# Does my speech rock? Automatic assessment of public speaking skills

**L. Azais, A. Payan, T. Sun, G. Vidal, T. Zhang, E. Coutinho, Florian Eyben, Björn Schuller**

**Angaben zur Veröffentlichung / Publication details:**

Azais, L., A. Payan, T. Sun, G. Vidal, T. Zhang, E. Coutinho, Florian Eyben, and Björn Schuller. 2015. "Does my speech rock? Automatic assessment of public speaking skills." In *INTERSPEECH 2015 - 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2519–23. ISCA Archive. https://doi.org/10.21437/Interspeech.2015-543.



# Does my Speech Rock? Automatic Assessment of Public Speaking Skills

*Lucas Azaïs*[1], *Adrien Payan*[1], *Tianjiao Sun*[1], *Guillaume Vidal*[1], *Tina Zhang*[1], *Eduardo Coutinho*[1], *Florian Eyben*[2], *Björn Schuller*[1]

[1]Department of Computing, Imperial College, London, UK
[2]audEERING UG (limited), Gilching, Germany

e.coutinho@imperial.ac.uk

10.21437/Interspeech.2015-543

## Abstract

In this paper, we introduce results for the task of Automatic Public Speech Assessment (APSA). Given the comparably sparse work carried out on this task up to this point, a novel database was required for training and evaluation of machine learning models. As a basis, the freely available oral presentations of the ICASSP conference in 2011 were selected due to their transcription including non-verbal vocalisations. The data was specifically labelled in terms of the perceived oratory ability of the speakers by five raters according to a 5-point Public Speaking Skill Rating Likert scale. We investigate the feasibility of speaker-independent APSA using different standardised acoustic feature sets computed per fixed chunk of an oral presentation in a series of ternary classification and continuous regression experiments. Further, we compare the relevance of different feature groups related to fluency (speech/hesitation rate), prosody, voice quality and a variety of spectral features. Our results demonstrate that oratory speaking skills can be reliably assessed using supra-segmental audio features, with prosodic ones being particularly suited.

**Index Terms**: Automatic Public Speech Assessment, database, classification, regression, prosody.

## 1. Introduction

Advances in signal processing and machine learning techniques have resulted in the development of a variety of applications of Computational Paralinguistics. These applications include affect recognition (e.g., [1, 2]), the automatic detection of speaking disabilities (e.g., [3, 4, 5]), or automatic evaluation of language proficiency (e.g., [6, 7, 8]). However, to the best of our knowledge, little work has been done in the field of automatic assessment of public speaking skills, despite the fact that such an application could help speakers to evaluate and improve their performance. This is particularly relevant as there are a myriad of circumstances where good public speaking skills can facilitate career advances (e.g., create new professional opportunities) and also bring benefits in other areas, including personal life (e.g., speech at a friend's wedding, inspire a group of volunteers at a charity event). Another indicator of the importance of automatically assessing public skills is the fact that there is commercial interest in such a product. Already in 2006, Hewlett Packard (HP) filed a patent [9] describing a "System and method of providing evaluation feedback to a speaker while giving a real-time oral presentation". To our knowledge such a system has yet to be developed, and will be the focus of this paper.

There are many components that come together to make a good speech - not only does the content, structure and language need to resonate with the audience, but the speaker who is projecting the message through his/her voice, body language and visual aids also needs to be engaging. In this paper we focus on what can be assessed from the voice of the speaker. According to the literature related to public speaking [10, 11, 12, 13], the most important voice-related characteristics to optimise for a speech are the volume (needs to be appropriate to room size), pauses and pacing, intonation, clarity (avoiding "filler" hesitation words) and energy. Interestingly, these attributes refer to paralinguistic aspects of speech and can be estimated from low level acoustic descriptors. Therefore it is plausible to hypothesise that paralinguistic analysis of speech can be used to predict the quality of the speaker.

