes of speech by 32 speakers. The individuals were asked to read six different sentences whereas the content of each sentence is a form of an apology. Moreover, each sentence was expressed in different prosodic styles: monotonic, non-monotonic, slow, and fast. Each instance of the dataset was rated by at least 13 annotators. The golden standard for sincerity consists of the average standardized ratings across all annotators.

3. Phone Recognition-based Features

Previous works have shown effects of speaker states and traits on speech disfluency respectively speaking rate and durations using ASR systems (e.g., [14], [15], and [16]). We composed a set of phone recognition-based features which, in combination with acoustic feature sets, are aimed to improve the prediction of deceptive and sincere speech.

3.1. Speech Transcription

We decoded speech using the open-source framework sphinx-4 [17]. The ASR system has been applied as an open-loop speech recognizer, i.e., without lexical constraints. In order to detect the phone sequences we used the pretrained US English generic acoustic model (cmusphinx-en-us-5.2) provided by CMU. The model comprises the following units:

- 39 phonemes: *AA*, *AE*, *AH*, *AO*, *AW*, *AY*, *B*, *CH*, *D*, *DH*, *EH*, *ER*, *EY*, *F*, *G*, *HH*, *IH*, *IY*, *JH*, *K*, *L*, *M*, *N*, *NG*, *OW*, *OY*, *P*, *R*, *S*, *SH*, *T*, *TH*, *UH*, *UW*, *V*, *W*, *Y*, *Z*, *ZH*
- 1 silent pause: *SIL*
- 6 filled pauses: *BREATH*, *COUGH*, *NOISE*, *SMACK*, *UH*, *UM*

We implemented the phone recognition system and obtained the speech transcriptions for each audio file including phonemes, silent pauses, and filled pauses with the corresponding timecodes in milliseconds. Finally, feature extraction was performed using these transcripts.

3.2. Feature Extraction

We considered each utterance from the first to the last phone based on voice activity detection of the phone recognizer, i.e., silence or noise at the beginning and the end of the audio file were left out. As a further indicator, vowels are detected (e.g., *AA* and *OW*) and considered as a separate feature group. Moreover, we derived pseudo syllable patterns based on the concept of the consonant-vowel (CV) structure [18]. For instance, *CV*, *V*, and *CCCV* are valid patterns. The four feature groups phonemes, vowels, pseudo syllables, and pauses were then considered for the extraction of the following 29 static features:

- Durations with statistics: Durations are derived from the timecodes of phonemes, vowels, pseudo syllables, and pauses. In order to take dynamic characteristics of an utterance into account, the functionals minimum, maximum, range, arithmetic mean, and standard deviation were applied to phonemes, vowels, and pseudo syllables. To pauses only the arithmetic mean was applied.
- Speaking rates: the average number of units per second computed for phonemes, vowels, and pseudo syllables.
- Vowel-to-phoneme occurrence ratio: the number of vowels divided by the number of phonemes (including vowels).
- Pause-to-phoneme occurrence ratio: the number of pauses divided by the number of phonemes.
- Pause duration ratio: the total duration of pauses divided by the total duration of the utterance.
- Silence occurrence: the total number of silent pauses.
- Silence-to-phoneme occurrence ratio: the number of silent pauses divided by the number of phonemes.
- Silence duration: the total duration of silent pauses.
- Silence duration ratio: the total duration of silent pauses divided by the total duration of the utterance.
- Filler occurrence: the total number of filled pauses.
- Filler-to-phoneme occurrence ratio: the number of filled pauses divided by the number of phonemes.
- Filler duration: the total duration of filled pauses.
- Filler-to-pause duration ratio: the total duration of filled pauses divided by the total duration of pauses (including filled pauses).

We hypothesize that there are detectable changes of speech disfluency respectively speaking rate and durations when speech is more deceptive or more sincere. Table 1 shows the comparison of deceptive vs. non-deceptive speech and sincere vs. insincere speech using the speaking rate with pseudo syllables, the arithmetic mean of the vowel durations, and the pause duration ratio averaged across the training sets of the corresponding databases. As the rating scales (golden standard) of instances in the SSC database contain continuous values we defined sincere speech with rating > 0 and insincere speech with rating < 0.

