A COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS FOR DETECTING DEPRESSION FROM SPONTANEOUS SPEECH

Sharifa Alghowinem[1,5], *Roland Goecke*[2,1], *Michael Wagner*[2], *Julien Epps*[3], *Tom Gedeon*[1], *Michael Breakspear*[4,3], *Gordon Parker*[3]

[1]Australian National University, Canberra, Australia
[2]University of Canberra, Canberra, Australia
[3]University of New South Wales, Sydney, Australia
[4]Queensland Institute of Medical Research, Brisbane, Australia
[5]Ministry of Higher Education: Kingdom of Saudi Arabia

sharifa.alghowinem@anu.edu.au, roland.goecke@ieee.org, michael.wagner@canberra.edu.au, j.epps@unsw.edu.au, tom@cs.anu.edu.au, mjbreaks@gmail.com, g.parker@blackdog.org.au

ABSTRACT

Accurate detection of depression from spontaneous speech could lead to an objective diagnostic aid to assist clinicians to better diagnose depression. Little thought has been given so far to which classifier performs best for this task. In this study, using a 60-subject real-world clinically validated dataset, we compare three popular classifiers from the affective computing literature – Gaussian Mixture Models (GMM), Support Vector Machines (SVM) and Multilayer Perceptron neural networks (MLP) – as well as the recently proposed Hierarchical Fuzzy Signature (HFS) classifier. Among these, a hybrid classifier using GMM models and SVM gave the best overall classification results. Comparing feature, score, and decision fusion, score fusion performed better for GMM, HFS and MLP, while decision fusion worked best for SVM (both for raw data and GMM models). Feature fusion performed worse than other fusion methods in this study. We found that loudness, root mean square, and intensity were the voice features that performed best to detect depression in this dataset.

Index Terms— Mood detection, clinical depression, classifier comparison, affective sensing

1. INTRODUCTION

Clinical depression is a common mental disorder and disabling condition that impairs an individual's ability to cope with daily life. Major depression is the leading cause of disability and the cause of more than two-thirds of suicides each year [1]. Therefore, failure to diagnose depression in primary care is a critical public health problem that results in high societal costs related to disability, morbidity, mortality, and excessive health care utilisation [2]. Moreover, effective depression treatment is limited by current assessment methods that rely almost exclusively on patient-reported or clinical judgements of symptom severity [3], risking a range of subjective biases. We believe that affective sensing technology will play a major role in providing an objective assessment. Our goal here is to investigate the utility of various classifiers for the detection of depression. Ultimately, we want to develop an objective affective sensing system that supports clinicians in their diagnosis and monitoring of clinical depression.

The main contribution of this paper is a comparative study of classifier performance – three popular classifiers from the literature (Gaussian Mixture Models (GMM), Support Vector Machines (SVM), multilayer Perceptron neural networks (MLP)) and the relatively new Hierarchical Fuzzy Signature (HFS) classifier – for the task of accurately detecting depression. Beside comparing the classifier performances, we also investigate which features or group of features perform better for this task and compare different fusion methods, namely feature, score and decision fusion.

2. RELATED WORK

Psychology research of depressed speech found several distinguishable prosodic features, such as differences in the pitch, loudness, speaking rate, and articulation [3]. Moreover, the research found that formants are a feature significantly distinguishing depressed from non-depressed speech [4, 5], with a noticeable decrease in the second formant for depressed compared to healthy controls [4]. There is convincing evidence that sadness and depression are associated with a decrease in loudness and energy [6]. Jitter and shimmer voice features were analysed for depression, finding higher jitter in depression caused by the irregularity of the vocal fold vibrations [6] and lower shimmer for depressed subjects [7]. Like the jitter feature, the harmonic-to-noise (HNR) feature is higher for depressed, as the patterns of air flow in the speech production differ for depressed and control subjects [8].

In recent years, the automatic detection of depression using artificial intelligent techniques has been investigated, e.g. [9, 10, 11]. While psychology investigations are concerned with the overall patterns of speech using statistical measurements of speech prosody, affective sensing approaches rely on frame-by-frame low-level features extracted from the speech signal, which has been shown to perform well for several features. The first 3 formants features gave good classification results in [10], as well as energy and loudness in [11]. Pitch or F0 classification results were not as good as expected [10, 11], only if compared with recordings of the same person after treatment (speaker-dependent classification) [9].