Previous research in speaker quality assessment has focused on multimodal analysis. For instance, [14] proposed a system that used vocal, visual and lexical inputs from audio, video and 3D motion capturing devices. Interestingly they found that among the three recorded modalities, speech features (related to speech delivery e.g. speaking rate, prosodic variations, pausing) provided the most information. In another work, [15] proposed a proof-of-concept system to develop a platform for public speaking training based on the analysis of speech flow (intrinsic rhythm) and clarity of intonation, gestures and gaze patterns. Regarding voice analysis, they show that descriptors such as number of pauses, average intensity, breathiness, number of pause fillers, fundamental frequency (F0) and spectral stationarity were highly correlated with expert assessments of speech quality (flow of speech, intonation, volume, interruptions and vocal variety). [16] proposed a presentation training system giving rule-based feedback to the user with respect to speaking rate, eye contact with audience and timing, using both speech and image processing techniques. Feedback on speech is given on both speaking rate and F0 in the cases where they respectively exceed or fall below a predefined threshold. Finally, [17] proposed the only method to assess public speaking skills exclusively from paralinguistic features. They trained a set of six Support Vector Machine (SVM) classifiers (Radial Basis Function kernels) each predicting a distinct public speaking quality dimension as defined by an expert evaluator. These qualities were *clear*, *competent*, *credible*, *dynamic*, *persuasive* and *pleasant*. The authors used a set of 6552 acoustic features (openSMILE's "*emo_large*" default feature set) - prosodic, voice quality and spectral features [18]), with feature selection The performance of their system (averaged across the six binary classifiers – for each label) reached 81% (chance level of 50%). Nonetheless, the database used was very small (124 instances).

In this paper we introduce a new annotated database of public speech quality, and present a series of classification and regression experiments for the automatic assessment of speech quality. The remainder of this paper is organised as follows:

Section 2 introduces the datasets of audio recordings used. Section 3 introduces the acoustic feature sets used for the automatic assessment of speech quality. Section 4 details our classification and regression experiments, and Section 5 concludes this paper with a summary of our findings and a discussion of future areas for follow-up research.

# 2. Database Description and Labelling Procedure

As far as we are aware, there is no database of single speaker speeches that has been labelled in a consistent fashion according to public speaking ability. Therefore, in this paper we introduce a new database that makes use of the ICASSP 2011 conference talks for the creation of a new labelled database for Automatic Public Speaking Assessment (APSA)[1]. The ICASSP 2011 conference talks were chosen in particular, as the speech transcripts of each recording include syllable count and annotations of hesitations (such as "uh" or "um"), which allows the calculation of speech and hesitation rates two important features of public speaking quality.

## 2.1. Recordings

We downloaded a set of ICASSP 2011 conference recordings from the Superlectures website (http://superlectures.com/). From the full list of available recordings, we have chosen only single speaker recordings (e.g., panel discussions were discarded), and one talk per speaker. For each talk, we removed the initial 60 seconds (to cut out speaker preparation and introduction) as well as the last 25% of the recording (this cut off was decided after a sample of videos indicated that the Q&A sessions were typically included in this section of the recordings). Then, for each speaker, we extracted three 30 seconds segments from the beginning, middle and end of each selected talk as it provides a sufficient and suitable time frame over which adequately assess a speech quality. Further, using the annotations on syllable counts included with the talks (which will be described later), we discard those speakers with at least one of the 30 seconds segments with fewer than 100 spoken syllables per minute as this is an unusually low syllable count and was found to be often the result of recording problems or a consequence of speakers making live demonstrations (e.g., playing a music recording). The final database, henceforth referenced as *APSA-ICASSP*, includes 228 distinct speakers (26 female, 202 male) and a total of 684 instances (3 per speaker).

## 2.2. Annotations and Ground Truth

Five postgraduate students from the Imperial College London (4 male, 1 female, aged between 22-30) were asked to label the *APSA-ICASSP* database in terms of perceived oratory ability (see Table 1). Labelling was carried out on an independent basis in various sessions (one hour maximum to avoid fatigue) over a 1 week period. Annotators used a 5-point Likert scale ranging from -2 (*very bad speaker*) to 2 (*very good speaker*) to provide their ratings. Henceforth, we refer to this scale as 5-point Public Speaking Skill Rating ($PSSR_5$). The meaning of each level of the rating scale was explained beforehand to all annotators as shown in Table 1. Annotators were specifically asked to rate the perceived oratory ability while ignoring the verbal contents of the talk and the pronunciation of the speaker.

