SPEECH LANDMARK BIGRAMS FOR DEPRESSION DETECTION FROM NATURALISTIC SMARTPHONE SPEECH

Zhaocheng Huang[1], Julien Epps[1], Dale Joachim[2]

[1]School of Electrical Engineering and Telecommunications, UNSW Sydney, Australia
[2]Sonde Health, Boston MA, USA

zhaocheng.huang@unsw.edu.au, j.epps@unsw.edu.au, djoachim@sondehealth.com

ABSTRACT

Detection of depression from speech has attracted significant research attention in recent years but remains a challenge, particularly for speech from diverse smartphones in natural environments. This paper proposes two sets of novel features based on speech landmark bigrams associated with abrupt speech articulatory events for depression detection from smartphone audio recordings. Combined with techniques adapted from natural language text processing, the proposed features further exploit landmark bigrams by discovering latent articulatory events. Experimental results on a large, naturalistic corpus containing various spoken tasks recorded from diverse smartphones suggest that speech landmark bigram features provide a 30.1% relative improvement in F1 (depressed) relative to an acoustic feature baseline system. As might be expected, a key finding was the importance of tailoring the choice of landmark bigrams to each elicitation task, revealing that different aspects of speech articulation are elicited by different tasks, which can be effectively captured by the landmark approaches.

Index Terms— Depression classification, landmark bigrams, speech articulation, smartphone speech, naturalistic environments.

1. INTRODUCTION

The increasing adoption of smartphones coupled with the emergence of voice assistants provides unprecedented opportunities for new automated medical screening methods through sampling of human voice [1], [2], [3], [4], motivated by the reported 10-15% of the population suffering from mental disorders [3]. However, lingering challenges to automated speech-based screening via smartphones remain, due in large part to the detrimental impact of various handset characteristics and noisy environments from which audio samples are recorded [5]. This impact of handset variability is particularly evident in systems using conventional spectral-based features [6], [7]. Thus, vocal biomarkers robust to these variabilities need exploration, and one candidate is speech landmarks.

Speech landmarks are event markers associated with articulation of speech. They offer an alternative speech processing framework focused on abrupt acoustic changes in speech articulation that remain relatively discernable even in the presence of variability introduced by diverse smartphone hardware and background environments. Moreover, studies have shown that speech production, which involves complex cognitive planning and motoric muscular actions, can be impacted by depression in various ways [8], including cognitive impairment, phonation and articulation errors, articulatory incoordination [9], disturbances in muscle tension, phoneme rates [10], and altered speech quality and prosody. Landmark biomarkers can indicate many of these attributes, and therefore offer a unique potential to capture depression-related cues in speech articulation in ways, which, to the best of our knowledge, have not been previously explored.

In this paper, we investigate two novel sets of features based on speech landmarks, i.e. landmark bigrams, and topic modelling of landmark bigrams via *Latent Dirichlet Allocation* (LDA) for depression classification in natural realistic environments.

2. RELATED WORK

Speech is produced by a series of articulator narrowings and releases [11]. The dominant speech processing methods derive frame-level acoustic features such as mel frequency cepstral coefficients (MFCCs) at fixed frame rates (such as 100Hz), within which the encapsulated signal is assumed time-invariant and stationary. By contrast with and independent of frames, landmark methods characterize articulatory elements of speech, and detect timestamp boundaries denoting sharp changes in speech articulation [12], [13], [14] (as seen in Figure 1). The introduction of landmarks dates back to Stevens *et al.* in 1992 [15], where landmarks were proposed to segment speech for lexical representation associated with articulators. Later, landmarks have been used in other fields, primarily for speech recognition [11], [12], [14].

While speech landmarks are relatively less common than their frame-based analysis counterparts, they are nonetheless increasingly probed. For instance, landmarks have been used to study both lexical content of speech [12], [13], [14] and non-lexical attributes of speech such as syllabic complexity [16] and voice-onset time [17]. Recently, landmarks have been investigated for paralinguistic content, e.g. children vocalization [18], emotion [19], Parkinson's disease and sleep deprivation [20]. In [19], landmark features were found to complement conventional acoustic features for emotion recognition, yet only three consonantal landmarks were used.

Compared with conventional acoustic features, landmarks offer a number of advantages: increased channel robustness and information related to the articulatory effects of depressed speakers, which are commonly reported to be important [8], [9], [10]. However, landmarks have mostly been treated as a tool to segment points in time about articulatory changes, while deriving more effective and meaningful representation from landmarks has yet to be investigated.