Table 1: *Results of pseudo syllable-based speaking rate (SR), pause duration ratio (PR), and the arithmetic mean of vowel durations in milliseconds (VD) averaged according to the type of speech on the DSD and SSC training sets.*

Type of Speech	SR	PR	VD
Deceptive	3.72	0.124	124
Non-deceptive	3.53	0.121	131
Sincere	3.68	0.099	96
Insincere	3.61	0.077	124

It can be seen that speaking rate has a higher value in deceptive than in non-deceptive speech, which correlates with the lower durations of vowels. Consequently, the utterance in deceptive speech seems to be more prepared by the speaker while the slightly higher pause duration ratio indicates more speech disfluency. Sincere speech has the same characteristics whereas insincere speech as a white lie cannot be mixed up with deceptive speech. In this study, we note that deceptive and insincere speech are not directly comparable. Moreover, the SSC

database was created with individuals who were asked to read the sentences, i.e., the pause duration ratio feature seems to be inappropriate for sincerity recognition.

4. Experiments

The main goals of the experiments are to verify the potential of the introduced automatic phone recognition-based features and to improve the baseline results of the Deception and Sincerity sub-challenge by fusing these features with state-of-the-art acoustic feature sets. Acoustic feature extraction was performed using the openSMILE toolkit (version 2.1 public release)[19] with its corresponding configuration files.

The evaluation measure for Deception is the unweighted average recall (UAR), because the distributions of instances in the DSD database are highly unbalanced among classes. We considered this fact by upscaling the instances of deceptive speech in the training set. The given development set was used for optimization and evaluation before testing. The official metric for the Sincerity sub-challenge is the Spearman's rank correlation coefficient ρ. As no development set is provided in the SSC database and the training set is rather small we performed leave-one-speaker-out cross-validation (LOSO-CV) on the training set in order to assess the performance before testing.

For both sub-challenges, all features in the training set were standardized using z-score normalization, i.e., the mean is 0 and the variance is 1. The standardization of the development and test set were computed with parameters only from the training set. For Sincerity, no separate standardization was performed in each LOSO-CV fold during development.

The prediction of deception, as a classification task, was performed using support vector machine (SVM) with a linear kernel function. Support vector regression (SVR) is used to predict the continuous values of sincerity, with linear kernels as well. For both tasks, Sequential Minimal Optimisation (SMO[20]/SMOreg[21]) is applied as training algorithm implemented in the Weka 3 data mining toolkit (revision 3.7.13)[22]. We optimized the complexity parameter C of the algorithm according to each applied feature set.

4.1. Phone-based Feature Group Evaluation

In this experiment we compare the performance of each proposed phone-based feature group on the datasets. As described in section 3.2, the introduced 29 features are partitioned into the four feature groups: phonemes (6 features), vowels (7 features), pseudo syllables (6 features), and pauses (10 features). The optimized complexity of the SMO during development using all features is C=10^{-2} for both databases. The results can be seen in Table 2.

Regarding DSD, using all features resulted in an UAR of 58.6%. The best result with 60.9% could be achieved solely by pseudo syllables. We find that features based on pseudo syllables are more appropriate for the representation of rhythm in speech. On the other hand, deriving these features, pseudo syllables can tolerate errors from the phone recognizer such as insertions.

For the SSC databases, the group with the highest value (0.323) is based on vowels. Studies have already shown relevance for the discrimination of speaker states using features on vowel-level (e.g., [23]). Combining all feature groups resulted in $\rho = 0.356$ which is the best result for the prediction of sincerity in this experiment.

However, as each feature group has potential for the prediction of deception as well as sincerity we decided to use all 29 features for further experimentation.