To complicate matters further, the little work on automatic detection of depression from speech in the literature used different classifiers and different measures applied on different datasets, which make the comparison of results even harder; a general problem of most emotion recognition papers [12, 13]. Therefore, there is a need for a comprehensive comparison of classifiers using the same dataset and measurements to identify the strongest feature (or group of features) and the most suitable classifier for depression detection. In this paper, we perform a comparative study of the performance of 4 classifiers, using 12 individual voice features, and also compare fusing these features using feature, score and decision fusion.

3. REAL-WORLD CLINICALLY VALIDATED DATA

For the experiments, we used data collected in an ongoing study at the Black Dog Institute, a clinical research facility in Sydney (Australia), offering specialist expertise in depression and bipolar disorder. Subjects included healthy controls and patients who had been clinically diagnosed with severe depression (Hamilton Depression Rating Scale (HAM-D) $>$ 15). To date, data from more than 40 depressed subjects and 40 controls (both females and males) has been collected after obtaining informed consent from the participants in accordance with approval from the local institutional ethics committee. In this study, a gender-matched subset of 30 depressed subjects and 30 healthy controls was analysed [11].

The audio-video experimental paradigm contains several parts, including an interview with each subject. The interview was conducted by asking specific open questions to describe events that had aroused significant emotions. The interview was manually labelled to extract pure subject speech, where the total duration was 290min. Since the durations differ for each subject, which may affect the comparison results, we only used an equal amount of speech data from each subject (92s) in this paper. We acknowledge that the amount of data used here is relatively small, but this is a common problem [5, 14] in similar studies. As we continue to collect more data, future studies will be able to report on a larger dataset.

For feature extraction, voice features can be categorised into acoustic and linguistic features [15]. Acoustic features can also be categorised into low-level descriptors (LLD) and statistical functionals, which are calculated based on the LLD over certain units (e.g. words, syllables, sentences). Here, we used the publicly available openSMILE software [16] to extract several LLD features (for each frame) and some functional features (e.g. the deltas of the MFCC) as listed in Table 1. The frame size was set to 25ms at a shift of 10ms with a Hamming window, which gave 9200 frames per subject.

4. CLASSIFIERS

Automatic emotion recognition approaches have been using a variety of classifiers, both descriptive (generative) and discriminative, but it is not clear, which one performs best for the depression detection from speech. We compare 4 classifiers in a binary (i.e. depressed/non-depressed) speaker-independent scenario. To mitigate the effect of the limited amount of data, a leave-one-subject-out cross-validation was used in all the classifiers without any overlap between training and testing data [17]. To measure the performance, several statistical methods could be calculated [17]; we used the average recall.

4.1. Gaussian Mixture Models

GMM is a generative classifier widely used in speaker and speech recognition as well as in recognising emotions [13]. Its advantage is modelling low-level (frame based) features directly regardless of speech duration differences. GMM were trained using a continuous Hidden Markov Model (HMM) with a single state that used 16 weighted mixtures of Gaussian densities [17] using the HTK software [18]. However, the major disadvantages of GMMs are that they require intensive computations and parameter optimisation, as well as the unclear choice of the number of mixtures. In this work, diagonal covariance matrix was used, and the choice of the number of mixtures was fixed to ensure consistency in the comparison of classifiers, acknowledging that some features benefit more from more detailed modelling.

4.2. Support Vector Machines

SVM is a discriminative classifier, which has been widely used in speech, vision and many other classification tasks [13]. It is often considered the state-of-the-art classifier, since it provides good generalisation, although it may not be the best for every case [17]. When it comes to low-level speech features, the dimensionality of the SVM super-vector depends on the duration of speech, which might be problematic for unbalanced speech durations. It also causes a very high-dimensional feature vector. As we used equal amounts of speech data from each subject, full low-level (frame-based) features were used to build the SVM super-vector (9200 values for each feature dimension), as well as the means of the 16 mixtures of each subject's GMM model as the SVM super-vector (16 values for each feature dimension). In this study,

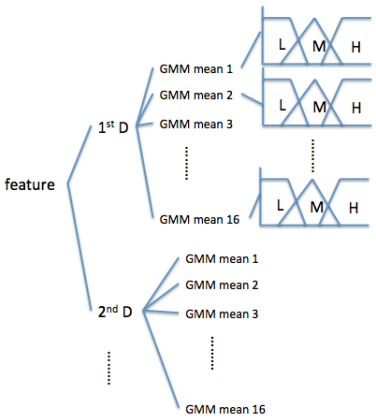

Fig. 1. Constructed Fuzzy Signature

using GMM models is a novel method for dimensionality reduction, enabling the use of hybrid generative and discriminative classifiers. To increase the accuracy of the results of SVMs, the cost and gamma parameters were optimised. We employed LibSVM [19] with a radial basis function kernel and a wide grid search range for the best parameters.

