INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Sincerity in Acted Speech: Presenting the Sincere Apology Corpus and Results

Alice Baird [1], *Eduardo Coutinho* [2], *Julia Hirschberg* [3], *Björn Schuller* [1,4]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, Univeristy of Augsburg, Germany
[2] Department of Music, University of Liverpool, UK
[3] Computer Science Department, Columbia University, USA
[4] GLAM – Group on Language, Audio and Music, Imperial College London, UK
alice.baird@informatik.uni-augsburg.de

Abstract

The ability to discern an individual's level of sincerity varies from person to person and across cultures. Sincerity is typically a key indication of personality traits such as trustworthiness, and portraying sincerity can be integral to an abundance of scenarios, e. g. , when apologising. Speech signals are one important factor when discerning sincerity and, with more modern interactions occurring remotely, automatic approaches for the recognition of sincerity from speech are beneficial during both interpersonal and professional scenarios. In this study we present details of the Sincere Apology Corpus (SINA-C). Annotated by 22 individuals for their perception of sincerity, SINA-C is an English acted-speech corpus of 32 speakers, apologising in multiple ways. To provide an updated baseline for the corpus, various machine learning experiments are conducted. Finding that extracting deep data-representations (utilising the DEEP SPECTRUM toolkit) from the speech signals is best suited. Classification results on the binary (sincere / not sincere) task are at best 79.2 % Unweighted Average Recall and for regression, in regards to the degree of sincerity, a Root Mean Square Error of 0.395 from the standardised range [-1.51; 1.72] is obtained.

Index Terms: sincerity, acoustic features, deep data-representations, acted speech, speech corpus.

1. Introduction

Across cultures, an individual's sincerity towards a particular subject or action can be an important factor in interpersonal communication and is a point of deep discussion by ancient philosophers, including Confucius [1] and Aristotle [2]. Although definitions for sincerity exist, i. e. , *the absence of pretence, deceit, or hypocrisy*[1], it is said that the notion of sincerity is inherent to understanding one's own consciousness; in this way, defining sincerity is itself insincere – and a natural portrayal of sincerity may only come through naivety [3]. However, it is well known that the inference of sincerity can be manipulated, particularly in a commercial setting [4, 5]. Along with other factors such as; reputation, social status, and lack of motivation for personal gain, sincerity is strongly linked to overall *trustworthiness* [6]. Trustworthiness is a personality trait which has great financial benefit, with heads of companies attempting to utilise tactics such as voice coaching in an endeavour to improve perceived emotional intelligence including sincerity [7].

In the same way as trustworthiness, sincerity is also linked to the existence of deceptive traits, and computational approaches for speech-based deception recognition is a popular field of research [8, 9, 10]. Recognising the deceptive nature of an individual by voice alone has been shown to be very promising for uses-cases such as online and over the phone banking fraud detection [11, 12]. Although deception may be a cause for concern, it has been shown that individuals give more sincere responses to their dyadic partner when not in person [13].

An apology is one such spoken action which does raise expectations in regards to sincerity, and the ability to apologise sincerely in public-facing roles, such as by sportsmen (e. g. , Tiger Woods in 2010) and politicians (e. g. , Bill Clinton 1995) is crucial to a favourable outcome [14]. In the age of the Internet-of-Things, the ability to make a sincere apology has now become a factor of social-media-based culture, with a businesses method for dealing with *bad news* online now having a much larger impact [15]. Public figures now tend to choose social media platforms such as YouTube to apologise to their followers, as it has been shown to offer a level of sincere interpersonal communication which is less available through conventional media outlet strategies [16].

