An Investigation of Depressed Speech Detection: Features and Normalization

Nicholas Cummins[1], *Julien Epps*[1,2], *Michael Breakspear*[3,4], *and Roland Goecke*[5]

[1] School of Elec. Eng. and Telecomm., The University of New South Wales, Sydney, Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia
[3] Black Dog Institute and School of Psychiatry, The University of New South Wales, Australia
[4] Division of Mental Health Research, Queensland Institute of Medical Research, Australia
[5] Faculty of Information Sciences and Engineering, University of Canberra, Australia
nicholas.cummins@student.unsw.edu.au, j.epps@unsw.edu.au

Abstract

In recent years, the problem of automatic detection of mental illness from the speech signal has gained some initial interest, however questions remaining include how speech segments should be selected, what features provide good discrimination, and what benefits feature normalization might bring given the speaker-specific nature of mental disorders. In this paper, these questions are addressed empirically using classifier configurations employed in emotion recognition from speech, evaluated on a 47-speaker depressed/neutral read sentence speech database. Results demonstrate that (1) detailed spectral features are well suited to the task, (2) speaker normalization provides benefits mainly for less detailed features, and (3) dynamic information appears to provide little benefit. Classification accuracy using a combination of MFCC and formant based features approached 80% for this database.

Index Terms: mental state recognition, depressed speech, feature comparison, MFCC, Gaussian mixture models

1. Introduction

Depression is a common mental disorder that presents persistent feelings of sadness, intrusive negative thoughts and, cognitive difficulties such as poor concentration, leading to functional impairment. Despite its high prevalence and enormous socio-economic burden, clinical practice remains rooted almost exclusively on the opinion of individual clinicians, risking a range of subjective biases. As a first step towards developing objective depression severity scales with clinical utility, the characterisation of depression using physiological and behavioural signals is needed. Speech is attractive since it can be measured cheaply, remotely, non-invasively and non-intrusively. However, it is also richly communicative and contains many sources of variability.

Early investigations of depressed speech found that depressed patients consistently demonstrated prosodic speech abnormalities, such as reduced variation in loudness, repetitious pitch inflections and stress patterns, and monotonous pitch and loudness [1]. A later study of 28 sufferers of depressive illness found similar evidence for a reduced variability in fundamental frequency and prosody [2]. In experiments by Stassen [3], speech features such as fundamental frequency and speech pause duration were found to be sufficiently highly correlated with the HAMD-17 depression score that simple speech analysis methods were trialled as an objective measure of patient recovery during a course of antidepressants. Flint [4] studied the effect of psychomotor retardation in depressed people and found that patients with a major depressive disorder had decreased second formant (F2) measurements when compared with a control group.

More recent work has seen the first steps towards automatic analysis of speech [5]. A range of acoustic features have already been identified for suitability in the classification of depression. Speech production cues such as pitch and formant measures are useful due in part to the effects of increased tension in the vocal tract associated with depression [4,5]. Spectral and energy based measures are also useful in classifiers, as depressive speech can contain more information in the higher energy bands when compared with neutral speech [5, 6]. Spectral centroid based methods including the sub-band spectral centroid features have recently shown promise in other applications [7], and other work [8] shows that these newer measures potentially include information useful in the classification of depression. Although systems for the classification of depressed speech have been proposed [9, 10], there is considerable further research to be conducted, particularly in light of the extensive speech-based emotion recognition literature, from which insight can be gained towards the classification of depressed speech.

In this paper, we examine the effect of segment selection and the choice of different speech characterization methods on the automatic classification of depressed and neutral speech.

2. Acoustic Characterization of Depression in Speech

2.1. Segment Selection

Beyond the usual isolation of speech-active regions in the signal using a voice activity detector (VAD), a question of interest is the accurate selection of speech segments that provide maximal depressed/neutral speech discrimination. To our knowledge, the decision between voiced-only, unvoiced-only or mixed-voicing speech is without empirical support. In this paper, we employ an energy-based VAD and confine ourselves to investigating the relative merits of voiced and unvoiced segments, using short term energy and F0 information to estimate the degree of voicing. Based on emotion recognition results and [8, 10], we expect to find that voiced segments provide the most effective discrimination.

