Probabilistic Acoustic Volume Analysis for Speech Affected by Depression

Nicholas Cummins^{1,2}, Vidhyasaharan Sethu¹, Julien Epps^{1,2}, Jarek Krajewski^{3,4}

¹School of Elec. Eng. and Telecomm., The University of New South Wales, Sydney Australia

² ATP Research Laboratory, National ICT Australia (NICTA), Australia

³ Experimental Industrial Psychology, University of Wuppertal, Wuppertal, Germany

⁴ Industrial Psychology, Rhenish University of Applied Sciences Cologne, Germany

n.p.cummins@unsw.edu.au, v.sethu@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Alterations in speech motor control in depressed individuals have been found to manifest as a reduction in spectral variability. In this paper we present a novel method for measuring acoustic volume - a model-based measure that is reflective of this decrease in spectral variability - and assess the ability of features resulting from this measure for indexing a speaker's level of depression. A Monte Carlo approximation that enables the computation of this measure is also outlined in this paper. Results found using the AVEC 2013 Challenge Dataset indicate there is a statistically significant reduction in acoustic variation with increasing levels of speaker depression, and using features designed to capture this change it is possible to outperform a range of conventional spectral measures when predicting a speaker's level of depression.

Index Terms: Depression, Gaussian Mixture Models, Acoustic Volume, Monte Carlo Approximation

1. Introduction

Depression has a heterogeneous clinical profile which produces a range of different physiological effects. The effect of many of these symptoms, such as a reduced cognitive ability, a continuous negative affect, fatigue and psychomotor retardation, is to alter speech motor control. Reductions in cognitive ability affect the planning and preparation of the muscular commands needs to produce speech causing phonetic and articulatory errors [1]. Continuous negative affect, fatigue and psychomotor retardation affect muscle tension, altering both vocal tract dynamics and laryngeal coordination as well as constraining articulatory movement [2].

Depressed speech is commonly described as having one or more of the following qualities: flat, dull, slurred, breathy or tense [3-7]. Despite this description, the use of pitch based features has seen mixed results; reductions in F0 range with increasing levels of depression are not uniformly reported in the literature [3-6]. Voice Quality features, as a result of a decrease in laryngeal coordination, have been shown to change significantly with changes in a speakers level of depression [6,7]. However, as in [6,7], suitable speech segments are required when extracting Voice Quality features to ensure clinically reliable results [8,9].

Changes in muscle tension and control, observable as a reduction in variability in facial features and head movement with increasing levels of depression [10], have consistently been linked with changes in vocal tract properties. Recent results published show that there is a lack of spectral variability with increasing level of depression [6,11,12]. In the recent AVEC Challenge [13], spectral based features such as Mel Frequency Cepstral Coefficients (MFCC), formants and MFCC-based supervectors demonstrated a high level of accuracy when used to predict a speaker's level of depression from a given multimedia file [14,15].

Decreases in spectral variability, as a result of depression, have been shown to be present in the MFCC/Gaussian Mixture

Model (GMM) representation of acoustic space. Cummins et al. report a statistically significant decrease in the correlation between depression severity and Average Weighted Variance (AWV), a measure which estimates the extent of localized acoustic variability present in a GMM [12]. Intuitively this result, combined with a lack of spectral variability, matches the flat or dull descriptions of speech affected by depression.

For the purpose of this paper, we use the term 'acoustic volume' to notionally refer to the 'size' of the feature space occupied by feature vectors extracted from some speech data. It is reasonable to expect that the relative size of this region will reflect acoustic variability and consequently will vary due to variations in speaker characteristics, phonetic content and paralinguistic phenomena modulated into the speech utterance/segment.

In a recently proposed approach to estimating acoustic volume from a GMM of the probability distribution of the feature vectors extracted from an utterance by Krishnamurthy and Hansen, acoustic volume was found to be proportional to the acoustic variation in the data [16]. Their results show that as the number of overlapping phonemes in a speech segment increase, acoustic volume decreases. The authors state that this reduction is due to a lack of distinguishing characteristics between spectra of individual phoneme, i.e. a reduction in acoustic variation, as the number of overlapping phonemes increases [16].

Motivated by the results in [12] we presently explore the hypothesis that due to a lack of speech motor control there is a reduction in acoustic variation of speech produced under greater levels of depression.