In this study, we present the Sincere Apology Corpus (SINA-C). SINA-C is a speech corpus of acted apologies which was first utilised in the research community for the INTERSPEECH 2016 Computational Paralinguistics challengE (COMPARE) [17]. Since then, additional studies have been made with SINA-C data [18, 19]. However, here we introduce a baseline binary classification task (Sincere or Not Sincere), along with updated baseline results for the regression task of the standardised sincerity ratings. For both tasks 3 core feature sets are explored, including the hand-crafted acoustic-features of COMPARE [20] and the extended Geneva Minimalistic Acoustic Parameter Set (EGEMAPS) [21]. Both, COMPARE and EGEMAPS have been applied with success to similar paralingusitic tasks, such as deception recognition [10] and likeability [22]. As a state of the art approach, we also extract deep data-representations from the speech instances, utilising the DEEP SPECTRUM toolkit [23], exploring the efficacy of utilising spectrogram and mel-spectrogram-based feature representations.

2. The Sincere Apology Corpus

With the purpose of investigating the attributes of the human voice which convey sincerity, the Sincere Apology Corpus (SINA-C) was initially gathered between 2015–2016 at the Columbia University Computer Music Centre (CCMC) in New York City, United States of America. The dataset was also included in the INTERSPEECH 2016 COMPARE challenge [17], and prior to that in 2015 a subset of the dataset was also exhibited as part of a graduate-school art exhibition.

SINA-C is an English speech corpus of acted apologies in various prosodic styles[2]. The corpus is comprised of 912 audio

[1] cf. 'Sincerity' Oxford English Dictionary

[2] SINA-C is available at http://doi.org/10.5281/zenodo.3241253 (open access). Please cite this paper and [24] when using the corpus.

http://dx.doi.org/10.21437/Interspeech.2019-1349

Figure 1: *Two subjects from* SINA-C *recording audio at the Columbia University Computer Music Center in 2016.*

files (∑ 01 h: 10 m: 20 s, μ 4.6 s, ± 2.6 s). During processing of the data, some recordings were discarded due to poor recording quality, and some subjects recorded utterances multiple times, hence the imbalance across aspects of the corpus (cf. Figure 2).

2.1. Subjects

There are 32 subjects in SINA-C[3], 15 male and 17 female, aged between 20–60 years old, (mean; 29.8 years ± 9.9years). Individuals in the dataset were mostly (27 /32) American born English native speakers[4], but all subjects have a fluent level of spoken English. Instructed in a controlled enviroment (i. e. , the recording booth), all subjects were first given a description of the study and short definition for each of the prosodic styles. As subjects recited utterances in a sequential order, the recordings were not spontaneous.

2.2. Audio

Subjects were recorded in a recording studio at the CCMC (cf. Figure 1). Utilising a sound-proof recording booth, audio was recorded with an AKG C414 dynamic microphone. The digital audio workstation Logic Pro 9 was used and audio was initially captured at 44.1 kHz and 16 bit in *aiff* format and later converted to stereo *WAV*, for the final version of SINA-C and prior to any additional labelling.

2.3. Utterances

Subjects from SINA-C recorded 6 types of apologies:

1. *Sorry*. (∑ 179 instances. μ 2.03 s, ± 0.9 s)
2. *I am sorry for everything I have done to you*. (∑ 163 instances. μ 3.47 s, ± 1.42 s)
3. *I can not tell you how sorry I am for everything I did*. (∑ 135 instances. μ 4.36 s, ± 1.48 s)
4. *Please allow me to apologise for everything I did to you. I was inappropriate and lacked respect*. (∑147 instances. μ 6.82 s, ± 1.74 s))
5. *It was never my intention to offend you, for this I am very sorry*. (∑ 141 instances. μ 5.13 s, ± 1.55 s)
6. *I am sorry but I am going to have to decline your generous offer. Thank you for considering me*. (∑ 147 instances. μ 6.33 s, ± 1.65 s)

The lexical content of the utterances used for SINA-C were adapted from formulations of apologies as discussed in [25].

