Automated Recognition of Sincere Apologies from Acoustics of Speech

Zafi Sherhan Syed[1]
Mehran University,
Pakistan

Muhammad Shehram Shah[2]
RMIT University,
Australia

Abbas Shah Syed[3]
University of Louisville,
USA

***Abstract*—Sincerity is an important characteristic of communicative behavior which represents an honest, truthful, and genuine display of verbal and non-verbal expressions. Individuals who are deemed sincere often appear more charismatic and can influence a large number of people. In this paper, we propose a multi-model fusion framework to identify sincerely delivered apologies by modelling difference between acoustics of sincere and insincere utterances. The efficacy of this framework is benchmarked using the Sincere Apology Corpus (SAC). We show that our proposed methods can improve the baseline classification performance (in terms of unweighted average recall) for SAC from 66.02% to 70.97% for the validation partition and 66.61% to 75.49% for the test partition. Moreover, as part of our investigation, we found that gender dependency can influence the classification performance of machine learning models, with models trained for male subjects performing better than those trained for female subjects.**

***Keywords*—Sincerity; affective computing; social signal processing**

I. Introduction

Sincerity is an act of being sincere. It is a quality of human beings that makes them free from pretense, deceit, and hypocrisy. Generally, it is believed that if a person is perceived to be sincere, more people will trust them. Sincerity and trust are at the heart of social interactions, be it in the form of relationships about business or personal life.

Sincerity is an important aspect of human behavior which is useful for many different day-to-day activities. For example, sincerity is a vital part of business dealing and along with honesty is considered to be one of its core values. Similarly, the perceived sincerity of public representatives, such as politicians, can significantly improve their chances of winning elections. Sincerity is also an important factor in healthcare where the trust needs to be established between clinicians and their patients. Finally, sincerity and truthfulness are important aspects of law prosecution where it needs to be ensured that the witness is being truthful in the court of law. To summarize, sincerity is an important aspect of human behavior that affects almost every part of society.

Recent advances in social signal processing have encouraged researchers to investigate aspects of human behavior that influence major sectors such as health and commerce. While sincerity recognition has not been investigated by the social signal processing research community in great detail, one can note that research into deception recognition has been a popular field of research [1], [2], [3], [4], [5], [6], [7], [8].

At the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH) held in late 2019, Baird et al. [9] published a relatively large corpus of speech recordings called the *Sincerity Apology Corpus (SAC)* which have been labeled explicitly for the task of sincerity recognition. To the best of our knowledge, this is the largest publicly available dataset in this field and therefore it provides researchers an opportunity to develop frameworks to recognize whether speech is perceived sincere or insincere.

In addition to releasing the dataset, Baird et al. [9] also investigated the efficacy of three types of audio features for the task of sincerity detection from speech. These features include, a) Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [10], b) Computational Paralinguistics Challenge (ComParE) feature set [11], and c) DeepSpectrum features [12]. While eGeMAPS and ComParE features are traditional non-deep learning based features, the DeepSpectrum features are generated by feeding audio signals into the AlexNet network for large scale image classification [13]. They reported classification performance in terms of unweighted average recall (UAR) for the three types of features which showed that the Deep Spectrum achieved 79.2% on the test partition of the cross-validation folds, whereas eGeMAPS and ComParE features achieved a UAR of 72.0% and 76.2%, respectively. As per Baird et al., these results are meant to serve as the baseline classification performance for further research in the field of sincerity detection, in particular research based on the SAC corpus. It is important to mention here that a part of the Sincerity Apology Corpus was also used for Computational Paralinguistics Challenge of 2016 [11]. There, the objective was to train machine learning models for a regression task with the objective to predict the sincerity score allocated to each audio recording by a group 16 annotators, whereas Baird et al. focus on a classification task to differentiate between sincere and insincere apologies.

In this paper, we propose a multi-model fusion framework for automated recognition of sincere apologies from acoustics of speech and test it using the Sincere Apology Corpus, a dataset publicly available for academic research. In addition to this, we investigate the influence of gender on the classification performance of machine learning models for the task at hand. The rest of this paper is organized as follows: In section II we provide a summary of the Sincere Apology Corpus, whereas in section III we detail the methodology for feature aggregation and model fusion methods. Experimental results and discussion is provided in section IV, and conclusion is provided in section V.