We computed inter-rater agreement in terms of pairwise Spearman rank correlation ($\rho$) and normalised Cronbach's Alpha ($\alpha$). Both statistics indicate an good level of inter-rater agreement – $\rho = .470$ and $\alpha = 0.840$. Then, from the individual ratings, we computed the ground truth using the Evaluator Weighted Estimator (EWE, [19]). This method computes a weighted average based on each rater's correlation with the mean ratings. The EWE will be used in our regression experiments (hereinafter $PSSR_{EWE}$). In order to derive discrete class labels from the continuous ratings, we rounded the $PSSR_{EWE}$ for each instance to the nearest discrete rating (see last column of Table 1). Due to the low number of instances with a $PSSR_5$ score of 2 and -2, we merged labels 1 and 2, and -1 and -2, which resulted in three cover classes {Positive, Neutral, Negative} (henceforth referred to as $PSSR_3$). This procedure led to a more balanced class distribution with {244, 321, 119} instances per class ({Positive, Neutral, Negative}, respectively).

Table 1: *Guidelines used as a reference by raters to assign $PSSR_5$ scores, total number of instances per score, number of instances with female speakers per score*

| $PSSR_5$ | Description | #Instances (#Female) |
|---|---|---|
| **-2** | **Very bad speech**, excessive pauses/hesitations that frustrate the listener and poor performance on the majority of prosodic aspects | 12 (0) |
| **-1** | **Bad speech**, some negative traits observed such as too monotonous, or too fast, or too many hesitations/pauses | 107 (5) |
| **0** | **Average/neutral speech**, some weaknesses, but generally intelligible and easy to listen to | 321 (25) |
| **1** | **Good speech**, few hesitations/pauses, good flow that sparks interest | 228 (45) |
| **2** | **Very good speech**, almost no hesitations, enthusiastic and natural speaker | 16 (3) |

# 3. Features

For the acoustic modelling of our public speaking databases, we used two standard features sets – the INTERSPEECH 2013 Computational Paralinguistics ChallengE (ComParE) [20] official feature set (COMPARE13), and the extended Geneva Minimalistic Acoustic Parameter Set(EGEMAPS; [21]). For the sake of reproducibility, feature extraction was carried out using the openSMILE toolkit [18]

The COMPARE13 feature set comprises 6 373 static features of functionals of low-level descriptor (LLD) contours (a full description of the feature set is available in [22]). The EGEMAPS is a much smaller feature set comprising 88 LLDs and respective functionals, and it was designed as a standard acoustic parameter set for automatic voice analysis (details shown in Table 2). With the purpose of evaluating the importance of prosody-related features for APSA, we divided both features sets into prosody-(COMPARE13-P and EGEMAPS-P) and non-prosody i.e. spectral/voice quality-related (COMPARE13-NP and EGEMAPS-NP) features. COMPARE13-P comprises 483 features corresponding to the LLDs in the "prosodic" group detailed in [22] and their related functionals. The remaining COMPARE13 features (5890) were assigned to the COMPARE13-NP set. As for EGEMAPS-P, it includes 26 features comprising all Pitch

[1]The database can be obtained from the authors on demand.

Table 2: *The* eGeMAPS *Feature Set is composed of 18 low-level descriptors (LLD) separated into 3 Parameter Groups, along with applied functionals. Asterisk denotes features included in the* eGeMAPS-P *feature set.*

| |
|---|
| **6 frequency related LLD** |
| Pitch*, Jitter, Formant 1,2, and 3 frequency, Formant 1 bandwidth |
| **3 energy/amplitude related LLD** |
| Shimmer, Loudness*, HNR (Harmonics-to-Noise Ratio) |
| **9 spectral related LLD** |
| Alpha Ratio, Hammarberg Index, Spectral Slope 0-500Hz and 500-1500Hz, Formant 1,2 and 3 relative energy, Harmonic difference H1-H2, Harmonic Difference H1-A3 |
| |
| **Functionals applied to all LLD (36)** |
| Arithmetic mean, coefficient of variation (standard deviation normalised by arithmetic mean) |
| **Functionals applied to loudness and pitch (52)** |
| 20th, 50th, 80th percentile, range of 20th-80th percentile, mean and standard deviation of the slope of rising/falling signal parts |
| **Functionals applied to Alpha Ratio, Hammarberg Index, Spectral Slope 0-500Hz and 500-1500Hz (56)** |
| Arithmetic mean over all unvoiced segments |
| **Additional functionals resulting in temporal features* (6)** |
| Rate of loudness peaks, mean length and standard deviation of voiced segments ($F_0 > 0$), mean length and standard deviation of unvoiced regions, number of continuous voiced regions per second |

