Investigation of Speech Landmark Patterns for Depression Detection

Article *in* IEEE Transactions on Affective Computing · October 2019

DOI: 10.1109/TAFFC.2019.2944380

Zhaocheng Huang
Amazon AWS AI

Julien Epps
UNSW Sydney

Investigation of Speech Landmark Patterns for Depression Detection

Zhaocheng Huang, *Member, IEEE*, Julien Epps, *Member, IEEE*, Dale Joachim, *Member, IEEE*

Abstract— The massive and growing burden imposed on modern society by depression has motivated investigations into early detection through automated, scalable and non-invasive methods, including those based on speech. However, speech-based methods that capture articulatory information effectively across different recording devices and in naturalistic environments are still needed. This article proposes two feature sets associated with speech articulation events based on counts and durations of sequential landmark groups or *n*-grams. Statistical analysis of the duration-based features reveals that durations from several consecutive landmark bigrams and onset-offset landmark pairs are significant in discriminating depressed from non-depressed speakers. In addition to investigating different normalization approaches and values of *n* for landmark *n*-gram features, experiments across different elicitation tasks suggest that the features can be tailored to capture different articulatory aspects of depressed voices. Evaluations of both landmark duration features and landmark *n*-gram features on the DAIC-WOZ and SH2 datasets show that they are highly effective, either alone or fused, relative to existing approaches.

Index Terms—Depression classification, landmark *n*-grams, speech articulation, smartphone speech, naturalistic environments

1 Introduction

DEPRESSION, a major mental disorder reported to afflict 10-15% of the world's population [1], places severe health, security, productivity and economic burdens on modern society. Early detection and treatment of depression can help relieve this economic burden while increasing the productivity and quality of life of depressed individuals. However, treatment of depression is expensive and often delayed due to the scarcity of trained psychological clinicians and often late diagnosis of mental disorder symptoms. Furthermore, the cost of early detection by either spot or large-scale screening is prohibitive due to the aforementioned reasons. Therefore, alternative technology-based screening methods have been sought in the form of inexpensive, automatic systems, to facilitate large scale early detection and connect with timely intervention.

The lack of effective depression screening candidate technologies has attracted research attention for more than a decade. To date, there have been a number of studies on automatic detection of depression ranging from voice, facial video, EEG signals, head pose, eye gaze, etc [1], [2], [3], [4]. Among these modalities, speech, which has demonstrated promising effectiveness and efficiency as an indicator of depression [1], remains notably non-invasive and easily accessible. However, most studies to date on speech-based depression detection have primarily focused on laboratory-collected data, recorded from a single channel in a clean environment.

The increasing adoption of smartphones coupled with the emergence of voice assistants provide unprecedented opportunities for new automated medical screening methods through sampling the human voice [5], [6], [7], [8] notably: 1) the ability to accumulate a sufficiently large quantity of data to statistically model variations in speech patterns for depressed and non-depressed individuals across populations and audio recording devices type; and 2) the ability to administer individual tailored questionnaires, analyze voice samples and provide clinical screening feedback across large populations. However, conventional features developed from clean lab-based datasets may not generalize as well in real-world applications due to the dramatic differences in speech recording such as noise conditions, handset hardware, design protocols, etc [7], [9], [10]. This shortcoming motivates the design of a new category of effective features for detecting depression under both environments.

Current speech processing methods typically segment speech into short 10-20 millisecond frames before extracting low-level descriptors (e.g. spectral, prosodic, and glottal features, etc. [1], [11], [12]) as well as high-level representations of those features such as statistical functionals (e.g. mean and percentiles, etc.), vocal tract coordination (VTC) features [13], *i*-vectors [14] and Fisher vectors [15].