II. Dataset

Audio recordings in the Sincere Apology Corpus are provided as dual-channel stereo audio files that are sampled at 44100 Hz. As per the convention of the field, we first converted the audio recordings into a mono-channel by taking the average signal value per sample for the two channels of the stereo signal. Next, the signal is downsampled to a sampling frequency of 16000 Hz using the Librosa library [14]. Finally, each audio recording is normalized such that the dynamic range of the signal lies between -1 and $+1$. Whereas details of the Sincere Apology Corpus are available in [15], a summary of dataset partitions is provided in Table I.

TABLE I. Summary of dataset partitions for gender independent and gender dependent settings

Gender Independent			
	NS	**S**	**Total**
Train	143	142	285
Val	186	184	370
Test	105	151	256
Male			
	NS	**S**	**Total**
Train	98	45	143
Val	59	61	120
Test	77	80	157
Female			
	NS	**S**	**Total**
Train	45	97	142
Val	127	123	250
Test	28	71	99

III. Methodology

In Fig. 1, we illustrate the process flow pipeline of our proposed multi-model fusion framework for automated recognition of sincere apologies. Here, one starts with speech based audio recordings of subjects which are preprocessed into a standard format as discussed in the previous section. The next step is to compute acoustic low-level descriptors (LLDs) which quantify characteristics of speech paralinguistics. In the current work, we use the IS10Paralinguistics feature set, the ComParE feature set and the eGeMAPS feature set. These LLDs need to undergo a process of feature aggregation which yields a higher level representation of speech acoustics. To this end, we use functionals, bag of audio words, and Fisher Vector encoding based feature aggregation. Next, machine learning models are trained using the training partition, their hyper-parameters are optimized using the validation, and the performance of machine learning models is compared against one-another in an unbiased manner using the test partition. Finally, model fusion approaches are used with the aim to improve the classification performance of the sincerity recognition framework.

A. Acoustic Features for Speech Paralinguistics

The speech signal is inherently non-stationary in nature and therefore acoustic features need to be computed over short intervals of time over which the speech signal demonstrates some form of stationarity. In speech signal processing, it is common to compute acoustic features over time intervals in the range of $15 - 30$ ms [16], and as a result, such features are called low-level descriptors (LLDs).

These LLDs quantify various acoustic characteristics of a speech signal such as its fundamental frequency (also known as pitch), the quality of voice, spectral characteristics, and more. Given that elementary discussion on the characteristics of acoustic features is beyond the scope of this paper, we refer the reader to [10] for further discussion. Typically, these acoustic LLDs are packaged together and used in the form of feature sets. In our work, we shall use feature sets that have been shown to be useful for a variety of tasks related to speech paralinguistics.

In the baseline paper for the SAC corpus, Baird et al. [15] had used four fundamental types of feature sets. These include two domain-knowledge based feature sets: the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and the Computational Paralinguistics Challenge (ComParE), as well as acoustic representations derived from the AlexNet deep neural network for image classification [13], [12]. In our work, we shall make use of eGeMAPS and ComParE feature sets (similar to the baseline paper) but also include the IS10-Paralinguistics feature set which we have previously found to be useful for tasks related to speech paralinguistics [17], [18]. We shall provide a brief description of these features in the following paragraphs.

The IS10-Paralinguistics feature set consists of 38 acoustic LLDs which include 31 LLDs that describe spectral characteristics of speech, 6 LLDs which describe voicing related characteristics, and an LLD to describe the energy of voice (in terms of loudness). As the name suggests, the IS10-Paralinguistics feature set was especially designed to characterize paralinguistics characteristics of speech. For further details of this feature set, we refer the reader to [19]. Meanwhile, the ComParE feature set consists of 65 acoustic LLDs of which 55 LLDs describe the spectral characteristics of the speech signal, 6 LLDs quantify voicing related characteristics, and 4 LLDs describe energy-related LLDs. The ComParE feature set is often called a brute force feature set since it provides a more holistic approach to modeling speech characteristics. Finally, the eGeMAPS feature set was proposed as an optimized version of the ComParE feature set in terms of feature set dimensionality. The eGeMAPS feature set consists of 23 acoustic LLDs which include 9 spectral LLDs, 13 voicing related LLDs, and 1 energy-related LLDs. For further details of this feature set, we refer the reader to [10].