and Loudness LLDs and related functionals, plus 6 functionals related to temporal features (denoted with an asterisk in table 2). All other features (62) were assigned to the eGeMAPS-NP set.

As mentioned earlier, the ICASSP 2011 conferences were transcribed with structural information that allowed us to compute two additional measures – speech and hesitation rates. Each recording had an accompanying XML file comprising a set of transcribed strings, each referenced with a timestamp. The speech rate was calculated using the syllable count over a given time segment by computing the number of vowels per string and linearly interpolating if the desired time segment started or ended in between 2 timestamps. The algorithm uses "a,e,i,o,u,y" as vowels, and counts the number of non-consecutive vowels. If the word ends in "es", or if the word ends in 'e' but the preceding character is not 'l', the syllable count is decremented by 1 count. All words shorter than 3 characters were considered as monosyllabic. Due to many peculiarities in the English language it is not possible to get a perfectly accurate syllable count unless a full English dictionary look up is used for every word. Nevertheless, we have tested that the syllable count algorithm gives correct results to a tolerance of 10%. The hesitation rate was calculated by counting the number of instances of transcribed "um" and "uh" interjections divided by the total number of syllables spoken over the segment. The hesitation rate, together with the speech rate described earlier, were used to create a new feature set (SHRATE).

## 4. Experiments and Results

The feature sets (ComParE13-P, ComParE13-NP, ComParE13, eGeMAPS-P, eGeMAPS-NP, eGeMAPS and SHRATE) and ground truths ($PSSR_{EWE}$ and $PSSR_3$) described earlier were used in a series of regression and classification experiments. For both classification and regression tasks we compared the performance of the models using all 7 feature sets in order to determine the most relevant types of features for APSA. For the classification tests we used Support Vector Machines (SVM), and for regression Support Vector Regression (SVR). Both models were implemented with WEKA [23]. For both SVM and SVR models we used polynomial Kernels of degree 1, and used the Sequential Minimal Optimization (SMO) algorithm for training. The database instances were split into equally sized, speaker-independent sets (using a modulo 3 scheme on the speaker IDs). Gender distribution as well as label distribution was checked to be relatively even across the three sets (see Table 3). Then, we used a 3-fold cross-validation schema to optimise the models' parameters and to estimate the performance on the test set. For the classifications tests, the training sets in each fold were up-sampled by integer values to achieve an even balance between classes (this implied upsampling the majority classes as well). No upsampling was performed for the regression tests. In each cross-validation fold, all sets were standardised to the mean and standard deviation of the respective training set. The validation set of each fold was used to estimate the best soft-margin complexity hyperparameter $C$ (we tested 0.0001, 0.001, 0.01, 0.1, and 1). After the best complexity hyperparameter was determined, we joined the training and development sets in each fold, retrained the models using the optimised $C$, and estimated the performance on the respective test set. In the regression tests we computed the unweighted average recall (UAR) as performance measure, whereas in the regression tests we used Spearman's rank correlation coefficient ($\rho$). The results of both classification and regression tests (averaged across all folds) are shown in Table 4.