However, there are a few drawbacks to the well-structured frame-based approach. First, all frames are treated equally, which undermines the fact that some frames contain less information than others. Second, frame-based features such as spectral features are vulnerable to channel variability, especially for smartphone speech, which is commonly collected various handset types. Moreover, the majority of acoustic features are extracted during 'steady-state' (in the 'middle' of phonemes), whereas a lot of important information related to speech production (and impairments in speech production) are related to changes in speech. By analogy,

- *Zhaocheng Huang and Julien Epps are with School of Electrical Engineering and Telecommunications, the University of New South Wales (UNSW Sydney), Australia, 2052. E-mail: {zhaocheng.huang, j.epps}@unsw.edu.au.*
- *Dale Joachim is with Sonde Health Inc., Boston, MA, United States. E-mail: djoachim@sondehealth.com.*

in image processing, edge detection is often a critical part of automatic systems [16]. Recognizing the link between articulatory changes and psychomotor retardation, some researchers have investigated transitions in acoustic/prosodic features [13], [17] or phoneme rate [18], however these approaches did not explicitly detect the timing and type of the articulatory changes, both of which are important for detecting depression.

Recently, landmark-based features were shown to provide discriminative information for speech-based depression classification [19], particularly using relatively simple counts of consecutive landmark bigrams. These encouraging results prompted extensive further investigations of speech landmark-based classification systems for depression reported herein, notably 1) landmark bigram durations, 2) use of *n*-grams for characterizing landmark patterns, 3) normalization methods for landmark *n*-grams, 4) which types of composite landmark patterns perform well in depression classification, 5) relationships between elicitation approaches and landmark analysis methods, 6) comparison of landmark-based systems with published systems, and 7) how performant landmark features can shed light on the impact of depression on speech articulation.

The broader research question in this paper investigates new ways with which to leverage changes and patterns in speech articulation rather than steady-state speech production, the basis for most acoustic features. Here we treat landmark features as proxies for speech articulation, and focus on two sets of landmark features - duration and *n*-gram count, as applied to specific patterns of consecutive landmarks. The evaluation is carried out across two different data sets: the first comprised of audio recorded in a clean controlled environment, and the second collected from noisy and unpredictable general smartphone environments.

2 Related Work

2.1 Depression and Speech Production

Many studies to date have shown that speech production, which involves complex cognitive planning and motoric muscular actions, can be impacted by depression in various ways [1], including cognitive impairment [20], phonation and articulation errors, articulatory incoordination [21], disturbances in muscle tension, psychomotor retardation, phoneme rates [18], and altered speech quality and prosody [17]. Thus, articulators, whose movements shape speech production, are expected to be informative for the changes incurred by depression. However, surprisingly there are not many studies investigating or adopting articulatory features for depression.

An important aspect of speech production relates to timing; depression affects articulatory movements and causes the slowing of speech, which has been found to be an indicator of psychomotor retardation [15], [18]. To this end, timing-based features such as speech rate, pause rate, and phoneme rate have been proposed for depression detection and prediction [18], [22], [23], [24]. More specifically, depressed speakers tend to have fewer pauses and increased speech errors [22], [25]. Besides, a more sophisticated set of neurologically-motivated Vocal Tract Coordination (VTC) features were proposed to capture psychomotor retardation by correlating different feature trajectory dynamics at different time scales, which showed great promise in predicting depression severity [13], [21]. However, the caveat is that psychomotor retardation does not always occur in depressed patients.

Despite the importance of articulation, the most commonly reported features for speech-based depression classification remain low level descriptors (e.g. spectral and prosodic features) [12], [22], [26]. These 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. The frame-level acoustic features allow extraction of higher level features, e.g. statistical functionals, vowel space area [27], [28], acoustic space modelling [29], etc. A general reduction in the vowel space area for depressed speakers was observed by Scherer *et al.* [28]. A similar reduction, but in the acoustic feature space, has also been found in depressed speech [29]. The aforementioned feature sets have also been trialled for other mental disorders such as Parkinson's disease [30], and PTSD (Post Traumatic Stress Disease) [28], Alzheimer's disease, schizophrenia, etc., which however have received less attention compared with depression in automatic systems.

Regardless of feature types, one challenge facing the choice or design of effective speech features to date is to generalize from lab-based research to real-world deployment (using smartphone speech in particular). The generalization can be undermined by many factors, primarily including handset variability, environment noise, etc. Hence, conventional features such as MFCCs, which are sensitive to channel variability, might be less than ideal [31]. Recently, the adoption of mobile phones as medical devices for mental health screening has gained attention, yet the investigation remains in its infancy, especially when it comes to effective speech features for depression screening [7], [10], [32], [33], [34].