4.3. Hierarchical Fuzzy Signature

HFS [20] is a recent classifier, especially to the area of emotion recognition. It overcomes the limitation of fuzzy rules, by handling problems with complex structure and dealing with missing data. Fuzzy signatures can be considered as special, multidimensional fuzzy data. It compositions data into vectors of fuzzy values, each of which can be a further vector [20]. HFS have been successfully applied to a number of applications, such as cooperative robot communication, personnel selection models, Severe Acute Respiratory Syndrome pre-clinical diagnosis system, etc. [21, 22]. We hypothesize that, given the continuous range nature of emotions in general and the overlap between them, fuzzy systems might be suitable for the task. However, the choice of the membership function, the aggregation function, and the number of fuzzy sets are critical for getting accurate results. In this study, we adopted the HFS construction approach based on the Levenberg-Marquardt method ([23]). The fuzzy signature was constructed using the 16 means of each subject's GMM model mentioned earlier as the branches. Each branch represents a fuzzy value calculated using Fuzzy C Mean (FCM) clustering into three fuzzy sets (Low-Med-High) (Figure 1).

4.4. Multilayer Perceptron Neural Network

MLP are a special case of the artificial neural network, which has been used in a wide range of applications, including pattern recognition and emotion recognition. Typically, an MLP consists of an input layer, one or more hidden layers, and an output layer, where each layer consists of nodes (perceptrons) that are connected to the nodes of the next layer. MLP networks are usually used for modelling complex relationships between inputs and outputs or to find patterns in data. However, the network topology including the number of hidden layers, the number of perceptrons in each layer, the choice of the activation function and the training algorithm, is not trivial and complicates the model. Therefore, MLPs are vulnerable to overfitting, requiring large amounts of training data [17]. In this paper, we implemented an MLP using two hidden layers, with 16 and 4 perceptrons respectively, chosen empirically and kept fixed for all features to ensure consistency in the comparison. The input for the MLP were the 16 mixture means of each subject's GMM model mentioned earlier and the target output was the binary label of the classes (1 for Depressed, 0 for Control). To create the MLP, we used Levenberg-Marquardt as the training function, a hyperbolic tangent sigmoid as activation function for the hidden layers and the mean squared error as the cost function.

5. FUSION

To determine the features best suited to classify depression, we investigated the performance of individual features without normalisation. We also investigated fusing those individual features using different methods: feature, score, and decision fusion. Feature fusion is a commonly used in multimodal systems, e.g. in audio-video speech processing. The feature vectors from each individual system have to be normalised before fusion. Although all features were extracted from the same modality, we fused them in the same manner as if they were from different modalities. We first normalised individual features using percentile normalisation (range from 0 to 1), then fused and fed them to the classifiers. Score fusion was performed by merging the score results (likelihood ratio in the case of GMM) of the individual features classification, using a weighted sum. Finally, the decision fusion used a weighted majority voting of the classification results of each individual feature. The weights in the score and decision fusion were selected using a grid search for the best results.

6. EXPERIMENTS AND COMPARISON

The results of the comparison of classifiers (Table 1) show that the hybrid SVM with GMM outperformed the other classifiers in each single feature, with F0, HNR and formants in SVM with raw data being exceptions. Remarkably, F0, formants, jitter and shimmer had a high recognition rate using both SVMs (with raw data and GMM), but not with GMM due to the sparse data, indicating that GMM benefit from continuous data streams [24]. In contrast, SVM easily separates sparse data. Note that although SVM using raw data came in second, it may not be suitable for unbalanced speech duration. The HFS and MLP classifiers performed not as well, but different topologies and structures need to be investigated.

Table 1. Average recall (in %) for individual and group features for different classifiers and different fusion methods