As shown in Table 4, all feature sets yielded a classification performance well above chance level (33%). In relation to the eGeMAPS features, we found that the prosodic subset (eGeMAPS-P) led to the best results both in classification

Table 3: *Distribution of $PSSR_3$ per speaker-independent split used in our experiments (228 instances per split). Number of speakers per gender in each split is also indicated – Total (Male/Female).*

| Split | $PSSR_3$ | | |
|---|---|---|---|
| | Positive | Normal | Negative |
| 1 | 72 (59/13) | 105 (100/5) | 51 (48/3) |
| 2 | 81 (59/22) | 111 (104/7) | 36 (35/1) |
| 3 | 91 (78/13) | 105 (92/13) | 32 (31/1) |

Table 4: *Performance measures on the test sets (averaged across all folds) for APSA classification (UAR) and regression ($\rho$) experiments using the various feature sets (and their combination). Chance level for the classification tests: 33%.*

| Feature Set | #Features | UAR (%) | $\rho$ |
|---|---|---|---|
| ComParE13 | 6373 | 52.9 | **0.60** |
| ComParE13-P | 483 | 51.7 | 0.48 |
| ComParE13-NP | 5890 | 52.6 | **0.60** |
| eGeMAPS | 88 | 54.3 | 0.52 |
| eGeMAPS-P | 26 | 56.2 | 0.58 |
| eGeMAPS-NP | 62 | 49.5 | 0.48 |
| SHRATE | 2 | **59.5** | **0.59** |

($UAR = 56.2\%$) as in regression tasks ($\rho = 0.58$). Both results indicate that prosodic features are particularly relevant for APSA. This observations receives further confirmation from the results obtained with the SHRATE feature set. The classification accuracy was the highest of all sets tested ($UAR = 59.5\%$), and the second best ($\rho = 0.59$) in the regression experiments (with the top performance being a small distance away – $\rho = 0.60$). This outcome is not surprising given that speech and hesitation rates are important features of speech quality (and were reported by the raters as two of the most important cues used to provide their assessments). However, it is interesting to observe that this feature set consisting only two features achieves comparable prediction accuracy. In relation to the large feature set (COMPARE13) and its subsets, we found that the classification performance was generally worst than all other sets. Nonetheless, COMPARE13-NP performed better than EGEMAPS-NP which may indicate that a larger set of non-prosodic features may also include relevant information for APSA. In the regression task, instead, we found that COMPARE13 and COMPARE13-NP performed better than all other sets ($\rho = 0.60$), but very close to the performance of the SHRATE ($\rho = 0.59$) and EGEMAPS-NP ($\rho = 0.58$) sets. Taken together these findings suggest that prosodic features are particularly well suited for APSA.

## 5. Conclusions

In this paper we have presented a new database of audio recordings for Automatic Public Speech Assessment, and the perceptual study conducted to obtain judgements of the perceived oratory ability of each recording with respect to supra-segmental characteristics. These annotations (shown to be highly consistent across raters) were then used to compute the gold standard used in a series of classification and regression experiments to predict quality of the various speakers in our database. Additionally, we also compared the use of different predictors related to prosody (and fluency in particular – speech/hesitation rate), voice quality and a variety of spectral features.

Overall, using a speaker-independent schema, we have demonstrated that it is feasible to predict the quality of public speeches from acoustic features alone. This was evident in both classification and regression experiments, where we achieved, respectively, a top performance of 59.5% (UAR) and 0.60 ($rho$). Prosodic features were generally better predictors of oratory ability, which is in line with a priori intuitive expectations as discussed in section 1 and consistent with the findings in [14]. In particular, we found that only two prosody-related features (speech and hesitation rates) can provide similar explanatory power to a larger feature sets that encapsulates a wider range of acoustic features. Taking into account the fact that measures of speech and hesitation rates were manually computed from human annotations, and therefore may not always be readily available for future inputs to the model, we find that the relatively compact EGEMAPS-P prosodic feature set is certainly a good alternative as it provides strong performance in both classification and regression models. Finally, the substantially larger COMPARE13 feature sets did not result in significantly stronger classification results, but the additional number of features did seem to have an effect in regression.