Table 2: *Development results for the prediction of deception (D) and sincerity (S) using phone-based feature groups.*

Feature Group	# Features	D(UAR[%])	S(ρ)
Phonemes	6	58.3	0.194
Vowels	7	50.3	0.323
Pseudo Syllables	6	60.9	0.193
Pauses	10	52.8	0.153
All	29	58.6	0.356

4.2. Feature Fusion and Selection

Obviously, the phone-based features are not sufficient enough for the prediction of deception and sincerity but can be complemented by other acoustic features. For development, we assessed different acoustic feature sets from the Interspeech challenges of prior years on the DSD and SSC database.

The COMPARE feature set [24] (IS13) includes the official baseline features. This set comprises 6373 static features of various functionals computed over low-level descriptor (LLD) contours. For comparison, we fused our features with the IS13 (IS13+Ph) for both databases. The optimal complexity parameter of the training algorithm was C=10^{-4}. After feature fusion we obtained an UAR of 63.4% (baseline is 61.9%) on DSD and ρ = 0.498 on SSC (baseline is 0.474). On DSD, the INTERSPEECH 2009 Emotion Challenge feature set [25] (IS09) performed best. This set contains 384 features as statistical functionals applied to LLD contours. Using parameter optimization C=10^{-2} we obtained an UAR of 65.8%. Fusing this feature set with our phone-based features (IS09+Ph) improves the results with 0.3% to 66.1%. On the SSC database, the updated version of the INTERSPEECH 2012 Speaker Trait Challenge feature set [26] (see openSMILE configuration file IS12_speaker_trait.conf), which contains 5757 static features computed over LLD contours, yields better results than the baseline. With C=10^{-4} we obtained ρ = 0.488. The results could be further improved by fusing this feature set with our features (IS12+Ph) which results in $\rho = 0.502$.

For both databases – DSD and SSC – some features can be irrelevant and redundant. A common way to obtain an appropriate subset of features is feature selection. In order to avoid overfitting we decided to apply a filter method on the three new feature sets (IS09+Ph, IS12+Ph, and IS13+Ph). The ReliefF algorithm [27, 28] was used to rank the features of each set in connection with the databases. An incremental evaluation with a step size of 100 was performed using the ranked features. The criterion for feature selection was the number of features which leads to the best result.

Figure 1 illustrates the UAR obtained using the feature sets IS09+Ph as well as IS13+Ph and the corresponding selected number of ranked features on the DSD database. It can be seen that the IS13+Ph feature set outperforms the baseline at the feature number of 400 with 62.0%. The highest value using these features is at 6000 (all Ph features are included) with an UAR of 64.0% which corresponds to an improvement of 2.1% over the baseline. In contrast, the same result could be achieved by the IS09+Ph features at the number of 200. However, the IS09+Ph feature set performed best with 66.7% and an improvement of 4.8% over the baseline using only 400 features (without filler

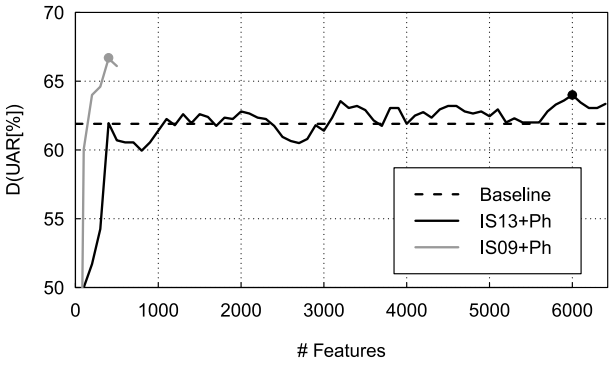

Figure 1: *Prediction of deception on the DSD development set according to the number of ranked features in the training set.*