Since landmarks are events, with a symbolic rather than numerical representation, they can potentially be effectively exploited by analysis approaches more akin to those from Natural Language Processing (NLP). It has been shown in literature that the text analysis is very effective for both depression classification and prediction [21], [22], [23]. Transcripts of interview questions together with topic-wise multimodal features yielded very promising results for depression prediction [21], and rich

information in text was found to be valuable for depression classification and prediction [22], [24].

3. PROPOSED LANDMARK BIGRAM FEATURES

3.1. What are landmarks?

This section introduces six landmarks adopted in this study, each with onset (+) and offset (−) states. They are '**g**(lottis)', '**p**(losives)', '**s**(onorant)', '**f**(ricative)', '**v**(oiced fricative)', and '**b**(ursts)', which essentially specify points in *time* for different abrupt articulatory events (summarized in Table 1).

Table 1: *Description of the six landmarks investigated.*

Landmark	Description
g	sustained vibration of vocal folds starts (+) or ends (−).
p	sustained periodicity begins (+) or ends (−)
s	opening (+) or closing (−) of the velopharyngeal port during a sonorant sound
f	frication onset (+) or offset (-)
v	voiced frication onset (+) or offset (-)
b	onset (+) or offset (-) of existence of turbulent noise during obstruent regions

They are detected once certain evidence of rapid changes (i.e. rises or falls) in power across multiple frequency ranges and multiple time scales is observed. Among the landmarks, 's' and 'v' relate to voiced speech, whereas 'f' and 'b' relate to unvoiced speech. Detailed descriptions for the landmark extraction process can be found in [25]. Examples of the landmarks identified from speech can be seen in Figure 1.

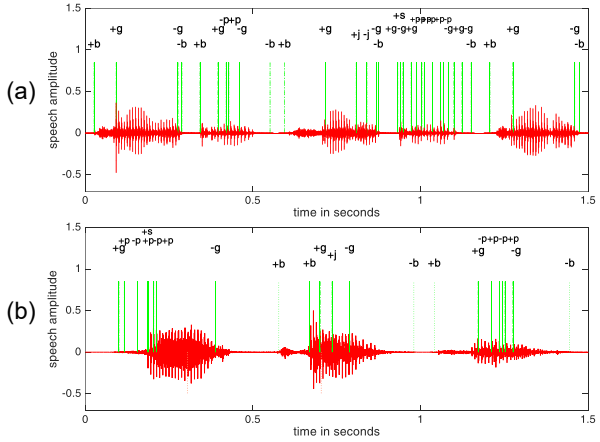


Figure 1: *Detected landmarks for the word "PaTaKa" uttered by two speakers: (a) healthy and (b) depressed, with PHQ-9 scores of 3 and 27 respectively. Both recordings were from males using Samsung smartphones. The healthy speaker uttered "PaTaKa" faster and more frequently than the depressed speaker.*

Although landmarks are informative when used alone, it is suggested by [12] that landmark bigrams carry more information regarding the speech articulation, for instance, the bigram (-b, +b) represents the number of pauses in speech.

3.2 Proposed Landmark Bigram Features

3.2.1. Bigram-Count

We define a set of L landmarks, each with onset (+) and offset (-) states, i.e. $2L$ states in total:

$$S = \{g_{\pm}, p_{\pm}, s_{\pm}, f_{\pm}, v_{\pm}, b_{\pm}\} \quad (1)$$

and associated with a speech file, the sequence of identified landmarks is $\{l_1, l_2, \dots, l_n, \dots l_{N+1}\}$, where $l_n \in S$ represents the n^{th} landmark (n is a non-uniform time index), and $N+1$ is the number of landmarks per speech recording (different across different files).