2. GMM Mean Acoustic Volume

In this section we give a brief description of the GMM Mean Acoustic Volume (GM-AV) technique, taken from [16], for estimating acoustic volume. This method uses a set of GMM means to define a hyper cuboid, the volume of which is an indicator of the acoustic variation present in the data. For the complete description of the technique see [16].

If $\mathbb{M} = \{\mu_1, \dots, \mu_M\}$ represents a set of mixture means of an *M*-mixture GMM which adequately describes the Ndimensional feature distribution of a given speech segment, there exists an *N*-dimensional hyper-cuboid minimal volume that contains \mathbb{M} within it. This cuboid has 2^N vertices characterized by the set of maxima and minima points evaluated for each of the *N* dimensions separately across all vectors in \mathbb{M} . It is possible to estimate the volume enclosed by this set by identifying the set of *N* edges $\{e_1, e_2, \dots, e_N\}$ which encloses the cuboid. Each edge e_k is defined by the maxima and minima points of the k^{th} feature dimension:

$$e_k = \max(n_k) - \min(n_k) \tag{1}$$

where $n_k = \left\{ \mu_1^{(k)}, \mu_2^{(k)}, \dots, \mu_M^{(k)} \right\}$ is a set of *M* elements and $\mu_i^{(k)}$ denotes the k^{th} dimension of the i^{th} mean vector, $\boldsymbol{\mu}_i \in \mathbb{M}$. The volume of cuboid (GM-AV) is then given by:

$$V = e_1 \times e_2 \times \dots \times e_N \tag{2}$$

It should be noted that this technique estimates acoustic volume using GMM mean vectors only. It has been shown for depressed speech classification that depression information is not only captured in the means but also in the covariance matrices and weight parameters [12].

3. Proposed Probabilistic Acoustic Volume

The Acoustic Volume measure given by (2) does not take into account the weight of each component of the GMM when calculating an edge. This makes it susceptible to potentially significant overestimation of acoustic volume if low-weight mixtures with means far from feature clusters are present as illustrated in Figure 1. Specifically, in the example shown at the top of the page, the difference between the extremely low weight mixture 3 being centered at X = 8 (Figure 1a) and being centered at X = 3 (Figure 1b) is negligible when the overall probability density modelled by the GMM is considered, however the difference in acoustic volume (GM-AV) estimated using equation (2) is significant.



Figure 1 (a) 1-dimensional 3-mixture GMM with low weight mixture 3 centered at X = 8; (b) 1-dimensional 3-mixture GMM with low weight mixture 3 centered at X = 3

An alternative and potentially more robust estimate of acoustic volume may be obtained by taking a probabilistic approach. Specifically, given an *N*-dimensional feature space, $X \in \mathbb{R}^N$, if the underlying distribution of the features is denoted by P(X), an estimate of the probabilistic acoustic volume (PAV), V_{θ} , can be obtained as the total volume of the space where P(X) is greater than some threshold θ (refer Figure 2). i.e.,

$$V_{\theta} = \int f(X)dX, \quad f(X) = \begin{cases} 1, & P(X) > \theta \\ 0, & P(X) \le \theta \end{cases}$$
(3)

An alternative view of the PAV, as can be seen from Figure 2, is that selecting a value for θ is akin to picking a cross-section of the feature distribution and V_{θ} is an estimate of the cross-sectional volume in the feature space. Further, by defining a series of thresholds, θ_i for $1 \le i \le L$, an array of probabilistic acoustic volumes, **V**, herein referred to as the probability acoustic volume profile (PAV profile), can be obtained which is characteristic of the 'concentration' of the feature vectors in the feature space.



Figure 2: Sample 1-dimensional feature distribution with volume of feature space corresponding to probability higher than a threshold $-\theta$

Given that outliers are by definition data points (feature vectors) of low probability, it can be expected that the PAV profile would not be distorted by outliers since a range of θ values are taken into consideration. Mathematically, the PAV profile may be defined as a set:

$$\mathbf{V} = \{ V_{\theta_i} \colon 1 \le i \le L \} \tag{4}$$

As can be seen from Figure 3 the shape of the PAV profile is representative of whether the feature vectors are tightly concentrated (Figure 3a) or spread out in the feature space (Figure 3b). In practice it is computationally efficient to approximate the underlying probability density function of the feature space with a GMM, which then allows for a Monte Carlo approach to computing V_{θ} . Specifically, the expression for PAV (eqn. 3) may be rewritten as:

$$V_{\theta} = \int \hat{f}(X) P(X) dX, \qquad (5)$$

where

$$\hat{f}(X) = \begin{cases} \frac{1}{P(X)}, & P(X) > \theta\\ 0, & otherwise \end{cases}$$
(6)

Consequently

$$V_{\theta} = E_P[\hat{f}(X)] \tag{7}$$

where, $E_P[\cdot]$ denotes the expected value with respect to the probability density function P(X).