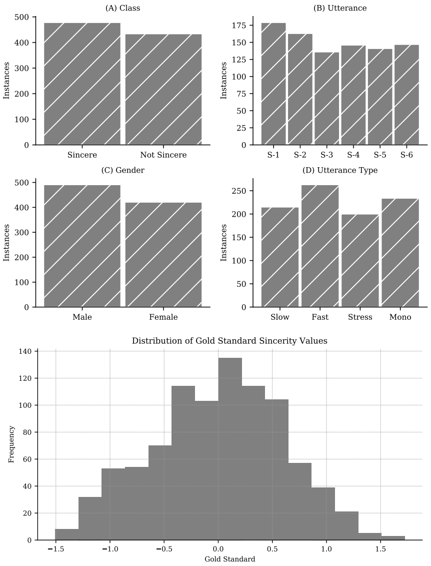


Figure 2: *Above, total instances for attributes of* SINA-C . *Class (Sincere or Not Sincere), utterance, gender and utterance type. Below, distribution of the gold standard ratings for Sincerity.*

Each sentence has a varied duration, which may allow for further prosodic analysis of the individual speakers. The apology was chosen as a control sentence, as this is quite extensively discussed in language research relating to sincerity [26, 27, 28]. Of note, a sincere apology in a corporate scenario is seen to be the only way to achieve a favourable outcome [14], and an individual who successfully portrays a sincere apology may have actually no feeling of moral responsibility towards the topic [29].

2.4. Utterance type

The subjects within the corpus were instructed to utter each of the sentences in 4 prosodic styles (1) monotonic (2) pitch prominence (labelled as 'Stress') (3) speaking rate: Fast (4) speaking rate: Slow. Although there is limited research in these particular prosodic styles in relation to sincerity, these styles were chosen due to their discussion within the literature of speech qualities similar to sincerity such as deception [30]. For example, it is well known to law enforcement that factors of speech such as speaking rate and raised pitch often occur during an interview with a deceptive partner [31]. Apposing to this, it has been shown that less deceptive speech may include a more monotonous prosodic flow [32]. In this regard, automatic speech-based deception recognition has successfully utilised prosodic-based features [8, 10].

2.5. Annotations

The SINA-C audio data was labelled in terms of the sincerity perceived by listeners (*'How sincere was the apology you just heard?'*) on a 5-point Likert scale ranging from 0 (Not Sincere) to 4 (Very Sincere) by 22 volunteers (13 male and 9 female; age range: 18-22; μ 19.5 ± 1.0). Of the 22 annotators, all reported to have *normal hearing* and all were English speakers (6 reported to be bilingual with at least one other language).

Raw annotations were standardised to zero mean and unit standard deviation on a per-subject basis in order to eliminate potential individual ratings biases. We then computed the mean across all subjects for each utterance. This resulted in a set of

[3] All subjects gave informed consent for the use of their recordings, as well as for dissemination within academic research.

[4] Other nationalities: Danish, Japanese, British, Irish, Puerto Rican.

Table 1: *Speaker (#), instance distribution for* COMPARE *2016 Speaker Independent Folds (C-SIF), and the Speaker Independent nested Cross Validation (SICV) schema. Indicating gender (M)ale:(F)emale and class, Sincere (S) and Not Sincere (NC).*

C-SIF				
	Train	Val	Test	∑
# M:F	11 5:6	11 5:6	10 5:5	32 15:17
S NS	142 143	184 186	152 105	478 434
∑	285	370	257	912
Min:Max	-1.51 :1.42	-1.41 :1.72	-1.48 :1.59	-1.51 :1.72
SICV				
	Fold 1	Fold 2	Fold 3	∑
# M:F	10 5:5	10 5:5	12 5:7	32 15:17
S NS	129 119	161 104	188 211	478 434
∑	248	265	399	912
Min:Max	-1.51 :1.54	-1.44 :1.60	-1.48 :1.72	-1.51 :1.72

ratings ranging from $[-1.51, 1.72]$ (mean: -0.002 $\pm$ 0.60) which are used as the gold-standard of our regression experiments (cf. Figure 2). We also converted these ratings to binary labels. Average ratings > 0 were then labelled as 'Sincere' (S), and those ≤ 0 were labelled as 'Not Sincere' (NS) (cf. Figure 2). This resulted in 478 instances labelled as S and 438 as NS. These labels were used as the gold-standard for the classification task.