Landmark bigrams are defined as pairs of consecutive landmarks $w_n^{i,j} = (l_n = i, l_{n+1} = j)$, where $i, j \in S$ represent a specific pair of landmarks. Accordingly, for the speech file $d \in \{1, \dots, D\}$, the landmark bigrams are

$$\boldsymbol{w}_d = \{w_{d,1}, \dots, w_{d,n}, \dots, w_{d,N}\} \quad (2)$$

The landmark bigram count for $w^{i,j}$ can then be defined:

$$c_d^{i,j} = \#(\boldsymbol{w}_d^{i,j}) \quad (3)$$

where $\#(\cdot)$ represents the counting operation per speech file d. Concatenating all possible bigrams gives a vector of bigram counts:

$$\boldsymbol{c}_d = [c_d^{g_+,g_+}, \dots, c_d^{i,j}, \dots, c_d^{b_-,b_-}]^T \in \mathbb{R}^{4L^2} \quad (4)$$

3.2.2. LDA-Bigram

Latent Dirichlet Allocation (LDA) has been widely used for topic modelling in NLP since its introduction in 2003 [26], [27], [28]. LDA generates a representation of latent topics (e.g. sports, travel, etc.) given documents consisting of words. Motivated by this, landmark bigrams herein were treated as 'words', from which meaningful articulatory events may be efficiently exploited using LDA. Figure 2 depicts the graphic model of LDA in the context of depression classification using landmark bigrams.

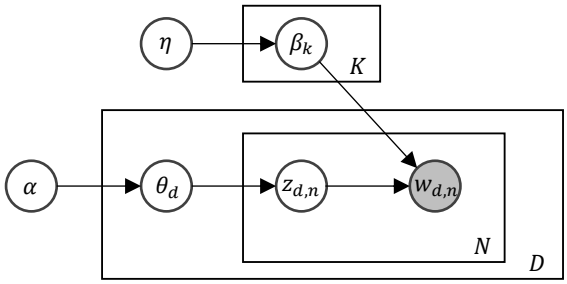


Figure 2: *The graphical model for LDA. There are D speech files ('documents'), N landmark bigrams ('words'), and K latent articulatory events ('topics'). $w_{d,n}$ is the n^{th} bigram in the d^{th} speech file. The latent variables $z_{d,n}$, β_k, θ_d are estimated from training, controlled by hyperparameters α and η.*

In Figure 2, β_k and θ_d follow a Dirichlet distribution, while $z_{d,n}$ and $w_{d,n}$ follow a Multinomial distribution.

$$\begin{gathered}\beta_k \sim \text{Dir}(\eta), \theta_d \sim \text{Dir}(\alpha) \\ z_{d,n} \sim \text{Multi}(\theta_d), w_{d,n} \sim \text{Multi}(\beta_{z_{d,n}})\end{gathered} \quad (5)$$

$\theta_d = \{\theta_{d,1}, \dots, \theta_{d,k}, \dots, \theta_{d,K}\}$, $\sum_{i=1}^{K} \theta_{d,i} = 1$ characterizes the probability distribution over K articulation events for the file $d \in \{1, \dots, D\}$, while $\beta_k = \{\beta_{k,1}, \dots, \beta_{k,n}, \dots, \beta_{k,N}\}$, $\sum_{n=1}^{N} \beta_{k,n} = 1$ characterizes the probability distribution over N bigrams for the articulation event $k \in \{1, \dots, K\}$. For $w_{d,n}$, an articulation event $z_{d,n} = k$ is sampled from θ_d to yield $\beta_{z_{d,n}=k}$, which specifies the probability of generating $w_{d,n}$. Overall, $z_{d,n}$, β_k, and θ_d together describe relationships for *bigram-articulation-speech*, which is similar to *word-topic-document* in topic modelling.

The objective of LDA is to learn probabilities for latent events θ_{d^*} from the observed bigrams $\boldsymbol{w}_{d^*}$ in a new document d^*. The training of LDA involves optimization of the posterior distribution

of latent variables given observations, $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \eta)$, which is however intractable. Thus, the optimization is done by Variational Bayesian Inference, i.e. approximating $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \eta)$ using a simpler distribution $q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$, so that the Kullback-Leibler (KL) divergence between the two distributions is minimized. $q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$ can be fully factorized as below:

$$\begin{gathered} q(z_{d,n} = k) \sim \text{Multi}(\phi_{d,n}^k) \\ q(\beta_k) \sim \text{Dir}(\lambda_k), q(\theta_d) \sim \text{Dir}(\gamma_d) \end{gathered} \tag{6}$$

As all distributions in (6) are exponential family distributions, they have analytical solutions for guaranteed convergence of KL divergence [26], [29]:

$$\begin{gathered} \phi_{d,n}^k \propto \mathbb{E}_{q(\theta_d)}[\log \theta_{d,k}] + \mathbb{E}_{q(\beta_k)}[\log \beta_{k,w_{d,n}}] \\ \gamma_{d,k} = \alpha + \textstyle\sum_w c_{d,w}\, \phi_{d,n}^k \\ \lambda_{k,w} = \eta + \textstyle\sum_d c_{d,w}\, \phi_{d,n}^k \end{gathered} \tag{7}$$

$c_{d,w}$ is the bigram counts, i.e. the number of times bigram *w* appears in document *d*, as in (4). The parameters $\phi_{d,n}^k$, λ_k, and γ_d are iteratively updated until convergence. However, it is worth noting that this optimization solution is non-convex, meaning that there exist multiple local maxima depending upon initialization of the latent variables.