2.2 Speech Landmarks

Speech landmarks are event markers associated with articulation of speech [19]. More precisely, they rely solely on the location of acoustic events in time, commonly occurring at times of consonant closures/releases, nasal closures/releases, glide minima and vowel maxima [35], thereby providing information about articulatory events (e.g. vibration of the vocal folds). By contrast with the frame-based processing framework and independent of frames, landmark methods characterize articulatory elements of speech, and detect timestamp boundaries denoting sharp changes in speech articulation [36], [37], [38] (as seen in Fig. 1). To this end, speech landmarks offer the potential to circumvent the aforementioned drawbacks underlying the frame-based processing, and an alternative speech processing framework that is focused on acoustically measurable changes in speech.

The introduction of landmarks in speech processing dates back to Stevens *et al.* in 1992 [39], who proposed them to segment speech for lexical representations associated with articulators. Later, landmarks were used in other fields, primarily for speech recognition [40], [36], [38]. This was mostly done via distinctive features [40], [39], [41], a binary primitive representation of speech, which bridges acoustic evidence and articulators: speech landmarks can be viewed as acoustic evidence of distinctive features [42], whereas some distinctive features may be mapped to particular articulators. Furthermore, the overlay of distinctive feature geometry [43], [44], a tree structure representation of feature sequence provides a framework for contextualizing not only static distinctive features and their acoustic observables but their dynamics over time (refer to *n*-grams discussed later in the paper).

Herein lies one of the main benefits and differentiators of landmarks: *the mechanism to link speech features to articulators*. If articulator motors are impacted by depression as suggested in the literature [18], then the identification of specific motor function deviation due to depression may lead to better tuned knowledge-based depression classification and furthermore, a better understanding of depression. This is in line with the argument in [45] that *"points around which articulatory information can be extracted, which thereby is useful to detect changes in motor coordination due to, more deeply about the motor planning, working memory and integration of auditory and proprioceptive feedback. These things are varied due to different speaking style, health status, or mental operation"*.

Landmark-based features can therefore offer unique potential to capture depression-related cues in speech articulation, which, to the best of our knowledge, have not been previously explored. 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 [36], [37], [38] and non-lexical attributes of speech such as syllabic complexity [45] and voice-onset time [46]. Recently, landmarks have been investigated for paralinguistic content, e.g. children's vocalization [47], emotion [48], Parkinson's disease and sleep deprivation [49]. In [48], landmark features were found to complement conventional acoustic features for emotion recognition, yet only three consonantal landmarks were investigated.

3 Landmark Features for Depression Characterization

3.1 Landmark Definitions

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

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 [50]. Examples of landmarks identified from speech can be seen in Figure 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	releases (+) or closures (–) of a nasal
f	frication onset (+) or offset (-)
v	voiced frication onset (+) or offset (-)
b	onset (+) or offset (-) of existence of turbulent noise during obstruent regions

We define a set of $L = 6$ landmarks, each with onset (+) and offset (-) states, i.e. $2L$ states in total:

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

and associated with a speech file, the sequence of identified landmarks is

$$W = \{w_i\}_{i=1}^{M} = \{w_1, w_2, \dots, w_i, \dots, w_M\}, \text{s.t.}\, w_i \in S \qquad (2)$$

where w_i represents the i^{th} landmark (i is a non-uniform time index), and M is the landmark count in a speech file. This type of representation is commonly referred to as a unigram in speech recognition and natural language processing.

3.2 Landmark Count Features

3.2.1 Landmark Count Features - n-gram Count

Beyond considering one landmark at a time (i.e. the unigram w_i), one can consider a sequence of n consecutive landmarks (referred to as n-grams, $\boldsymbol{w}_{i\to j}$). Accordingly, the sequence of identified n-grams for a speech file is:

$$W_n = \{\boldsymbol{w}_{i\to j}\}_{i=1}^{j=M}, \text{s.t.}\, j = i + n - 1, \qquad (3)$$

$$\boldsymbol{w}_{i\to j} = \{w_i, w_{i+1}, \dots, w_j\}, \text{s.t.}\, i \ge 1, j \le M, \qquad (4)$$

where $\boldsymbol{w}_{i\to j}$ is a sequence of n landmarks (n-grams) from the i^{th} to the j^{th} landmark, i and j are time indices satisfying $n = j - i + 1$. If $n = 1$, then $i = j$ and $W_{n=1}$ represents a sequence of unigrams as in (2).