A Monte Carlo approximation of this value is given by

$$\hat{V}_{\theta} = \frac{1}{N} \sum_{x_i \sim P(X)} \hat{f}(x_i), \qquad 1 \le i \le N$$
(8)

where x_i denotes points in the feature space drawn from the probability density function P(X). When a GMM, $\mathcal{G}(X)$, is used to approximate P(X) the process of drawing samples from it is greatly simplified. Specifically, given an *M*-mixture GMM, *N* points are drawn from it by drawing N_i points from each Gaussian component such that $N_i/N = \omega_i$, where ω_i is the



Figure 3: (a) Estimate of probability acoustic volume profile for a sample 1-dimensional feature distribution reflecting two small 'highly densely concentrated' feature clusters (b) Estimate of probability acoustic volume profile for an example 1-dimensional feature distribution reflecting two wide low density feature clusters exhibiting a 'low concentration' of features.

weight associated with the i^{th} component and $\sum_i N_i = N$. The PAV estimate \hat{V}_{θ} , given by eqn (8), then reduces to a sum involving the fraction of points whose probability, elevated using the GMM as $\mathcal{G}(x_i)$, is greater than the threshold θ .



Figure 4: (a) Example 2-mixture GMM and points drawn from the mixtures of the GMM – black denotes points with probability less than θ =0.15 and blue denotes points with probability greater than θ =0.15. (b) Comparison of Probabilistic Acoustic Volume (PAV) and GMM mean acoustic volume (GM-AV) – Points drawn from example 2mixture GMM, with blue denoting the points considered in computing PAV. The violet box represents GM-AV computed from mixture means.

Comparing the proposed probabilistic acoustic volume (PAV) and the GMM mean acoustic volume (GM-AV), it can be seen that they do not estimate the same quantity (Figure 4). The GM-AV is an estimate of the hypervolume spanned by all mixture components of a GMM. It is dependent on intermixture distances and the volume may contain sparse regions of the feature space. The PAV on the other hand is independent of

distances between feature clusters and is dependent only on the size and density of feature clusters.

4. Experimental Settings

4.1 Depression Corpus

All experimental results in this paper are reported on the Audio/Visual Emotion Challenge and Workshop (AVEC) 2013 corpus. The full data set comprises 150 recordings, each labelled with a Beck Depression Inventory (BDI) score [17], divided into training, development and testing partitions each of 50 files (recordings) with a mean file length of 14 min 52 sec containing a mix of different speech types: read speech including an excerpt from the novel "Homo Faber" by Max Frisch, free response speech and vocal exercises. Papers published on this corpus include [13-15,18].

4.2 Experimental Settings

The experimental settings (unless otherwise stated) were as follows: to retain phonetic consistency whilst maintaining a suitably long speech segment all evaluations were performed on "Homo Faber" excerpts with a mean length of 3 min, 24 sec. This excerpt was unavailable for two speakers in the test set, for these files an appropriate length of continuous German speech was used, this approach was also used in [15]. Frame-level MFCC features were extracted as per [14]. All UBMs were trained with 10 iterations of the EM algorithm. As per [12,14], speaker specific GMM's were formed using full adaption, with five iterations of the MAP algorithm.

A PAV profile, with L = 21 was extracted per file using 100,000 points (*N*) for the Monte Carlo approximation using (8). The θ values were chosen experimentally; if a given θ is too small the resulting V_{θ} will not contain sufficient discriminatory volume information, conversely if θ is too high the resulting V_{θ} will equal zero. Our θ values were equally spaced on the log probability scale between 0 and the highest probability found in the combined training and development set, using 30 values. We then removed the top 4 values and bottom 5 values, decided experimentally on the development set, to reduce the effect of outliers in our tests. The resulting array was compressed to rescale volume estimates in order to make them independent of the feature space dimensionality:

$$\mathbf{V} = (V_{\theta_1}^{\frac{1}{D}}, \cdots, V_{\theta_{21}}^{\frac{1}{D}})$$
(9)

where D = 39 was the dimensionality of the underlying feature space.