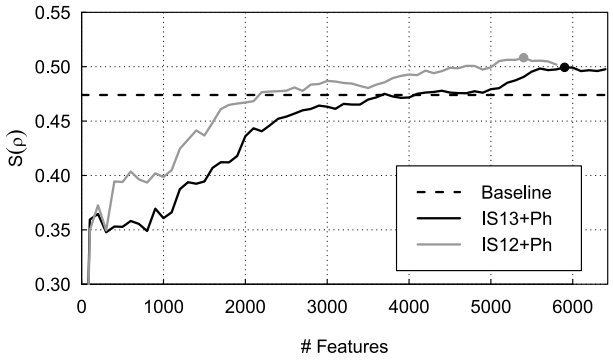

Figure 2: *Prediction of sincerity on the SSC training set (LOSO-CV) according to the number of ranked features.*

duration and filler-to-phoneme occurrence ratio).

Figure 2 shows the results of the Spearman's rank correlation coefficient ρ using the regressional ReliefF algorithm according to the IS13+Ph and IS12+Ph feature sets on the SSC database. Concerning the IS13+Ph, the baseline could be outperformed at the number of 3700 (0.001 over the baseline). The optimized feature number is located at 5900 (with all Ph features) and yields 0.499. The IS12+Ph performs best with 0.508 at the feature number of 5400 including all Ph features. Nevertheless, at the number of 2200 this feature set already achieved an improvement compared with the baseline.

4.3. Results

A series of experiments was carried out for the prediction of deception and sincerity using the databases of the corresponding sub-challenges. The results can be seen in Table 3. For both sub-challenges, the development results (Dev) refer to the sections 4.1 and 4.2.

Regarding the Deception sub-challenge, each configured feature set, except our phone-based features (Ph), achieved better UAR results than the baseline (61.9%) on the development set. However, the best performance was obtained by feature fusion of the IS09 feature set with our phone-based features and feature selection (IS09+Ph & feature selection). The result is 66.7%, 4.8% over the baseline. In order to obtain the performance on the test set we concatenated the training and development sets to a new training set for each feature configuration. The values of the optimized complexity parameter C were the same as in the development phase. IS09+Ph and IS13+Ph performed better than the baseline on the test set. Moreover, using feature selection on IS13+Ph the UAR increases up to 69.3% with a corresponding improvement of 1.0%.

For the Sincerity sub-challenge, all proposed new feature sets, except phone-based features on its own, gave better correlation results than the baseline (ρ = 0.474) in the development phase. The highest value with ρ = 0.508 was obtained using feature fusion of IS12 with phone-based features and feature selection (IS12+Ph & feature selection). Concerning the test set, we utilized the full training set for model construction with optimized complexity parameters determined in the development phase. Only the IS13+Ph method achieved better results than the baseline on the test set. Feature selection decreases the performance, but the result of the Spearman's rank correlation coefficient ρ is still above the baseline. Without feature selection we obtained the best result with ρ = 0.611 which corresponds to an improvement of 0.009.

Table 3: *Summary of the final results for the Deception (D) and Sincerity (S) sub-challenge. The best development and test results are highlighted in bold.*

Method	Dev	Test
D(UAR[%])		
IS13 (baseline)	61.9	68.3
Ph	58.6	-
IS09+Ph	66.1	68.7
IS09+Ph & feature selection	**66.7**	67.8
IS13+Ph	63.4	68.5
IS13+Ph & feature selection	64.0	**69.3**
S(ρ)		
IS13 (baseline)	0.474	0.602
Ph	0.356	-
IS12+Ph	0.502	0.598
IS12+Ph & feature selection	**0.508**	0.599
IS13+Ph	0.498	**0.611**
IS13+Ph & feature selection	0.499	0.608

5. Conclusions

In this paper we have investigated the potential of automatic phone recognition-based features for the prediction of deception and sincerity from speech. Therefore, we designed a feature set including 29 static features associated with durations, speaking rates, and ratios. These features are applied as indicators regarding changes in speech disfluency respectively speaking rate and rhythms. Obviously, the features are important but not sufficient enough on its own for the prediction of deception and sincerity. Development results show a difference with 3.3% of UAR for deception and Spearman's rank correlation coefficient ρ = 0.118 for sincerity compared to the baseline. For optimization we fused our features with carefully selected acoustic feature sets. Moreover, we demonstrated how to further boost the performance of the predictions for both sub-challenges by refining these fused features using feature selection. Finally, experimental results show that the combination of phone-based features with the baseline acoustic feature set of the challenge performed best on the test set for both, deception and sincerity. For deception, we obtained the best result by applying feature selection (UAR = 69.3%). For sincerity, the best result obtained on the test set was ρ = 0.611 without feature selection.