		Classifier:	GMM	SVM raw data	SVM + GMM	HFS + GMM	MLP + GMM	
Feature Group	Feature	Feature Dimension	9200 × feature dimension	9200 × feature dimension	16 × feature dimension	16 × 3 Fuzzy sets × feature dimension	16 × feature dimension	**Feature Performance Average**
Pitch	F0	1	25.00	66.36	65.43	31.25	41.58	45.92
	Voicing Probability	1	55.14	62.82	70.09	48.29	55.01	58.27
	F0 quality	1	72.10	71.53	72.91	55.05	60.00	66.32
MFCC	MFCC	13	56.70	65.15	69.39	25.00	56.70	54.59
	MFCC, Deltas	39	62.00	63.39	75.25	25.00	56.67	56.46
Energy	Log	1	66.74	70.36	76.79	**72.29**	60.65	69.36
	RMS	1	72.88	69.39	80.54	66.48	**71.69**	**72.20**
Intensity	Intensity	1	**76.79**	**77.95**	78.62	62.78	60.72	**71.37**
	Loudness	1	69.48	75.12	**85.04**	68.86	68.52	**73.40**
Formants	3 formants	6	25.00	66.97	63.39	62.78	53.35	54.30
Voice Quality	Jitter	2	1.61	70.00	77.00	25.00	69.48	48.62
	Shimmer	1	1.61	70.00	73.29	25.00	66.34	47.25
	HNR	1	55.50	75.42	68.52	57.50	50.00	61.39
Average for Features			49.27	69.57	73.56	48.10	59.28	-
Feature Fusion: Percentile Normalisation	Pitch Group	3	25.00	63.48	68.86	48.33	56.70	52.47
	Energy Group	2	72.88	**71.53**	80.00	68.63	68.86	**72.38**
	Intensity Group	2	**77.78**	76.44	83.48	53.75	**73.33**	**72.96**
	Voice Quality Group	4	0.00	59.26	70.83	**81.25**	56.70	53.61
	All Group	30	65.63	67.94	73.33	25.00	63.33	59.05
Average for Feature Fusion			48.26	67.73	75.30	55.39	63.78	-
Score Fusion: Weighted Sum	Pitch Group	3	**75.86**	71.53	72.91	**56.79**	**65.15**	68.45
	Energy Group	2	**75.57**	**76.44**	**81.99**	72.29	**73.44**	**75.94**
	Intensity Group	2	**77.78**	77.95	85.04	**71.31**	**68.86**	**76.19**
	Voice Quality Group	3	55.49	70.00	77.00	**65.86**	69.48	67.57
	All Group	13	**78.85**	66.36	72.91	**75.71**	**76.79**	**74.12**
Average for Score Fusion			72.71	72.46	77.97	68.39	70.74	-
Decision Fusion: Weighted Majority Voting	Pitch Group	3	72.09	**72.09**	**73.81**	**55.78**	58.34	66.42
	Energy Group	2	72.88	70.36	80.54	72.29	**76.67**	**74.55**
	Intensity Group	2	76.79	**79.14**	85.04	68.86	65.15	**75.00**
	Voice Quality Group	3	55.49	75.42	77.00	57.50	69.48	66.98
	All Group	13	**78.85**	**84.88**	**91.67**	**73.81**	**76.44**	**81.13**
Average for Decision Fusion			71.22	76.38	81.61	65.65	69.22	-

To increase the accuracy, we compared unimodal fusion methods, such as feature, score, and decision fusion. As can be seen in Table 1, (where the bold numbers indicate improvements from using single features), on average fusing features at the feature level gave either similar or even less accurate results than fusing them at the score or decision level. As an exception, the intensity group feature fusion using MLP and the voice quality group feature fusion using HFS gave better results than score or decision fusion. On average, score fusion performed better than decision fusion with most classifiers but SVM, where decision fusion performed better in both raw data and GMM. Fusing all features in a decision level majority voting increased the results strongly. Both HFS and MLP score-level fusion outperformed their single features classification, indicating that better topology, structure and optimization would potentially lead to better recognition rate.

The experiments showed that loudness, RMS energy, and intensity were the strongest features for detecting depression, showing high results with every classifier used, even when fusing them as groups, which is in line with psychological findings that depression is associated with a decrease in loudness and energy [6]. Although depressed speech has a monotone characteristic indicated by changes in pitch and formants [4, 3, 25], their performances here were not as good as expected. That might be an indication that pitch and formants features are more suitable in a speaker-dependent comparison. The voice quality group performed relatively better than pitch and formants when using SVMs, which is in line with psychological conclusions that depressed speech is characterised by irregularities in the vocal fold vibrations [6, 7, 8].

7. CONCLUSIONS

We are interested in accurate detection of depression from spontaneous speech, which could lead to an objective diagnostic aid to assist clinicians. We compared the performance of various acoustic and prosodic features and classifiers for this task (GMM, SVM, MLP, and HFS) on a 60-subject real-world, clinically validated dataset. We also investigated the usage of a hybrid classifier, using GMM models as input to the other classifiers, which also reduced the dimensionality. Among the 4 classifiers, the hybrid classifier using GMM with SVM performed best overall. Amongst the fusion methods, score fusion performed better when combined with GMM, HFS and MLP classifiers, while decision fusion worked best for SVM (both for raw data and GMM models). Feature fusion exhibited weak performance compared to other fusion methods. Loudness, root mean square, and intensity were the strongest voice features to detect depression using the classifiers in this study. In future work, these findings will be investigated for generalisation across cultures (American, Saudi) and languages (Arabic).