To gain a one-dimensional parameterization of PAVprofile for comparison with GM-AV, we fitted a negative linear slope to each of the profiles. Whilst the resulting coefficient (PAVslope) is not acoustic volume, in the same sense as GM-AV, it is a measure of the spread of data in the feature space, the steeper this slope, the more tightly concentrated the feature space (Figure 3).

The discriminatory strength of the PAVslope coefficient was compared with that of GM-AV in a series of correlations and 2-class, speaker independent *t*-tests with results reported in terms of Spearman's rank coefficient and Hedge's *g* coefficient for the two tests respectively as well as the associated *p*-value. The UBM's were trained using the entire testing partition, approx. 13hrs of data. Correlations were performed using all 100 training and development files. For the *t*-test, the first class contained 38 files with BDI < 10 (low levels of depression), the second contained a further 38 files, with a BDI > 18 (moderate to severe depression).

For the score level prediction results, all UBM's were trained using the entire training partition, approximately 12hrs of data, and prediction results were recorded in terms of Root Mean Square Error (RMSE). Prediction results for PAVslope and GM-AV were found using linear regression whilst PAV profile results were generated using a Linear Support Vector Regressor (SVR), as in [14]. All prediction results were compared with the brute forcing (BF), Kullback-Leibler (KLmeans) supervector, Vocal Tract Correlation (VTC) feature and i-vector results from [13], [14], [15] and [18] respectively.

5. Results

5.1 Statistical Analysis

The negative correlations, seen in Table 1, show that GM-AV decreases with increasing levels of depression, and significant correlations (p<0.01) are seen for 8 and 16 mixtures. However the results from the 2 class *t*-test results ($|g| \le 0.5$, $p \ge 0.05$) show that GM-AV cannot sufficiently differentiate between low/high levels of depression.

For PAVslope, the correlations (p < 0.001) and T-test results (|g| > 0.5, p < 0.01) provide strong evidence for the suitability of this feature as a marker of depression (Table 1). The positive correlations show that the steepness of the PAVprofile increases with increasing levels of depression. This is a strong indicator that the MFCC feature space becomes more tightly concentrated with increasing levels of depression.

Table 1: GM-AV and PAVslope statistical test results, for three different GMM sizes, calculated on two different partitionings of the AVEC 2013 dataset

Feat.	GMM	Corre	lation	T-Test		
	Mixes	Rho	р	Hedge's g	р	
GM- AV	8	-0.32	1.3E-03	0.15	0.52	
	16	-0.28	4.5E-03	-0.09	0.68	
	32	-0.24	1.8E-02	0.31	0.17	
PAV slope	8	0.45	3.1E-06	-0.69	3.3E-03	
	16	0.50	1.2E-07	-0.95	7.9E-05	
	32	0.49	1.7E-07	-1.02	2.7E-05	

5.2 Score Level Prediction

To test the performance of GM-AV, PAVslope and PAVprofile as markers of depression we ran a series of score level prediction tests. Results from this analysis showed that GM-AV was not well suited to performing score level prediction (Table 2); this is not surprising given the results in section 5.1. The PAVslope scores, given their strong statistical results, are disappointing; although it is worth noting this single dimensional feature was able to outperform the high dimensional (2268 features), brute forcing approach used in [13]. PAVprofile matches performance with conventional spectral features in the development set, however the weaker test set results indicate potential overfitting when setting the thresholds. It should also be noted that whilst direct comparison with [15] is possible, a direct comparison of results in [13], [14], and [18] is not straightfoward; these results were calculated using the entire files, not just the "Homo Faber" passage.