B. Feature Aggregation: Functionals

As mentioned earlier, due to the non-stationary nature of speech, acoustic features are computed for short duration frames of the audio signal (typically in the range of $20 - 30$ ms). These acoustic features, called low-level descriptors (LLDs) only provide low-level information. Therefore, in order to generate a global representation for an audio recording, the information provided by acoustic LLDs needs to be aggregated by appropriate methods. The simplest and commonly used feature aggregation method uses functionals of descriptive statistics such as mean, variance, range et cetera. In this work, we use a set of standard functionals as defined in the openSmile toolkit [20]. The toolkit is the defacto standard in the field of social signal processing due to its open-source nature and free availability for academic research.

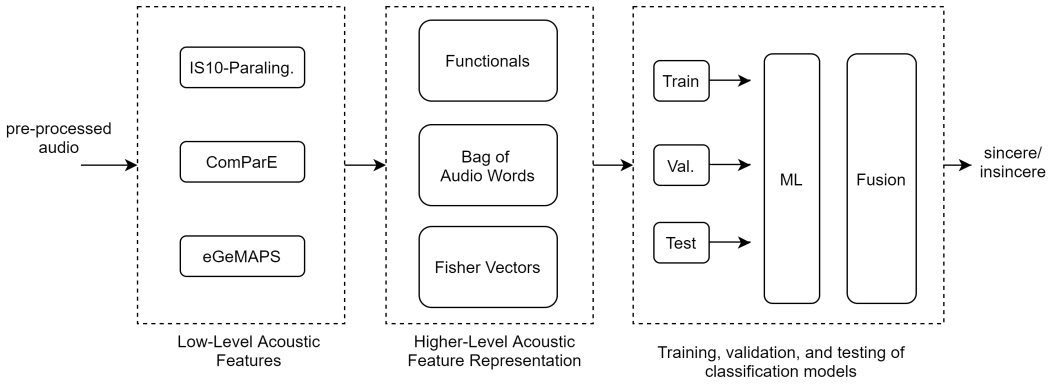

Fig. 1. Illustration of the Pipeline for Baseline Classification

C. Feature Aggregation: Bag of Audio Words

An alternate to functionals based feature aggregation is the Bag-of-Audio-Words (BoAW) method which is an extension of the bag-of-words (BoW) method from the field of natural language processing. BoW has been a popular approach to generate word-frequency histogram-based representation of text documents for applications related to text processing. The same concept has also been extended in audio signal processing to yield a global histogram-based representation for audio recordings [21], [22]. Unlike the text domain where textual words naturally exist, one needs to compute *audio words* through a process called vector quantization. Here, acoustic LLDs from all audio files are concatenated into a matrix and a clustering algorithm is used to learn representative clusters for the LLDs. Each cluster is called an audio word and the set of clusters for acoustic LLDs is called the codebook. It was common to use the k-means clustering algorithm, however Rawat et al. discovered that clustering based on random selection performs just as well with the advantage of a considerably smaller computational complexity [23]. Henceforth, it has become common practice to use random sampling for learning the codebook. As a result, we shall also use a random sampling approach for clustering in this work. The BoAW approach requires tuning of hyperparameters such as the codebook size, the number of simultaneous assignments to multiple audio words (in case an acoustic LLD is close in terms of Euclidean distance to multiple audio words), and normalization. We shall perform hyperparameter optimization for BoAW using the validation partition and from there select the best performing model for predicting test partition labels. In order to compute BoAW, we use the openXBOW toolkit [24].

D. Feature Aggregation: Fisher Vectors

Fisher vector encoding is a feature aggregation method which was initially proposed by Perronin et al. [25], [26] for applications in the field of computer vision, achieving state-of-the-art performance [27] for object recognition before the advent of deep learning for computer vision era [13]. This method has also been found useful for applications related to speech paralingusitics [28], [29], [30]. In our previous works, we found the Fisher Vector based feature aggregation to be useful for speech screening of depression [31] and bipolar disorder [17]. Fisher Vector representation combines the advantages of generative models i.e. the ability to work with variable-length data and discriminative models i.e. ability to learn class-specific boundaries.