After having the trained parameters, i.e. $\phi_{d,n}^k$, λ_k, and γ_d, the latent structure for speech articulation given bigrams is characterized. Given a new audio file d^*, which has landmark bigram counts $c_{d^*,w}$, the resultant LDA-bigram features are θ_{d^*}:

$$\begin{gathered} \gamma_{d^*,k} = \alpha + \textstyle\sum_w c_{d^*,w}\, \phi_n^k \\ \theta_{d^*} \sim \text{Dirichlet}(\gamma_{d^*,1}, \ldots, \gamma_{d^*,K}) \end{gathered} \tag{8}$$

where $\theta_{d^*} = \{\theta_{d^*,1}, \ldots, \theta_{d^*,K}\}$ and $\sum_{k=1}^K \theta_{d^*,k} = 1$.

Accordingly, the proposed landmark bigram counts $\boldsymbol{c}_{d^*}$ and the LDA representation of bigrams $\boldsymbol{\theta}_{d^*}$ (referred to as LDA-bigram) were used as features for depression detection.

4. EVALUATION

4.1. Experimental Settings

As per [5], the experiments were conducted on the SH2 corpus, containing 16 hours of speech data. This corpus is a collection of audio recordings (durations from 4-30s) in naturalistic environments, and self-reported Patient Health Questionnaire (PHQ-9) scores gathered through an interactive Android™ smartphone app. It contains 5863 audio files for 887 speakers (437 females and 450 males), each of whom completed up to six elicitation tasks, i.e. sustained vowel ('Ut'), diadochokinetic ('Wo'), free speech (FS), rainbow passage ('Pa'), cognitive load ('CL') and sentence ('Se'). The SH2 corpus has the same training and testing partition as [5]: 4584 files (695 speakers) for training and 1279 files (192 speakers) for testing. As a result of applying a PHQ-9 threshold of 10 to separate healthy (PHQ-9<10) and depressed (PHQ-9≥10) speakers (suggested by [30]), 122 and 35 depressed speakers were respectively found in the training and test data partitions.

The landmarks were extracted using the SpeechMark® toolbox [16], a publicly available, representative landmark extraction software. Note that bigrams that did not occur within the training data were removed from the bigram list, leading to actual feature dimensions much smaller than the full bigram list. For LDA, the number of latent articulation events *K* is an important parameter, optimized from 2 to 40 unless specified.

For comparison with acoustic features in [5], all experiments in this study adopted a linear Support Vector Machine (SVM) [31] classifier, which was fine-tuned through parameter sweeps of *C* from 10^{-5} to 10 in a log space in a 3-fold cross validation scheme within the training data, and the best parameter was adopted for testing on the test data. During training, *C* was weighted inversely proportional to class frequencies to handle imbalanced training data for the healthy and depressed classes, as per [5]. Since each speaker conducted up to six elicitation tasks, task-based decisions were fused per speaker via majority voting, except for Section 4.3 (where depression detection was investigated per task). F1 score (for depressed speakers), accuracy, and *Unweighted Averaged Recall* (UAR) were used for evaluating performance for speakers. Parameter *C* was optimized for F1 on the training data.

4.2. Landmark Bigram Features

The first question addressed was which landmarks are effective for differentiating depressed and healthy speakers. Further, experiments were carried out to study the importance of normalization specific to task and gender, which have shown great benefits for depression detection [5]. Gender or task normalization was proposed in [5] to normalize training and test data specific to a certain gender or task with coefficients learnt from the training data. Figure 3 compares the proposed bigram-count and LDA-bigram features with/out task/gender normalization. Moreover, landmarks were appended one-by-one to find an optimal set of landmarks to derive bigram-count, starting from "*s*" to all six landmarks "*s,v,p,f,g,b*". The same set of landmarks were then used for LDA-bigram to examine the benefit of LDA representation.