3. Experimental methodology

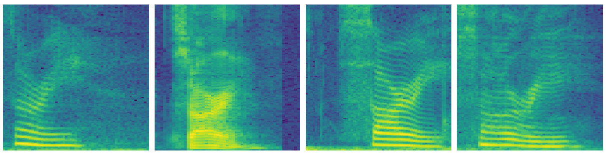
In this section, we describe the development of our machine learning model for the automatic inference of sincerity via audio using various features sets. We conducted both regression and classification experiments using the two gold-standards described in the previous section.

3.1. Feature sets

For the classification and regression sincerity-based tasks, we have extracted both hand-crafted speech based features (COMPARE and EGEMAPS), as well as deep representations (DEEP SPECTRUM). All features were extracted over the entire instance of audio and EGEMAPS and COMPARE feature sets were standardised to the mean and standard deviation of the training sets for each split. We do not perform additional feature selection on the extracted features (although this would be of interest to future work), instead we implement a brute-force approach utilising all features of each feature set.

The COMPARE feature set [20] of 6 329 hand-crafted dimensional speech-based features is used for our first approach. COMPARE has shown success for an abundance of paralinguistic tasks [33]. Additionally, we extract the 88 dimensional feature set EGEMAPS [21]. Similar to the COMPARE feature set, EGEMAPS has shown efficacy for similar classification and prediction tasks and is consistently effective for automatic approaches to classifying affective human states [34]. From each audio recording, the COMPARE and EGEMAPS acoustic features are extracted using OPENSMILE toolkit [35], and low level descriptors (LLDs) are left at the default setting, resulting in 2 feature set of 6 329 and 88 dimensions, respectively.

To obtain deep data-representations from the input audio, we use the feature extraction DEEP SPECTRUM toolkit[5]. DEEP SPECTRUM uses pre-trained convolutional neural networks (CNNs) to extract visual representations from audio [23]. We extract 4 types of DEEP SPECTRUM feature sets. Audio signals are transformed into spectrogram or mel-spectrogram plots (cf. spectrogram plots in Figure 3), and the generated representations are then forwarded through AlexNet [36]. For this study we extract features from the activations of the third and second to last (*fc6, fc7*), fully connected layer of the network. Resulting in 4 features sets of 4 096 dimensions; DS-1 - (Mel-Spectrogram, *fc6*) , DS-2 (Spectrogram, *fc6*), DS-3 (Mel-Spectrogram *fc7*) and DS-4 (Spectrogram, *fc7*). No other parameters for the LLDs are altered from the default provided with the toolkit.



(a) *597_train* (b) *290_train* (c) *810_train* (d) *438_train*

Figure 3: *Spectrogram (1 second, 5 Khz) representation of (1) 'Sorry'. monotonic (a, b) and speech rate: slow (c, d). Sincere (a, c) and Not Sincere (b, d).*

3.2. Training procedure

For our experiments we used Support Vector Machines (SVM) for classification tests and linear Support Vector Regression (SVR) for the regression ones. In both cases we used linear kernels and both SVM and SVR were implemented using the open-source machine learning toolkit Scikit-Learn [37].

We developed our models using two different approaches. First, and as with the INTERSPEECH 2016 COMPARE challenge [17], we split the dataset into two speaker independent (SI) sets: development and test (hereinafter we will refer to this approach as C-SIF). Both sets have the same instances for development and testing, as was used in the COMPARE 2016 challenge. However, unlike in the challenge (where a leave-one-speaker-out cross validation (LOSO-CV) schema was used during development), we opted to divide the development set into a single training and a single validation set. These sets were created by splitting development data in two SI folds (10 speakers each (cf. Table 1). During the development phase, we trained various models (using the training set) with different complexity parameters ($C \in 10^{-7}$, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1), and evaluated their performance on the validation set. After determining the optimal C, we concatenated the training and validation sets, re-trained the model with this enlarged training set, and evaluated the performance on the test set.