2.2. Feature Extraction

Our framework for investigating acoustic features for characterising depressed speech is based on that adopted by [11]. The motivation is similar, namely to understand the relative contribution of speech production cues, detailed spectral information and broad spectral information to depressed/neutral speech discrimination.

Among speech production cues, pitch, energy and formants are obvious feature choices, for which insights into depressed speech exist already [5, 6, 8, 10]. In this work, we use the RAPT Algorithm to estimate F0 and Linear Prediction Analysis using VOICEBOX to estimate and track the first three formant frequencies.

Detailed spectral information can be represented by a variety of different features, including the widely used mel frequency cepstral coefficients (MFCCs), linear predictive group delay and the spectral centroid frequencies and amplitudes (SCF / SCA). SCF is a measure of the average weighted frequency for the *k*th sub-band, where the weights are the normalized energy of each frequency component *S*[*f*] in that sub-band:

$$F_k = \frac{\sum_{f=l_k}^{\mu_k} f\left|S[f]\omega_k[f]\right|}{\sum_{f=l_k}^{\mu_k} \left|S[f]\omega_k[f]\right|} \qquad (1)$$

where $\omega_k[f]$ represents the sub-band filter. SCA are simply the average magnitude of a given sub-band weighted on a component-wise basis by the frequency of each magnitude component in the *k*th sub-band:

$$A_k = \frac{\sum_{f=l_k}^{\mu_k} f\left|S[f]\omega_k[f]\right|}{\sum_{f=l_k}^{\mu_k} f} \qquad (2)$$

Both SCF and SCA were extracted using a mel scale Gabor filterbank, and calculated as described in [7].

Broad spectral information is of interest since detailed spectral information can be expected to contain substantial variability due to the phonetic composition of an utterance and the speaker identity. Typically, the former source of variability is accounted for by employing a Gaussian mixture model with many mixtures, which represent different acoustic regions within which depressed/neutral speech discrimination can be more effectively characterized. The latter source of variability can be mitigated using feature normalization methods. In this work, we investigate energy slope (ES – extracted by using least squares analysis of the magnitude spectrum to calculate the linear slope coefficient for a given frame), zero-crossing rate (ZC), and spectral centroid (SC – extracted across the full speech bandwidth).

2.3. Feature Normalization

As explained above, feature normalization can be applied to reduce the mismatch in feature distributions between different speakers. Research has shown that speaker variability is a stronger confounder for emotion recognition than phonetic variability [11], and we adopt a similar hypothesis for depressed/neutral speech classification.

An important distinction between emotion recognition and depressed/neutral speech classification must be made: unlike emotion, which can be transient and hence easily elicited from a single speaker in a variety of forms, depression is a condition that is sustained for weeks to months. Finding speakers able to produce both depressed and neutral speech, even over a long period of time, is very challenging. Hence, the depressed and neutral speaker utterances might often be mutually exclusive with respect to speaker identity. Normalizing on a per-speaker basis for depressed/neutral speech will prove beneficial if the speaker variability is substantially larger than the depressed/neutral speech variability (this seems likely [11]). On the other hand, if the depressed/neutral speech variability is larger than the speaker variability, then per-speaker normalization might prove detrimental to depressed/neutral classification.

2.4. Modeling of Depressed Speech

Following the approach of many paralinguistic speech classification systems based on acoustic features, including some focused on depressed speech recognition, we employ Gaussian mixture models (GMMs) to model depressed and neutral speech. In contemporary systems, researchers often make use of multiple subsystems and score-level fusion, to combine the benefits of individual systems and advance the state of the art. The overall system described in Section 2 is summarised in Fig. 1.

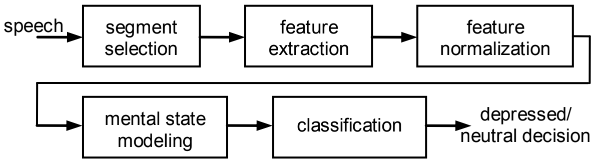

Figure 1: *Depression detection system configuration*

3. Experimental Configuration

3.1. Database

The database of 23 depressed and 24 control subjects (approximately 50%/50% male and female) was obtained from audio-video data collected during an ongoing study into measuring the facial activity in depressed patients [12] at the Black Dog Institute. No formal measure was used to differentiate between the depressed and control groups. Subjects were excluded from the control group if they had any personal or family history of mental illness. All subjects gave informed consent and the study was conducted in accordance with local institutional ethics committee approval.