In order to compute Fisher Vector features for the task of sincerity recognition, we follow the approach detailed by Perronin et al. [25], [26] but adapt it for audio signals. To this end, we first concatenate acoustic LLDs from all audio recordings and train a Gaussian Mixture Model (GMM) [32], which serves as the generative model of the Fisher Vector framework. Next, the first and second-order statistics for gradients between acoustic LLDs from an audio recording and the generative model are computed. These statistics are then concatenated to yield a single feature vector and called Fisher Vector (FV) features. In the current work, we use the VLfeat toolkit [33] in order to train GMMs as well as compute FV features.

E. Fusion

Fusion is a method through which it is possible to combine information from multiple machine learning models with the aim to improve overall classification performance[34]. There are two fundamental ways to fuse such information i.e. label fusion and confidence fusion. In label fusion, class-label predictions from machine learning models are stacked and the class which is predicted by the majority of models is deemed to be the correctly predicted class. For example, if three models in five-model fusion predict the label for a speech recording to be *Sincere* whereas two models predict the label as *Not-Sincere* then the final label will be decided as *Sincere* based on a majority vote. Meanwhile, confidence fusion takes place on probabilistic or confidence outputs of the classifier. It is reminded that each classification model returns a confidence metric for the predicted label which essentially quantifies how sure it is about its predictions. Naturally, it predicts the label for which it has the most confidence. Given that different models are trained with different features and parameters, the confidence may be different and such information may help to improve the overall accuracy of prediction. In confidence fusion, the idea is to decide on the class-label based on the confidence of multiple machine learning models. The simplest way to implement confidence fusion is to take the arithmetic average of the confidence metric from multiple machine learning models and predict the label which has more confidence.

IV. Experimentation, Results and Discussion

We use the implementations of logistic regression (LR), support vector machine (SVM), and random forest (RF) classifier which are available in the scikit-learn toolkit [1]. The complexity value of the LR and SVM is optimized over a logarithmically spaced grid between 10^{-7} to 10^{7}. The RF classifier has a number of hyperparameters which need to be optimized, such as a) the number of trees in the forest (num_est), b) the maximum depth of each tree (*max_depth*), c) the minimum number of samples before splitting at a node (*min_samps_split*), and d) the minimum number of samples required to be at a leaf node (*min_samps_leaf*). To this end, we conduct gridsearch based optimization with the following parameter values: *num_est* = $\{25, 50, 100, 200, 400\}$, *max_depth* = $\{2, 5, 10, 15, 20\}$, *min_samps_split* = $\{2, 5, 10, 15, 20\}$, and *min_samps_leaf* = $\{2, 5, 10, 15, 20\}$. These classifiers are trained using the training partition, their hyperparameters are optimized using the validation partition, and the classification results being compared against the test partition. For the sake of completeness, we report the results for both validation and test partitions.

A. Baseline Classification Performance

In Table II, we provide a summary of the baseline classification performance for the Sincere Apology Corpus which was reported by Baird et al. [15]. Furthermore, we also report the results we achieved using the same feature set (note that along with audio recordings, Baird et al. also provided their features) albeit with three different classifiers, that are LR, SVM, and RF whereas Baird et al. only used SVM.

Ideally, one expects that the results provided in the baseline paper and those computed by us with the SVM classifier would be the same given that features and the SVM classifier are similar; but these are not. In fact, we find that all results of machine learning models reported by Baird et al. achieve a greater classification accuracy on the test partition as compared to the results we computed. One can think of two possible reasons: 1) the random seed and the number of iterations for training the SVM classifier could be different between our implementation and that of [15] which can lead to a difference in results, and 2) Baird et al. optimized the SVM for results on the test partition directly whereas we optimized the SVM classifier for the validation partition and only used the test partition for comparing classification performance across different models. Furthermore, one should also note that Baird et al. did not report results for the validation partition which would have otherwise made their results easier to interpret.

Given this, we shall use classification performance achieved through our experiments as the baseline and shall carry out our investigations on feature aggregation, gender dependency, and model fusion for binary classification between sincere and insincere apologies. To this end, we report that eGeMAPS functionals when used with the SVM classifier, provide the best classification performance of UAR = 66.20% for the development partition, and the corresponding model achieves a UAR = 66.61% for the test partition. This shall be the baseline classification performance.