Our second approach consists of a 3-fold SI nested Cross Validation (SICV) schema to limit the effect of over-fitting the models to the training data as well as avoiding any subject-specific activity. In these experiments, we divided the whole database into 3 folds, balancing the number of subjects per fold as well as the gender of speakers (cf. Table 1).

Then, each fold was used successively as training, validation or test set. C was optimised based on the performance of the

[5] https://github.com/DeepSpectrum/DeepSpectrum

Table 2: *SVR and SVM with linear kernal, results on the test partition for all experiments made on* SINA-C . *Speaker Independent Folds (C-SIF) and nest Cross Fold Validation (SICV) - for both classification (Sincere /Not Sincere) and regression, range -1.51 :+1.71. Feature sets, (eGe)MAPS, (Com)PaRe, and 4 variations of* DEEP SPECTRUM *DS1–4. As performance metric for regression we report Root Mean Square Error (RMSE) and normalised RMSE (nRMSE) with the addition of Spearman's Correlation Coefficient (ρ). Unweighted Average Recall (UAR) is used for classification.*

Task	Schema	Metric	eGe	Com	DS 1	DS 2	DS 3	DS 4
Regression	C-SIF	RMSE	0.674	**0.571**	0.602	**0.572**	0.646	0.597
		nRMSE (%)	20.9	17.7	18.6	17.7	20.0	18.5
		ρ	0.390	0.420	0.432	**0.487**	0.340	0.417
	SICV	RMSE	0.989	0.489	0.454	**0.395**	0.449	0.416
		nRMSE (%)	30.6	15.1	14.1	12.2	13.9	12.9
		ρ	0.416	0.542	0.487	**0.594**	0.491	0.572
Classification	C-SIF	UAR (%)	**70.2**	70.0	69.8	69.9	65.9	68.6
	SICV	UAR (%)	72.0	76.2	75.9	78.8	75.1	**79.2**

validation sets over the 3 folds. Once the best value of C had been determined, we joined the training and development sets of each fold, retrained the models with the optimised C, and estimated the generalisation performance on the respective tests sets. For the classification experiments, due to some imbalance across the classes, we attempted to balance the class distribution by up-sampling the training set in each fold.

For experiments which used the COMPARE amd EGEMAPS features sets, the features inputs were standardised to the mean zero and unit variance (using the parameters of the training sets in each cross-validation fold – for SICV – or the single training set – for C-SIF). This method was not applied to the DEEP SPECTRUM features as it has not been shown to alter the model output [38].

4. Results and discussion

All results across all experiments for the task of both sincerity 2-class classification and the degree of sincerity regression task are presented in Table 2. We report both Unweighted Average Recall (UAR) and Root Mean Square Error (RMSE) as the measure of performance for classification and regression, respectively.

From the classification results of the C-SIF schema, it can be seen that the best UAR for all subjects is 70.2 % utilising the EGEMAPS features set, results with all other feature sets (aside from DS-3 and DS-4) are within a percentage point from the EGEMAPS result. In contrast to this, a much higher classification result is obtained though the SICV approach, with a 79.2 % UAR from the spectrogram-based features of DS-4.

When observing the regression results from the C-SIFs schema we see a similar trend with the spectrogram-based features (DS-2), achieving the lowest percentage of error (17.7 %). DEEP SPECTRUM features also perform best for the SICV regression task, with most prominence from the spectrogram features (DS-2 12.2 % and DS-4 12.9 %) over the mel-based results by approximately 2 percent points.

Although across all experimental settings we see that the conventional hand-crafted (COMPARE and EGEMAPS) features are competitive with the state of the art DEEP SPECTRUM approach, it is only in the first SVM experiment where they achieve best results, and even this is only marginal. This suggests that deep representations are able to capture more subtle attributes of sincerity, which may not include known speech features.

When evaluating the representation type for the DEEP SPECTRUM features – mel-spectrogram and spectrogram – it appears that, although the mel-spectrogram has had success in other emotion related tasks [38], the spectrogram is best suited over all. In regards to the CNN activation layer for extraction, little difference is seen between *fc6* and *fc7* although it would appear that for regression the *fc6*, performs over all best.