As part of the experimental setup, participants were asked to read sentences containing affective content. In the readings two sets of sentences, as used in [13] to examine the effects of emotional content and context on verbal memory, were read aloud. The first set contained emotionally arousing 'target' words. The second set replaced the target word with a well matched neutral word.

3.2. System Configuration

The experimental settings (unless otherwise stated) of the depression detection system were as follows: the VAD employed was energy based, retaining the 80% highest energy frames from each utterance. Thirteen MFCC coefficients were extracted, and the delta and delta-delta coefficients (Δ,ΔΔ) were extracted by the standard regression equation over seven consecutive frames. The shifted delta coefficients (SDC) were extracted with parameters *N –d –p – k* equal to 7-1-3-7. For the spectral measures of SCF/SCA 20 coefficients were extracted. For normalization, feature warping (cumulative distribution mapping) was used. This technique maps each feature to a pre-determined (normal) distribution. HTK was used to train the GMMs using 16 mixtures and 10 iterations of the EM algorithm, with a variance flooring threshold set to 0.01. The choice of the number of mixtures (fixed to ensure consistent comparison) is a difficult one since detailed spectral measures benefit more from more detailed modeling (assuming sufficient data are available) than lower dimension features.

3.3. Evaluation Methods

Each of the 47 speakers' recordings contained 20 individual sentences (referred to herein as utterances), of approximately 40-60s total duration, per speaker. All identified acoustic parameters were tested for their effectiveness in both *speaker dependent* and *speaker independent* systems. In the speaker dependent tests, utterances were allocated to training and test databases by randomly selecting 14 utterances for training and 6 for test per speaker. For speaker independent tests, utterances from 20 random speakers for the depressed group and 21 random speakers from the control group were aggregated for training, and the remaining set of utterances from the unused 6 speakers was used for testing. Different features may have different levels of speaker dependent and depression dependent characteristics, therefore testing both dependent and independent systems will allow us to further compare the speaker variability of these features. Ten-fold cross-validation was applied to the test database, and the average accuracy was reported. To calculate the overall accuracy results reported, the log-likelihood of each test utterance was calculated with respect to both a depressed and control (neutral) GMM, followed by a maximum-likelihood decision. Unless otherwise stated the reported accuracies are for the speaker independent system.

4. Results

4.1. Feature Variation for Depressed and Control Patients

The discriminative capabilities of one-dimensional features can be directly visualized. Here we show as an example the distribution of F1 for depressed and neutral speech (Fig. 2). Formant features have already been shown to capture information useful in distinguishing between the two classes [4,5]. Here, the difference between the two classes for the F1 measure is also reflected by the depressed class having 5% greater standard deviation than the neutral class.

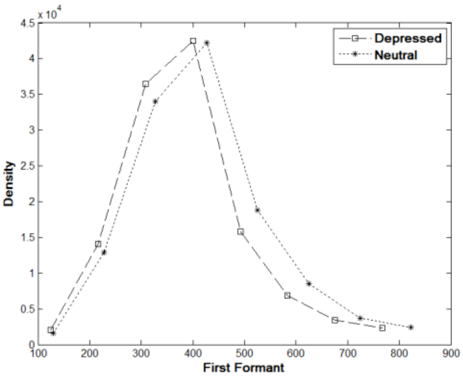

Figure 2: *Distributions of feature F1 for depressed and neutral speakers (both equally gender balanced).*

4.2. Segment Selection

The speech database employed was annotated carefully to minimize the amount of silence/background noise. Hence, as seen in Table 1, accuracy was high across a range of threshold values, and using either energy-based or F0-based criteria. For this database, providing the lowest energy frames are discarded, the choice of criterion and threshold is not critical, although voiced frames seem to be preferable to unvoiced.