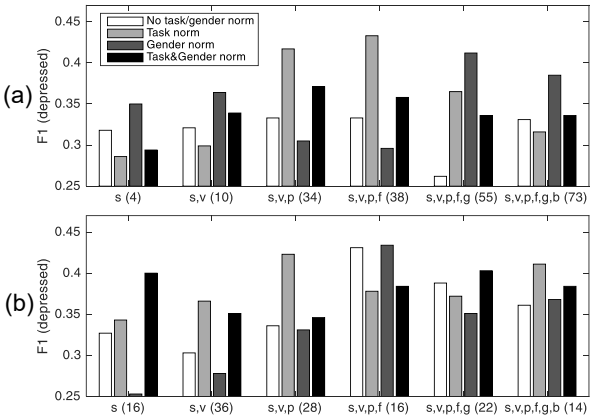


Figure 3: *F1 (depression) scores (chance level=0.267) for (a) bigram-count and (b) LDA-bigram. Within the brackets are the feature dimensions.*

In Figure 3, the 's' (sonorant) appeared important and achieved 0.35 and 0.4 in F1 scores for bigram-count and LDA-bigram. Also, it was beneficial to apply task/gender normalization in most cases for bigram-count (achieving 0.433 and 0.412 within each case), while gender normalization showed more importance than task normalization for LDA-bigram. The benefit of including gender normalization was not surprising, as this is in line with the gender dependency of landmarks for Parkinson's disease reported in [20]. Results also suggest that larger numbers of landmarks improved learning of the LDA representation, which achieved 0.431.

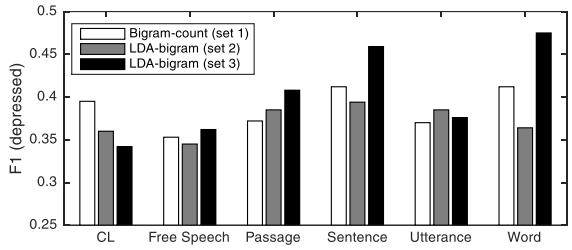
The fact that task normalization was helpful for bigram-count was also expected, since there are certain unique sets of landmark bigrams involved when people conduct different elicitation tasks. For instance, a reading task consistently produce roughly the same number of the same landmarks, whereas landmark occurrence during free speech will depend on word choice. Utterances of 'PATAKA' produce landmarks depending on how quickly people say it in a fixed specified time window (as in Figure 1).

The observation above, on the other hand, suggests that it is sub-optimal to rely on the same set of landmark bigrams for all elicitation tasks, and benefits can be obtained from tailoring bigrams for each task. The latter point motivated the search for optimum bigrams for each task in the following experiments.

4.3. Landmark Bigrams Optimized for Elicitation Tasks

In this section, choices of landmarks were tailored for each elicitation task to consider the uniqueness of landmark bigrams for each task, as well as to better understand the underlying articulatory aspects for each task via the best-performing landmark bigrams. For bigram-count, gender normalization was applied, except for the 'CL' task. For LDA-bigram, gender normalization was only applied to the 'Pa' and 'Ut' tasks. z-normalization was applied to normalize training and testing data in all systems.

There are three sets of experiments in Figure 4: set 1 looked for the best landmarks for bigram-count; set 2 used the same set of landmarks for LDA-bigram, as per Figure 3; In set 3, the event number *K* was empirically set to 4 (i.e. the resultant features have 4 dimensions), and the landmarks were tailored for LDA-bigram. The motivation for set 3 was that there might exist a different set of landmarks that are associated more closely with latent articulation events than using those optimized from bigram-count.



	CL	*FS*	*Pa*	*Se*	*Ut*	*Wo*
#D/#H	24/108	23/105	32/140	19/89	34/150	30/137
set 1	g,p,s,f (27)	f,v,b (16)	g,s,b (18)	s,f,b (18)	s,f,v,b (31)	g,s,v,b (27)
set 2	g,p,s,f (6)	f,v,b (36)	g,s,b (36)	s,f,b (14)	s,f,v,b (10)	g,s,v,b (30)
set 3	s,f (4)	g,p,f (4)	f,s,v (4)	g,p,b,v (4)	s,f,b (4)	p,f,v (4)

Figure 4: *F1 (depression) score (chance-level=0.267) for bigram-count and LDA-bigram within each elicitation task, which has various most effective landmarks. The table summarizes the number of depressed and healthy speakers (i.e. #D and #H) per task, selected landmarks, and feature dimensions for each set of experiments.*

LDA-bigram performed the best for "Wo", "Se" and "Pa" tasks with merely 4 features extracted from the tailored landmarks, achieving 0.475, 0.459 and 0.408 in F1 score respectively. Also, it was beneficial to tailor landmark choices for each task, which gave improvements for across three sets of experiments. It is also worth mentioning that within each task, mostly 3-4 landmarks were effective in characterizing the unique speech articulation for depression classification.