Additionally of note for the regression task, we have not beaten the original COMPARE 2016 challenge baseline, which was presented as 0.609 ρ. Although we do utilise the same train and test partition we did not use a LOSO-CV development and chose here to optimise using RMSE instead of ρ which we speculate may be the reason for this.

5. Conclusion and future work

In this study we presented the detail of SinA-C, including a full description of the data, as well as an updated classification and regression approach. The results from the study have shown that although conventional hand-craft features remain competitive, deep unsupervised data-representations (provided by DEEP SPECTRUM) are better suited to the task. Additionally, we see that SICV partitioning gives overall better results by 9 percent points over the C-SIF for the classification task, suggesting that features relating to sincerity may have a speaker-based bias.

With many use-cases to consider in regards to the automatic recognition of sincerity, such as non-verbal communication training i. e. , to the assist individuals who may have difficulties in inferring sincerity, further investigation into aspects including gender may be fruitful. As well as through a balanced expansion of the dataset, aspects including age and nativeness could also be explored, in this save way analysing more closely aspects prosody and lexical content for each utterances. Given the inherent similarity to other human traits, e. g. , likeability, trustworthiness and deception, transfer learning approaches to extract knowledge from larger data sets may offer insights into sincerity. In a similar way, given the success of data augmentation in the vision community [39] this approach may also be beneficial, due to the scarce sincerity data available. Likewise, another approach for data expansion is to crawl related YouTube data in regards to sincerity and authenticity. As mentioned previously, the sincerity of content providers is a prominent topic, offering an abundance of in-the-wild data sources.

6. Acknowledgements

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B) We would also like to thank the Columbia Computer Music Center for their contribution to the work.

7. References

[1] Y. An, "Western sincerity and confucian cheng," *Asian Philosophy*, vol. 14, no. 2, pp. 155–169, 2004.

[2] L. Trilling, *Sincerity and authenticity*. Boston, USA: Harvard University Press, 2009.

[3] H. Read, "The cult of sincerity," *The Hudson Review*, vol. 21, no. 1, pp. 53–74, 1968.

[4] S. Varga, "Authenticity," *The Encyclopedia of Political Thought*, pp. 215–225, 2014.

[5] J. R. Beniger, "Personalization of mass media and the growth of pseudo-community," *Communication research*, vol. 14, no. 3, pp. 352–371, 1987.

[6] S. R. Strong, "Counseling: An interpersonal influence process." *Journal of Counseling Psychology*, vol. 15, no. 3, p. 215, 1968.

[7] S. Neale, L. Spencer-Arnell, and L. Wilson, *Emotional intelligence coaching: Improving performance for leaders, coaches and the individual*, London, UK, 2011.

[8] J. B. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, and L. Michaelis, "Distinguishing deceptive from non-deceptive speech," in *Proc. INTERSPEECH 2005*, 2005, pp. 1833–1836.

[9] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proc. of International Conference on Multimodal Interaction*, 2015, pp. 59–66.

[10] S. Amiriparian, J. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. W. Schuller, "Is deception emotional? an emotion-driven predictive approach." in *Proc. INTERSPEECH 2016*, 2016, pp. 2011–2015.

[11] S. Mhamane and L. Lobo, "Fraud detection in online banking using hmm," in *Proc. of Computer Science & Information Technology*, 2012, pp. 200–204.

[12] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, "Sequence classification for credit-card fraud detection," *Expert Systems with Applications*, vol. 100, pp. 234–245, 2018.

[13] R. J. Voogt and W. E. Saris, "Mixed mode designs: Finding the balance between nonresponse bias and mode effects," *Journal of official statistics*, vol. 21, no. 3, p. 367, 2005.

[14] J. G. Knight, D. Mather, and B. Mathieson, *The key role of sincerity in restoring trust in a brand with a corporate apology. In Marketing Dynamism & Sustainability: Things Change, Things Stay the Same.* Springer International Publishing, 2015.