8. REFERENCES

[1] US Department of Health and Human Services, *Healthy People 2010: Understanding and improving health*, vol. 2, US Government Printing Office, Washington, DC, 2000.

[2] S Baik, B J Bowers, L D Oakley, and J L Susman, "The recognition of depression: The primary care clinicians perspective," *Annals Of Family Medicine*, vol. 3, no. 1, pp. 31–37, 2005.

[3] J C Mundt, P J Snyder, M S Cannizzaro, K Chappie, and D S Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response technology," *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[4] A J. Flint, S E. Black, I Campbell-Taylor, G F. Gailey, and C Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, July 1993.

[5] E Moore, M Clements, J Peifer, and L Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech.," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[6] K.R. Scherer, "Vocal assessment of affective disorders," in *Depression and Expressive Behavior*, J.D. Maser, Ed., pp. 57–82. Lawrence Erlbaum Associates, 1987.

[7] A Nunes, L Coimbra, and A Teixeira, "Voice quality of european portuguese emotional speech corresponding author," *Computational Processing of the Portuguese Language Lecture Notes in Computer Science*, vol. 6001/2010, pp. 142–151, 2010.

[8] L A Low, N C Maddage, M Lech, L B Sheeber, and N B Allen, "Detection of clinical depression in adolescents speech during family interactions.," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2011.

[9] J F. Cohn, T S Kruez, I Matthews, Y Yang, M H Nguyen, M T Padilla, F Zhou, and F De la Torre, "Detecting depression from facial actions and vocal prosody," *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–7, Sept. 2009.

[10] N Cummins, J Epps, M Breakspear, and R Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in *Proc. Interspeech*, 2011, pp. 2997–3000.

[11] S Alghowinem, R Goecke, M Wagner, J Epps, M Breakspear, and G Parker, "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech," in *Proc. FLAIRS-25*, 2012, pp. 141–146.

[12] S G Koolagudi and K S Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[13] Z Zeng, M Pantic, G I Roisman, and T S Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions.," *IEEE Trans. on PAMI*, vol. 31, no. 1, pp. 39–58, 2007.

[14] A. Ozdas, R.G. Shiavi, S.E. Silverman, M.K. Silverman, and D.M. Wilkes, "Analysis of fundamental frequency for near term suicidal risk assessment," *IEEE Conf. Systems, Man, Cybernetics*, pp. 1853–1858, 2000.

[15] T Polzehl, A Schmitt, F Metze, and M Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Communication*, vol. 53, no. 910, pp. 1198 – 1209, 2011.

[16] F Eyben, M Wöllmer, and B Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM'10)*, Oct. 2010, pp. 1459–1462.

[17] B Schuller, A Batliner, S Steidl, and D Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9–10, pp. 1062–1087, Nov. 2011.

[18] S Young, G Evermann, M Gales, T Hain, D Kershaw, X A Liu, G Moore, J Odell, D Ollason, D Povey, V Valtchev, and P Woodland, *The HTK Book (for Version 3.4)*, Cambridge University Engineering Department, 2006.

[19] C C Chang and C J Lin, "LIBSVM: a library for SVM," *2006-03-04. http://www.csic.ntu.edu.tw/ rcjlin/papers/lib.svm*, pp. 1–30, 2001.

[20] K. Tamás and L. T. Kóczy, "Mamdani-type inference in fuzzy signature based rule bases," in *8th International Symposium of Hungarian Researchers on CINTI*, 2007, pp. 513–525.

[21] B.S.U. Mendis and T.D. Gedeon, "A comparison: Fuzzy signatures and choquet integral," in *Fuzzy Systems. WCCI*, june 2008, pp. 1464–1471.

[22] H. Ben Mahmoud, R. Ketata, T. Ben Romdhane, and S. Ben Ahmed, "Hierarchical Fuzzy Signatures approach for a piloted quality management system," in *Systems, Signals and Devices*, Mar. 2011, pp. 1–6.

[23] B.S.U. Mendis, T.D. Gedeon, and L.T. Koczy, "Learning Generalized Weighted Relevance Aggregation Operators Using Levenberg-Marquardt Method," in *Proc. Hybrid Intelligent Systems*, Dec. 2006, p. 34.

[24] K Yu and S Young, "Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 19, no. 5, pp. 1071–1079, July 2011.

[25] E Moore, M Clements, J Peifer, and L Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," *Proc. 26th Ann. Conf. Eng. Med. Biol.*, vol. 1, pp. 17–20, Jan. 2004.