4.4. Fusion of Proposed Landmark Features

Choices for landmark bigrams per task were found important for both bigram-count and LDA-bigram, and it is expected that better performances can be obtained via fusion of the task-optimized systems. This section accomplishes this using majority voting across all task-based decisions for each speaker.

As shown in Table 2, fusion of the optimized landmarks per task gave significant improvements over those using landmarks optimized over all tasks, 0.506 vs 0.433 for bigram-count, and 0.549 vs 0.431 for LDA-bigram, in F1 scores. Compared with acoustic features used in [5], the use of LDA-bigram yielded significantly better performances, confirming the effectiveness of landmark bigrams as well as the LDA technique for classifying depression.

Table 2: *Fusion of task decisions using Majority Voting. The 2nd and 3rd rows fused all tasks which adopted the same landmark bigrams, whereas the 4th and 5th rows fused task-optimized landmark bigrams. The three best performing tasks were fused for LDA-bigram* (i.e. "Wo", "Se", and "Pa"), and five tasks were fused for bigram-count* (i.e. excluding only "Ut").*

	F1	**Accuracy**	**UAR**	**Confusion Matrix**
Acoustic features [5]	0.422	72.9%	0.657	$\begin{bmatrix}121 & 36\\ 16 & 19\end{bmatrix}$
Bigram-count#	0.433	71.4%	0.669	$\begin{bmatrix}116 & 41\\ 14 & 21\end{bmatrix}$
LDA-bigram#	0.431	65.6%	0.679	$\begin{bmatrix}101 & 56\\ 10 & 25\end{bmatrix}$
Bigram-count*	0.506	78.7%	0.714	$\begin{bmatrix}130 & 27\\ 14 & 21\end{bmatrix}$
LDA-bigram*	**0.549**	**78.7%**	**0.758**	$\begin{bmatrix}126 & 31\\ 10 & 25\end{bmatrix}$

"#" means the same landmarks for all tasks (Figure 3), whereas "" means different optimized landmarks for each task (Figure 4).*

It is worth noting that the acoustic features in [5] require careful selection of voice activity detectors to remove background environment noise, whereas the landmark detectors operate directly on the raw speech signal to pick up certain articulatory events that are robust to the noise.

5. CONCLUSION

This research has presented two novel sets of features based on speech landmark bigrams for depression detection under naturalistic environment, i.e. bigram-count, which counts the occurrences of landmark bigrams, and LDA-bigram, which effectively discovers latent articulation patterns from landmark bigrams. The proposed features are appealing, since they capture information from speech articulation, possibly exhibit increased robustness to channel variability and environment noise, have reduced feature dimensionality, and improved performance. In particular, the most effective sets of landmarks and investigations of task-specific settings were presented, yielding significant improvements over systems without consideration of the landmark selection and task-specific information.

Apart from the improved performance, this research is significant in a number of ways: 1) the first study to our knowledge to apply landmark bigrams for depression detection, showing great promise; 2) yielding further improved results over those on datasets which have clean recordings, a single recording environment and long utterances [32]; 3) results were evaluated on a larger number of speakers compared with previous studies; 4) there is no gap between PHQ-9 in separating the depressed and healthy speakers.

Future work involves in-depth analysis and interpretability of articulatory disfunction for depressed speakers using the proposed features. The proposed features will be tested on new depression datasets to validate their effectiveness and generalization. Investigations into word representation as well as classification methods from NLP may improve detection performance further.

6. ACKNOWLEDGEMENT

This work was supported by Australian Research Council Linkage Project LP160101360. Julien Epps is also partly supported by Data61, CSIRO, Australia.

7. REFERENCES

[1] Insel, T. R., "Digital phenotyping: Technology for a new science of behavior," *Journal of the American Medical Association*, vol. 318, no. 13, pp. 1215–1216, 2017.