[15] J. Park, M. Cha, H. Kim, and J. Jeong, "Managing bad news in social media: A case study on dominos pizza crisis," in *Proc. of Conference on Weblogs and Social Media*, p. no pagination.

[16] J. K. Sandlin and M. L. Gracyalny, "Seeking sincerity, finding forgiveness: Youtube apologies as image repair," *Public Relations Review*, vol. 44, no. 3, pp. 393 – 406, 2018.

[17] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language." in *Proc. of INTERSPEECH*, 2016, pp. 2001–2005.

[18] Y. Zhang, F. Weninger, Z. Ren, and B. W. Schuller, "Sincerity and deception in speech: Two sides of the same coin? a transfer-and multi-task learning perspective." in *Proc. INTERSPEECH 2016*, 2016, pp. 2041–2045.

[19] G. Gosztolya, T. Grósz, G. Szaszák, and L. Tóth, "Estimating the sincerity of apologies in speech by dnn rank learning and prosodic analysis," in *Proc. INTERSPEECH 2016*, 2016, pp. 2026–2030.

[20] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*, 1st ed. Basel, Switzerland: Springer, 2016.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[22] F. Eyben, F. Weninger, E. Marchi, and B. Schuller, "Likability of human voices: A feature analysis and a neural network regression approach to automatic likability estimation," in *Proc. of International Workshop on Image Analysis for Multimedia Interactive Services*, 2013, pp. 1–4.

[23] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. of INTERSPEECH*, 2017, pp. 3512 – 3516.

[24] A. Baird and E. Coutinho, "The sincere apology corpus (sina-c) [data set]," *Zenodo*, 2019.

[25] D. Kramer-Moore and M. Moore, "Pardon me for breathing: Seven types of apology," *ETC: A Review of General Semantics*, p. 160, 2003.

[26] G. F. Bachman and L. K. Guerrero, "Forgiveness, apology, and communicative responses to hurtful events," *Communication Reports*, vol. 19, no. 1, pp. 45–56, 2006.

[27] T. E. Basford, L. R. Offermann, and T. S. Behrend, "Please accept my sincerest apologies: Examining follower reactions to leader apology," *Journal of Business Ethics*, vol. 119, no. 1, pp. 99–117, 2014.

[28] L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, and T. Lenaerts, "Apology and forgiveness evolve to resolve failures in cooperative agreements," *Scientific reports*, vol. 5, p. 10639, 2015.

[29] B. N. Waller, "Sincere apology without moral responsibility," *Social Theory and Practice*, vol. 33, no. 3, pp. 441–465, 2007.

[30] L. Akehurst, G. Köhnken, A. Vrij, and R. Bull, "Lay persons' and police officers' beliefs regarding deceptive behaviour," *Applied Cognitive Psychology*, vol. 10, no. 6, pp. 461–471, 1996.

[31] J. P. Buckley, *The Reid technique of interviewing and interrogation*. John E. Reid & Associates, Inc, 2000.

[32] L. Mnatsakanyan, "Falsehood in speech and some means of its expression," *Armenian Folia Anglistika*, pp. 79–84, 2012.

[33] N. Cummins, A. Baird, and B. W. Schuller, "The increasing impact of deep learning on speech analysis for health: Challenges and Opportunities," *Methods, Special Issue on on Translational data analytics and health informatics*, vol. 151, pp. 41–54, 2018.

[34] G. Keren and B. Schuller, "Convolutional RNN: an enhanced model for extracting features from sequential data," in *Proc. of 2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, Canada, 2016, pp. 3412–3419.

[35] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc the International Conference on Multimedia*, 2013, pp. 835–838.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, pp. 2825–2830, 2011.

[38] S. Amiriparian, S. Julka, N. Cummins, and B. Schuller, "Deep convolutional recurrent neural network for rare acoustic event detection," in *Proc. of the Challenge for Detection and Classification of Acoustic Scenes and Events*, 2018, p. no pagination.

[39] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Adaptive data augmentation for image classification," in *Proc. of International Conference on Image Processing*, 2016, pp. 3688–3692.