[1] https://scikit-learn.org

B. Experiments with Gender Independent Partitions

We first compare the classification performance of feature aggregation methods in a gender-independent setting. In Table III, we provide a summary of results for classification between sincere and insincere apologies for functionals, BoAW, and FV based feature aggregation of IS10-Paralinguistics, eGeMAPS, and ComParE features for a gender independent setting. It is clear to note that the baseline UAR = 66.02% for the development partition can be improved by all three feature aggregation methods. The top performing models from each method are Funcs-IS10Paraling-RF, BoAW-ComParE-RF, and FV-ComParE-RF. Overall, the best performing model for the validation partition is BoAW-ComParE-RF with a UAR = 67.08% which goes on to achieve a UAR = 68.78% on the test partition.

C. Experiments with Gender Dependent Partitions

It is known that the subject's gender can influence the paralinguistic characteristics of speech for applications such as emotion valance recognition [35] and depression recognition [36], [37]. We, therefore, investigate the effect of gender on the accuracy with which machine learning models can differentiate between sincere and insincere apologies. To this end, we conducted experiments as discussed previously with gender-dependent partitions and provide a summary of results for male gender in Table IV and female gender in Table V. It is important to mention here that since Baird et al. did not report results of classification performance under a gender-dependent setting, therefore, a baseline does not exist, and we shall introduce a baseline for gender-dependent settings as part of our work.

From Table IV, we note that the best performance for the validation partition is achieved by Funcs-eGeMAPS-RF with a UAR = 80.79% although the performance drops significantly to 55.52% on the test partition. It is interesting to note that a number of models achieve similar performance as Funcs-eGeMAPS-RF on the validation partition, such as BoAW-IS10Paraling-RF with UAR = 80.02%, FV-ComParE-SVM with UAR = 79.91%, FV-ComParE-LR with UAR = 79.04%, and FV-IS10Paraling-SVM with UAR = 79.01%. Amongst these, FV-IS10Paraling-SVM achieves the highest performance on the test partition with a UAR = 69.34% whereas FV-ComParE-SVM and FV-ComParE-LR achieve a UAR of approximately 67.50%. These results suggest overfitting on the validation partition since there exists a large difference between the UAR values achieved for validation and test partition.

As far as recognition of sincere and insincere apologies for female subjects is concerned, we find somewhat poorer classification performance of machine learning models as compared to the case of male subjects. Here, the best performing model FV-IS10Paralig-LR achieves a UAR = 67.89% on the validation partition and a UAR = 72.33% on the test partition. The best performing model amongst BoAW based feature aggregation is the BoAW-IS10Paraling-LR which achieved a UAR = 64.88% for the validation partition whereas the best performing with functionals based aggregation achieved a UAR = 64.82% for the same partition. This suggests that gender does influence the paralinguistic characteristics of sincere and insincere speech.

TABLE II. SUMMARY OF RESULTS PROVIDED AS BASELINE IN [15] AND FROM EXPERIMENT PERFORMED BY THESE AUTHORS USING BASELINE FEATURES

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Results provided	ComParE-funcs	-	70.00	-	-	-	-
	eGeMAPS-funcs	-	70.20	-	-	-	-
	DeepSpectrum-MelSpec-fc6	-	69.80	-	-	-	-
	DeepSpectrum-LinSpec-fc6	-	69.90	-	-	-	-
	DeepSpectrum-MelSpec-fc7	-	65.90	-	-	-	-
	DeepSpectrum-LinSpec-fc7	-	68.60	-	-	-	-
Our experiments	ComParE-funcs	62.07	66.40	64.46	64.50	**65.24**	62.80
	eGeMAPS-funcs	**66.02**	66.61	63.62	63.75	63.07	63.90
	DeepSpectrum-MelSpec-fc6	51.87	53.95	58.17	66.05	57.88	66.34
	DeepSpectrum-LinSpec-fc6	57.23	57.76	61.65	65.16	58.98	65.74
	DeepSpectrum-MelSpec-fc7	59.85	65.80	60.11	65.43	58.46	64.64
	DeepSpectrum-LinSpec-fc7	53.48	53.00	58.20	66.05	**59.81**	65.30