[2] Ben-Zeev, D., E. A. Scherer, R. Wang, H. Xie, Andrew, and T. Campbell, "Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health," *Psychiatric Rehabilitation Journal*, vol. 38, no. 3, pp. 218–226, 2015.

[3] Walker, J., K. Burke, M. Wanat, R. Fisher, J. Fielding, A. Mulick, S. Puntis, J. Sharpe, M. Degli Esposti, and E. Harriss, "The Prevalence of Depression in General Hospital Inpatients: A Systematic Review and Meta-Analysis of Interview Based Studies," *Psychological Medicine*, 2018.

[4] Cohn, J. F., N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal Assessment of Depression from Behavioral Signals," in *Handbook of Multi-Modal Multi-Sensor Interfaces*, D. Oviatt, S., Schuller, B., Cohen, P., and Sonntag, Ed. Morgan and Claypool, 2017, pp. 113–155.

[5] Huang, Z., J. Epps, D. Joachim, and M. Chen, "Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions," in *INTERSPEECH*, 2018, pp. 3393–3397.

[6] Stasak, B. and J. Epps, "Differential performance of automatic speech-based depression classification across smartphones," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 171–175.

[7] Mitra, V., A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *IEEE ICASSP*, 2016, pp. 5795–5799.

[8] Cummins, N., S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, Jul. 2015.

[9] Williamson, J., T. Quatieri, and B. Helfer, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on AVEC, ACM MM*, 2014.

[10] Trevino, A. C., T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 42, 2011.

[11] Liu, S. A., "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[12] Park, C., "Consonant Landmark Detection for Speech Recognition," *PhD Thesis, MIT, USA*, 2002.

[13] Stevens, K. N., "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[14] Hasegawa-johnson, M., J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, C. Mellon, J. Hopkins, and G. Tech, "Landmark-Based Speech Recognition : Report of the 2004 Johns Hopkins Summer Workshop," in *ICASSP*, 2005, pp. 213–216.

[15] Stevens, K. N., S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a Model for Lexical Access based on Features," in *ICSLP*, 1992, no. October, pp. 499–502.

[16] Boyce, S., H. J. Fell, and J. MacAuslan, "SpeechMark: Landmark Detection Tool for Speech Analysis.," in *INTERSPEECH*, 2012, pp. 1894–1897.

[17] Chenausky, K., J. MacAuslan, and R. Goldhor, "Acoustic Analysis of PD Speech," *Parkinson's Disease*, vol. 2011, pp. 1–13, 2011.

[18] Fell, H. J., L. J. Ferrier, and S. G. Worst, "Vocalization Age as A Clinical Tool," in *ICSLP*, 2002, pp. 1–4.

[19] Dai, K., H. Fell, and J. MacAuslan, "Recognizing emotion in speech using neural networks," *Proceedings of the Fourth IASTED International Conference*, pp. 31--36, 2008.

[20] Ishikawa, K., J. MacAuslan, and S. Boyce, "Toward clinical application of landmark-based speech analysis: Landmark expression in normal adult speech," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. EL441-EL447, 2017.

[21] Gong, Y. and C. Poellabauer, "Topic Modeling Based Multi-modal Depression Detection," in *ACM Multimedia, AVEC '17*, 2017, pp. 69–76.

[22] Huang, Z., B. Stasak, T. Dang, K. Wataraka Gamage, L. Phu, V. Sethu, and J. Epps, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in *Proceedings of the 6th International Workshop on AVEC, ACM MM*, 2016.

[23] Williamson, J. R., E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting Depression using Vocal, Facial and Semantic Communication Cues," *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 11–18, 2016.

[24] Dang, T., B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps, "Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017," in *ACM Multimedia, AVEC '17*, 2017, pp. 27–35.

[25] Macauslan, J. and F. P. Landmarks, "What Are Acoustic Landmarks , and What Do They Describe ?," pp. 15–18, 2016.

[26] Blei, D. M., A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[27] Blei, D. M. and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[28] Blei, D., L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65, 2010.

[29] Hoffman, M., D. Blei, and F. Bach, "Online Learning for Latent Dirichlet Allocation," in *NIPS*, 2010, pp. 1–9.

[30] Kroenke, K., R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[31] Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[32] Valstar, M., J. Gratch, F. Ringeval, M. T. Torres, S. Scherer, and R. Cowie, "AVEC 2016 – Depression , Mood , and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on AVEC, ACM MM*, 2016, pp. 3–10.