In addition to expanding our database to include more speech corpora, there are other potential directions for future research that would be of interest. First, and although there was strong inter-rater agreement for *APSA-ICASSP* database, annotations were performed by non-expert annotators. Additional labelling by, for instance, speech coaches (such as those used in [24]) could provide a more solid ground truth. Furthermore, the use of separate labels for individual oratory quality features (e.g., fluency, pace, monotonicity, etc.) could also prove beneficial for establishing a more informative assessment of public speaking skills. In this case, each individual prosodic label could then be predicted with a targeted selection of features, and potentially lead to the development of a system that would give more granular feedback on the aspects of the public speaking skills of the speaker that should be developed. Second, given the explanatory power of speech and hesitation rates alone, further work should explore the use of algorithms for the automatic calculation of these measures (or the development of new ones). Third, given the relevance of prosodic features in general to the prediction of oratory quality, it would be important to investigate in more detail other specialised prosodic features such as those proposed in [25]. Finally, and although it was not the focus of this paper, alternative machine learning models could also be tested. Possible candidates are random forests ([26]) and convolutional neural networks (CNNs; [27]). In particular, the latter would be of special interest as CNNs have been shown to perform well for image recognition [28] due to their invariance to translations and other small transformations. This has analogies in audio processing if a 2 dimensional input layer was used (time versus the extracted features extracted), such that there would be overlapping receptive fields (each field as a matrix of all the features over a small chunk of time) across time.

## 6. Acknowledgements

The work of E. Coutinho and B. Schuller is funded by European Union's Horizon 2020 research and innovation programme under grant agreement No 645378, ARIA-VALUSPA. We are also very thankful to Felix Weninger for his valuable comments and suggestions throughout the development of this work.

## 7. References

[1] T. Sobol-Shikler and P. Robinson, "Classification of complex information: Inference of co-occurring affective states from their expressions in speech," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 7, pp. 1284–1297, 2010.

[2] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–577.

[3] D. Le and E. M. Provost, "Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 1563–1567.

[4] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer speech & language*, vol. 29, no. 1, pp. 132–144, 2015.

[5] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[6] K. Zechner and I. I. Bejar, "Towards automatic scoring of non-native spontaneous speech," in *Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics*. Association for Computational Linguistics, 2006, pp. 216–223.

[7] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in english pronunciation." in *ICSLP*, vol. 90, 1990, pp. 1185–1188.

[8] X. Xi, D. Higgins, K. Zechner, and D. M. Williamson, "Automated scoring of spontaneous speech using speechratersm v1. 0," *ETS Research Report Series*, vol. 2008, no. 2, pp. i–102, 2008.

[9] D. A. Silverstein and T. Zhang, "System and method of providing evaluation feedback to a speaker while giving a real-time oral presentation," May 23 2006, uS Patent 7,050,978.

[10] S. Lloyd-Hughes, *How to Be Brilliant at Public Speaking: Any Audience. Any Situation*. FT Press, 2011.

[11] P. Lieberman, "A study of prosodic features," in *Status report on speech research, SR-23*. Haskins Laboratories, New Haven, CT, 1970, pp. 179–208.

[12] J. H. Byrns, *Speak for yourself: An introduction to public speaking*. McGraw-Hill, 1994.

[13] L. M. Schreiber, G. D. Paul, and L. R. Shibley, "The development and test of the public speaking competence rubric," *Communication Education*, vol. 61, no. 3, pp. 205–233, 2012.

[14] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee, "Towards automated assessment of public speaking skills using multimodal cues," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 200–203.

[15] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, "Cicero-towards a multimodal virtual audience platform for public speaking training," in *Intelligent Virtual Agents*. Springer, 2013, pp. 116–128.

[16] K. Kurihara, M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi, "Presentation sensei: a presentation training system using speech and image processing," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 358–365.

[17] T. Pfister and P. Robinson, "Speech emotion classification and public speaking skill assessment," in *Human Behavior Understanding*. Springer, 2010, pp. 151–162.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[19] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 381–385.

[20] 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 INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, ISCA. Lyon, France: ISCA, August 2013, pp. 148–152, (acceptance rate: 52 %, IF* 1.05 (2010)), 63 citations.

[21] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *Transactions on Affective Computing*, submitted.

[22] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Psychology, Emotion Science, Special Issue on Expression of emotion in music and vocal communication*, vol. 4, no. Article ID 292, pp. 1–12, May 2013.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[24] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 66–78, 2011.

[25] F. Hönig, A. Batliner, T. Booklet, G. Stemmer, E. Nöth, S. Schnieder, and J. Krajewski, "Are men more sleepy than women or does it only look like – automatic analysis of sleepy speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 995–999.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *Multimedia, IEEE Transactions on*, vol. 16, no. 8, pp. 2203–2213, Dec 2014.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.