TABLE III. SUMMARY OF RESULTS FOR FUNCTIONALS, BAG-OF-AUDIO WORDS, AND FISHER VECTOR FEATURES FOR GENDER INDEPENDENT SETTING OF TRAINING, VALIDATION, AND TEST PARTITIONS

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Functionals	IS10Paral.	64.46	70.75	64.42	68.97	**66.84**	70.01
	ComParE	62.84	66.87	65.23	65.37	64.71	64.54
	eGeMAPS	63.02	59.47	62.57	66.34	**66.05**	63.13
BoAW	IS10Paral.	**66.00**	62.68	65.73	65.00	64.67	65.29
	ComParE	61.59	56.01	62.98	62.18	**67.08**	68.78
	eGeMAPS	59.48	64.11	59.19	67.48	64.66	60.96
FV	IS10Paral.	62.72	71.87	63.78	70.01	63.33	67.50
	ComParE	65.73	69.39	65.45	69.82	**66.60**	69.40
	eGeMAPS	63.82	62.70	64.08	65.16	64.17	68.45

TABLE IV. SUMMARY OF RESULTS FOR FUNCTIONALS, BAG-OF-AUDIO WORDS, AND FISHER VECTOR FEATURES FOR TRAINING, VALIDATION, AND TEST PARTITIONS WITH MALE SUBJECTS ONLY

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Functionals	IS10Paraling	72.06	69.72	71.33	71.14	74.91	52.32
	ComParE	73.81	69.05	73.02	66.50	72.05	57.23
	eGeMAPS	68.78	60.78	75.62	64.15	**80.79**	55.74
BoAW	IS10Paraling	76.55	66.82	78.19	66.70	**80.02**	60.76
	ComParE	69.03	62.32	68.30	57.32	77.42	66.19
	eGeMAPS	71.77	62.35	67.81	55.06	75.90	50.64
FV	IS10Paraling	**79.01**	69.34	79.01	67.42	73.63	54.16
	ComParE	**79.91**	67.54	**79.04**	67.47	76.09	67.18
	eGeMAPS	75.70	59.82	78.27	61.31	74.34	62.09

TABLE V. SUMMARY OF RESULTS FOR FUNCTIONALS, BAG-OF-AUDIO WORDS, AND FISHER VECTOR FEATURES FOR TRAINING, VALIDATION, AND TEST PARTITIONS WITH FEMALE SUBJECTS ONLY

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Functionals	IS10Paraling	59.97	59.68	60.43	58.98	59.32	64.71
	ComParE	56.64	58.60	56.98	67.91	58.22	67.91
	eGeMAPS	**64.82**	50.48	**65.21**	56.54	62.68	64.34
BoAW	IS10Paraling	64.22	69.52	**64.88**	64.81	**64.86**	64.76
	ComParE	63.66	68.86	63.38	62.75	62.33	68.66
	eGeMAPS	61.83	50.96	62.61	54.15	64.31	68.66
FV	IS10Paraling	60.65	63.63	**67.89**	72.33	63.17	64.71
	ComParE	61.61	72.56	60.50	62.93	60.88	66.83
	eGeMAPS	63.43	58.32	63.78	62.22	61.79	58.27

D. Model Fusion

Finally, in Table VI, we provide a summary of results for label- and confidence-based fusion for predicting sincerity for the gender independent setting. Here, we chose to fuse the results from top-5 performing models. The results show that both fusion approaches can help improve the classification performance for validation as well as test partitions. Interestingly, there is little difference between the UAR achieved by label-based and confidence-based fusion approaches for the validation partitions but for the test partition, label-based fusion provides a better UAR with 75.49% compared to 73.22% as achieved by confidence-based fusion.

TABLE VI. SUMMARY OF RESULTS FOR LABEL- AND CONFIDENCE-BASED FOR TOP-5 PERFORMING MODELS

Model Name	**UAR (%)**	
	Val.	*Test*
BoAW-ComParE-RF	**67.08**	68.78
Funcs-IS10Paraling-RF	66.84	**70.01**
FV-ComParE-RF	66.60	69.40
Funcs-eGeMAPS-RF	66.05	63.13
BoAW-IS10Paralig-LinSVM	66.00	62.68
Label Fusion	70.79	**75.49**
Conf. Fusion	**70.97**	73.22

V. CONCLUSION

The purpose of the current study was to propose a multi-model fusion based framework for identifying speech recordings which carry insincere apologies amongst a corpus which also contains recordings of sincere apologies. To this end, our proposed methods were able to improve the classification performance for the Sincere Apology Corpus from 66.02% to 70.97% for the validation partition and 66.61% to 75.49% for the test partition. We also proposed new baselines for gender dependent classification between sincere and insincere apologies and report that classification models tend to perform better for male subjects as compared to female subjects.