Table 1. *Classification accuracy for different segment selection approaches*

Segment selection	Frames retained (%)				
	20	40	60	80	100
High energy frames ret.	73	76	78	77	78
Low energy frames ret.	51	69	75	77	
Voiced frames retained	77	76	77	-	
Unvoiced frames ret.	69	74	-	-	

4.3. Feature Extraction

Results from the feature comparison, seen in Table 2, broadly agree with those in [11]. Specifically, detailed spectral measures performed best for the speaker dependent case, capturing depression related information effectively. Their performance in the speaker independent case was slightly less convincing, however, suggesting that depression is manifested somewhat differently in different speakers for those features. Single-dimension features on the whole seemed less affected by the speaker independent conditions, as might be expected (they carry less speaker-specific information). The linguistic constraints on the database may have been an influence on these results.

Interestingly the best performing single dimensional feature was the first formant. It has been previously reported in the literature that the second formant location is most affected by depressive speech [4]. F2 has also been linked with emotional and cognitive information. The performance of speech production cues F0 and energy was lower than expected, in agreement with findings in [10], where it is suggested that whilst F0 and energy offer a high level view of vocal tract dynamics they do not provide true information on vocal tract tension.

Table 2. *Comparison of feature types using two-class depressed/neutral speech classification accuracy, evaluated on a 47-speaker database.*

Features	**Classification Accuracy**		
	Spk Dep.	**Spk Indep.**	
		No Warp	Warp
Single Dimensional Features			
F0	53	48	64
Short Term Energy (E)	58	58	60
Energy Slope (ES)	51	50	70
Zero Crossing Rate (Z)	58	57	59
F1	**64**	**60**	72
F2	56	50	70
F3	60	58	**79**
Spectral Centroid (SC)	60	58	52
Multidimensional Features			
F0 + Energy (F0E)	54	42	56
Energy Slope + ZCR(SZ)	61	56	61
MFCC	**80**	**77**	65
SCF	69	51	58
SCA	**80**	76	51
Group Delay (GD)	76	73	**72**
Formants (F1,F2,F3,A1,A2,A3)	79	74	70

When normalization was introduced, the accuracies of single dimensional features nearly all increased, many substantially. By contrast, there was little improvement (and even a drop in performance from MFCC) in the detailed spectral measures when normalization was introduced, broadly in line with [11]. This result may be consistent with prior research in that depression manifests differently in different speakers for detailed spectral measures. As in [11], whilst detailed spectral measures are able to distinguish emotions and in our case depression, they are also very characteristic of their speaker. Feature warping attempts to reduce the variation in data due to differences in speaker variability. But if a proportion of the between-speaker variability removed in warping is due to the effects of depression, the ability to distinguish between the two classes will also be reduced. The improvement seen in some of the single-dimension features could be explained by their lack of speaker specific information. Mean and mean-variance normalization results (not shown) were substantially poorer than for warping.

4.4. Dynamic Information

To test whether temporal information provides benefit to depressed/neutral speech classification, three separate speaker independent experiments were conducted, as seen in Table 3. From this preliminary result, it appears that including dynamic information provides no classification benefit. This result is consistent with the results reported in [8], where inclusion of the first and second order derivatives increased classification accuracy by just 3%.

Table 3. *Classification accuracy for dynamic information features included in a MFCC based system*

Features	MFCC+Δ	MFCC+Δ+ΔΔ	MFCC+ SDC
Accuracy (%)	78	78	77

4.5. Comparison of Fused Systems

The foregoing results indicate that the detailed spectral feature of MFCCs can be taken as a baseline system. To attempt to improve on this performance, various pairs of features were combined using score level fusion. The logistic regression method employed for this was trained on the evaluation data, so results should be interpreted as upper bound rather than typical. It can be seen from results in Table 4 that combining different features with broad spectral measures does not significantly improve classification accuracy. For example, the combination of F0 and energy measures with other features does not improve accuracy beyond the classification of formants on their own. This may reflect the earlier observation that these measures don't provided detailed information on vocal tract tension.

Table 4. *Comparison of two class classifier accuracies for various feature-pair combinations, evaluated on the 47-speaker database*

Features	**Classification Accuracy (%)**		
	Spk Dep.	**Spk Indep.**	
		No Warp	Warp
MFCC + SC	82	80	64
MFCC + F0E + SZ	81	79	67
MFCC + Formants	81	79	70
GD + Formants	81	78	72
Formants + F0E + SZ	77	74	70
Formant + SC	76	72	70

Combining with detailed spectral measures with formant information seems a useful approach. This makes sense as detailed spectral features are complemented by the vocal tract information contained in the formant feature. If amplitude information is omitted from the formant feature vector, classification accuracy drops by approximately 10%, so that both the frequency and amplitude of the formants are needed.