Table 2: Comparing AVEC 2013 Development and Test RMSE's for acoustic volume measures with accuracies reported in [13], [14], [15] and [18]

Sustam		Devel.		Test		
System	8	16	32	8	16	32
GM-AV	12.55	11.89	11.85	11.45	11.53	11.64
PAVslope	11.26	10.48	10.55	11.36	11.37	11.47
PAV profile	10.50	10.60	10.75	11.87	11.86	12.09
BF [13]		10.75		14.12		
KLmean [14]	9.60			10.17		
VTC [15]	7.42			9.49		
i-vector [18]	10.34			11.37		

6. Conclusion

A decrease in feature space variability and in localized model domain variability have been previously reported when using the MFCC/GMM paradigm to parameterize and model depressed speech [12]. To parameterize these reductions in acoustic variation we proposed the Probabilistic Acoustic Volume, and a technique for its calculation using Monte Carlo methods for estimating the volume of a region of space where the corresponding probability distribution is greater than some given threshold. This technique resulted in two features which strongly relate to a speaker's level of depression.

PAVslope performed well in statistical testing; the strong correlations provide evidence for reduced variations in the MFCC feature space. This result, and arguably also the GM-AV correlations, complement the results presented in [12] and provides strong evidence for our hypothesis that altered speech motor control in depression causes a reduction in acoustic variations. Both PAVslope and the full PAV profile performed adequately in the score level prediction results, outperforming the AVEC 2013 baseline [13].

Future work will include developing ways of automatically setting threshold values and experimenting with different methods to collapse the PAVprofile into a smaller dimensional feature space with fewer redundancies. Given the results here and in [12] we will also be exploring other measures for characterizing how changes in speech motor control with increasing levels of speaker depression are manifested in acoustic space.

7. Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This work was partly funded by ARC Discovery Project DP130101094 (Epps) and the German Research Foundation (KR3698/4-1) (Krajewski). The authors wish to thank Dr Michel Valstar and Kirsty Smith from the University of Nottingham for calculating the AVEC 2013 test set results.

8. References

- G. Christopher and J. MacDonald, "The impact of clinical depression on working memory," *Cogn. Neuropsychiatry*, vol. 10, no. 5, pp. 379–399, 2005.
- [2] 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," *J. Psychiatr. Res.*, vol. 27, no. 3, pp. 309–319, 1993.
- [3] A. Nilsonne, "Speech characteristics as indicators of depressive illness," *Acta Psychiatr. Scand.*, vol. 77, no. 3, pp. 253–263, 1988.
- [4] 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 (IVR) technology," *J. Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.
- [5] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biol. Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.
- [6] T. F. Quatieri and N. Malyska, "Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity," in 13th Annual Conference of the International Speech Communication Association Interspeech2012, 2012, pp. 1059–1062.
- [7] S. Scherer, G. Stratou, J. Gratch, and L. Morency, "Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD," in 14th Annual Conference of the International Speech Communication Association Interspeech2013, 2013, pp. 847–851.
- [8] R. F. Orlikoff and J. C. Kahane, "Influence of mean sound pressure level on jitter and shimmer measures," *J. Voice*, vol. 5, no. 2, pp. 113–119, 1991.
- [9] J. Laver, S. Hiller, and J. M. Beck, "Acoustic waveform perturbations and voice disorders," *J. Voice*, vol. 6, no. 2, pp. 115–126, 1992.
- [10] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image Vis. Comput.*, Dec. 2013.
- [11] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-Temporal Analysis of Speech Affected by Depression and Psychomotor Retardation," in 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7542–7546.
- [12] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling Spectral Variability for the Classification of Depressed Speech," in 14th Annual Conference of the International Speech Communication Association Interspeech2013, 2013, pp. 857–861.
- [13] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge," in *Proceedings of the 3rd ACM International Workshop* on Audio/Visual Emotion Challenge, 2013, pp. 3–10.
- [14] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of Depression by Behavioural

Signals: A Multimodal Approach," in *Proceedings of* the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 2013, pp. 11–20.

- [15] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal Biomarkers of Depression Based on Motor Incoordination," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 41–48.
- [16] N. Krishnamurthy and J. Hansen, "Babble noise: modeling, analysis, and applications," *Audio, Speech, Lang. Process. IEEE Trans.*, vol. 17, no. 7, pp. 1394–1407, 2009.
- [17] D. Maust, M. Cristancho, L. Gray, S. Rushing, C. Tjoa, and M. E. Thase, "Chapter 13 - Psychiatric rating scales," in *Handbook of Clinical Neurology*, vol. Volume 106, F. B. Michael J. Aminoff and F. S. Dick, Eds. Elsevier, 2012, pp. 227–237.
- [18] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability Compensation in Small Data: Oversampled Extraction of I-Vectors for the Classification of Depressed Speech," in 2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 970–974.