6. References

[1] S. Rigoulot, K. Fish, and M. D. Pell, "Neural correlates of inferring speaker sincerity from white lies: An event-related potential source localization study," *Brain research*, vol. 1565, pp. 48–62, 2014.

[2] 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.

[3] 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*. Association for Computational Linguistics, 2009, pp. 309–312.

[4] 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.

[5] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple, "Pitch changes during attempted deception." *Journal of personality and social psychology*, vol. 35, no. 5, p. 345, 1977.

[6] K. Gopalan and S. Wenndt, "Speech analysis using modulation based features for detecting deception," in *The 15th International Conference on Digital Signal Processing*, 2007, pp. 619–622.

[7] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[8] J. Eriksson, "Straight talk: conceptions of sincerity in speech," *Philosophical studies*, vol. 153, no. 2, pp. 213–234, 2011.

[9] J. Trouvain, S. Schmidt, M. Schröder, M. Schmitz, and W. J. Barry, "Modelling personality features by changing prosody in synthetic speech," 2008.

[10] K. P. Rankin, A. Salazar, M. L. Gorno-Tempini, M. Sollberger, S. M. Wilson, D. Pavlic, C. M. Stanley, S. Glenn, M. W. Weiner, and B. L. Miller, "Detecting sarcasm from paralinguistic cues: anatomic and cognitive correlates in neurodegenerative disease," *Neuroimage*, vol. 47, no. 4, pp. 2005–2015, 2009.

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

[12] R. Rakov and A. Rosenberg, "" sure, i did the right thing": a system for sarcasm detection in speech." in *INTERSPEECH*, 2013, pp. 842–846.

[13] H. S. Cheang and M. D. Pell, "The sound of sarcasm," *Speech communication*, vol. 50, no. 5, pp. 366–381, 2008.

[14] A. Zlotnik, J. M. Montero, R. San-Segundo, and A. Gallardo-Antolín, "Random forest-based prediction of parkinson's disease progression using acoustic, asr and intelligibility features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] C. Montacié and M.-J. Caraty, "High-level speech event analysis for cognitive load classification." in *INTERSPEECH*, 2014, pp. 731–735.

[16] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, and G. Tur, "Speech recognition as feature extraction for speaker recognition," in *Signal Processing Applications for Public Security and Forensics, 2007. SAFE'07. IEEE Workshop on*. IET, 2007, pp. 1–5.

[17] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Mountain View, CA, USA, Tech. Rep., 2004.

[18] J. Farinas and F. Pellegrino, "Automatic rhythm modeling for language identification." in *INTERSPEECH*, 2001, pp. 2539–2542.

[19] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[20] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to platt's smo algorithm for svm classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[21] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the smo algorithm for svm regression," *Neural Networks, IEEE Transactions on*, vol. 11, no. 5, pp. 1188–1193, 2000.

[22] 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.

[23] F. Ringeval and M. Chetouani, "A vowel based approach for acted emotion recognition." in *INTERSPEECH*, 2008, pp. 2763–2766.

[24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[25] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009. Citeseer, 2009, pp. 312–315.

[26] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge." in *INTERSPEECH*, vol. 2012.

[27] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Machine Learning: ECML-94*. Springer, 1994, pp. 171–182.

[28] M. Robnik-Šikonja and I. Kononenko, "An adaptation of relief for attribute estimation in regression," in *Machine Learning: Proceedings of the Fourteenth International Conference (ICML97)*, 1997, pp. 296–304.