4.6. Detection Error Curves

The results reported to this point are accuracies based on maximum likelihood classification, however the classification threshold is a design choice, and it is instructive to consider different error trade-offs at different system operating points. The DET curve in Figure 4 shows the improved classification performance when combining MFCC and formant information. It also shows the negative effect of warping on MFCCs.

5. Conclusion

Several insights have been gained from this study of depressed/neutral classification. Voiced speech segments appear to be mildly preferable for the purpose, however segment selection is not critical. Detailed spectral features are very well suited to the speaker-dependent problem, but are also the feature of choice for the speaker-independent case. Feature warping needs to be used with care, however its behaviour in this investigation is similar to that for emotion recognition, despite the differences in the structure of the respective recognition problems. Temporal feature evolution appears not to provide any benefit for depressed/neutral discrimination. Future work will consider glottal features and improved normalization methods.

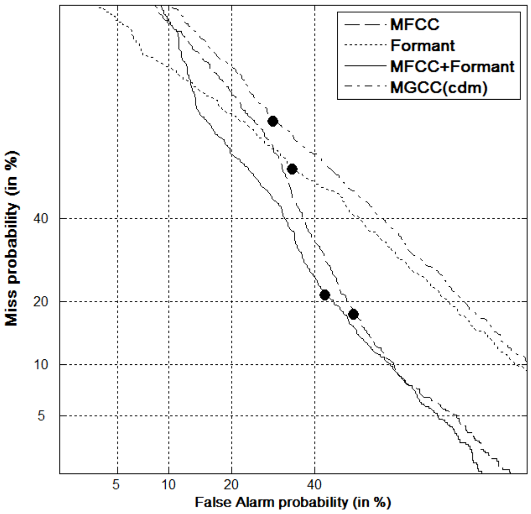

Figure 4. *Comparison of DET curves showing the optimum detection cost points for system configurations based on MFCC and formant features*

6. Acknowledgements

This research was funded by the ARC Discovery Projects scheme (DP 110105240), the NHMRC and the Ronald Geoffrey Arnold Foundation.

7. References

[1] Darby, J. K., "Speech and voice parameters of depression: A pilot study", *J. Commun. Disord.*, 17, 1984, pp. 75-85.

[2] Nilsonne, A., "Speech characteristics as indicators of depressive illness", *Acta Psych. Scandinavia*, vol. 77, 1988, pp. 253-263.

[3] Stassen, H., Kuny, S., and Hell, D., "The speech analysis approach to determining onset of improvement under antidepressants", *Eur. Neuropsych.*, vol. 8, 1998, pp. 303-310.

[4] Flint, A.J., et al., "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression", *Journal of Psychiatric Research*, 1993. 27(3): pp. 309-319.

[5] France, D. J., et al., "Acoustical properties of speech as indicators of depression and suicidal risk", *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, July 2000, pp. 829-837.

[6] Ozdas, A.,et al., "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk", *IEEE Trans. Biomed. Eng.*, vol. 51, no. 9, 2004, pp. 1530-1540.

[7] Kua, J. M. K., et al. "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition." *Proc. Odyssey: Speaker and Lang. Rec. Workshop*, 2010, pp. 34 - 39.

[8] Low, L. S. A., N. C. Maddage, et al. "Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents." in *Proc. IEEE Int. Conf. on Cognitive Informatics*, 2009, pp. 346-350.

[9] Yingthawornsuk, T., et al., "Direct Acoustic Feature Using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech." in *Proc. Interspeech, 2007.*

[10] E. Moore, et al. "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, 2008, pp. 96-107.

[11] Sethu, V., et al., "Speaker dependency of spectral features and speech production cues for automatic emotion classification", in *Proc. IEEE ICASSP*, 2009, pp. 4693-4696.

[12] McIntyre, G., R. Göcke, et al., "An approach for automatically measuring facial activity in depressed subjects", in *Proc. Int. Conf. on Affective Computing and Intelligent Interaction and Workshops*, 2009.

[13] Brierley, B.N. Medford., et al., "Emotional memory for words: Separating content and context", *Cognition & Emotion*, 2007. 21(3): p. 495-521.