REFERENCES

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

[2] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2006, pp. 1033–1036.

[3] J. F. Torres, E. Moore, and E. Bryant, "A study of glottal waveform features for deceptive speech classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4489–4492.

[4] N. Raiman, H. Hung, and G. Englebienne, "Move, and I will tell you who you are: Detecting deceptive roles in low-quality data," in *ACM International Conference on Multimodal Interaction*, 2011, pp. 201–204.

[5] C. Fan, H. Zhao, X. Chen, X. Fan, and S. Chen, "Distinguishing deception from non-deception in Chinese speech," in *International Conference on Intelligent Control and Information Processing (ICICIP)*, 2015, pp. 268–273.

[6] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," in *WMDD 2015 - Proceedings of the ACM Workshop on Multimodal Deception Detection, co-located with ICMI 2015*, 2015.

[7] C. Montacie and M. J. Caraty, "Prosodic cues and answer type detection for the deception sub-challenge," in *INTERSPEECH*, 2016, pp. 2016–2020.

[8] G. Mendels, S. I. Levitan, K. Z. Lee, and J. Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection," in *INTERSPEECH*, 2017, pp. 1472–1476.

[9] A. Baird, S. Amiriparian, N. Cummins, S. Sturmbauer, J. Janson, E.-M. Messner, H. Baumeister, N. Rohleder, and B. Schuller, "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," in *INTERSPEECH*, 2019, pp. 534–538.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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.

[11] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native language," in *INTERSPEECH*, 2016, pp. 2001–2005.

[12] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *ACM on Multimedia Conference*, 2017, pp. 478–484.

[13] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1–9.

[14] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "c," in *14th Python in Science Conference*, 2015, pp. 18–24.

[15] A. Baird, E. Coutinho, J. Hirschberg, and B. Schuller, "Sincerity in Acted Speech: Presenting the Sincere Apology Corpus and Results," in *INTERSPEECH*, 2019, pp. 539–543.

[16] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis*, 1st ed. Academic Press Inc. (London) Ltd., 2014.

[17] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 39–45.

[18] Z. S. Syed, S. A. Memon, M. S. Shah, and A. S. Syed, "Introducing the Urdu-Sindhi Speech Emotion Corpus: A novel dataset of speech recordings for emotion recognition for two low-resource languages," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 1–6, 2020.

[19] S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, S. Language, P. Group, D. Telekom, and A. G. Laboratories, "The INTERSPEECH 2010 Paralinguistic Challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

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

[21] Y. Liu, W. L. Zhao, C. W. Ngo, C. S. Xu, and H. Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *ACM International Conference on Image and Video Retrieval*, Xi'an, China, 2010, pp. 89–96.

[22] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words Approach for Multimedia Event Classification." in *INTERSPEECH*, Portland, Oregon, USA, 2012, pp. 2105–2108.

[23] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust Audio-Codebooks for Large-Scale Event Detection in Consumer Videos," in *INTERSPEECH*, 2013, pp. 2929–2933.

[24] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[25] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[26] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Lecture Notes in Computer Science*, vol. 6314, 2010, pp. 143–156.

[27] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[28] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression Estimation Using Audiovisual Features and Fisher Vector Encoding," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 87–91.

[29] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 255–259.

[30] H. Kaya, F. Gurpinar, S. Afshar, and A. A. Salah, "Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 459–466.

[31] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 37–43.

[32] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[33] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," in *ACM International Conference on Multimedia*, 2010, pp. 1469–1472.

[34] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, no. 1, pp. 115–129, 2020.

[35] H. Sagha, J. Deng, and B. Schuller, "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition," in *ACM International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 1–5.

[36] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, P. Le, V. Sethu, and J. Epps, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 19–26.

[37] G. Stratou, S. Scherer, J. Gratch, and L. P. Morency, "Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences," in *IEEE Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 147–152.