In addition to $\boldsymbol{w}_{i\to j}$, we define $\boldsymbol{w}^{m,l}$ as a particular n-gram occurring in W_n:

$$\boldsymbol{w}^{m,l} = \{w_{i\to j} | w_{i\to j-1} = m, w_j = l\}, \text{s.t.}\, m, l \in S, \qquad (5)$$

where m and l denote a certain landmark or landmark sequence, in which each landmark belongs to the landmark set S. $\boldsymbol{w}^{m,l}$ specifies a certain landmark n-gram. For instance, if $m = \{g_+, p_+, p_-\}$ and $l = \{s_+\}$, then $\boldsymbol{w}^{m,l} = \{g_+, p_+, p_-, s_+\}$, as in Fig. 1. Given a sequence of n-grams for a speech file, the counts for all possible n-gram sequences are

$$\boldsymbol{C}_n = \#(W_n) \in \mathbb{R}^{(2L)^n}, \qquad (6)$$

where $\#(\cdot)$ is the counting operation applied to the whole speech recording. That is, all the possible unique n-grams $\boldsymbol{w}^{m,l}$ are counted, leading to:

$$\boldsymbol{C}_n = \{\#(\boldsymbol{w}^{m,l})\}_{\forall m\in S, l\in S} \tag{7}$$

$\boldsymbol{C}_n$ is referred to as the n-gram count. The n-gram count is expected to be large for frequently-occurring unique patterns of landmarks, which might be distinct between depressed and healthy speakers. Furthermore, the n-gram count will be related to the amount of speech produced in a speech file (i.e. a longer speech-active file length will increase the n-gram counts). It may be informative about timing differences during speech articulation, or for tightly specified speech elicitation tasks that are sensitive to the speed of articulation (e.g. PaTaKa). It was reported in [19] that the 2-gram count yields promising performance for depression classification on smartphone speech, but the effect of occurrence frequency was not investigated, and neither were the n-gram counts for n=1, 3, 4. Further, in some contexts (e.g. free speech tasks), it may make more sense to normalize for duration or occurrence.

3.2.2 Landmark Count Features - Normalization

Two normalization methods are proposed: the first taking into account time and the second focusing on the relative occurrence of different landmark transition types, for application to the n-gram count features.

The 'n-gram rate' can be obtained by dividing the n-gram count by the speech duration in seconds l_s, which varies for different files.

$$\bar{\boldsymbol{C}}_n = \frac{\boldsymbol{C}_n}{l_s} \tag{8}$$

The time-normalized n-gram rate $\bar{C}_n$ represents how often each n-gram is produced per unit time. According to the literature, timing information (such as phoneme rate, associated with psychomotor retardation [13], [18]) is discriminative for depression, and hence n-gram rate is expected to carry unique landmark-specific timing information associated with speech articulation for depressed speakers.

The 'n-gram probability' can be obtained by calculating the probability of transition from the previous (n-1)-grams to the last landmark (analogous to the HMM transition probability):

$$p_n(l|m) = \frac{\#(\boldsymbol{w}^{m,l})}{\#(\boldsymbol{w}^{m})}, \tag{9}$$

where $p_n(l|m)$ represents the transition probability from landmark sequence m to l. The counting operation is applied per speech file. The normalized n-gram probabilities are the concatenation of all the possible n-grams.

$$\tilde{\boldsymbol{C}}_n = \{p_n(l|m)\}_{\forall m\in S, l\in S} \tag{10}$$

The n-gram probability represents how diverse the range of landmark transitions is. The n-gram probability potentially characterizes changes from one landmark (sequence) to the other. It is expected that depressed speakers tend to have relatively constrained transitions; more constrained than those of healthy speakers whose voice can more easily produce a wider range of sounds [27], [29].

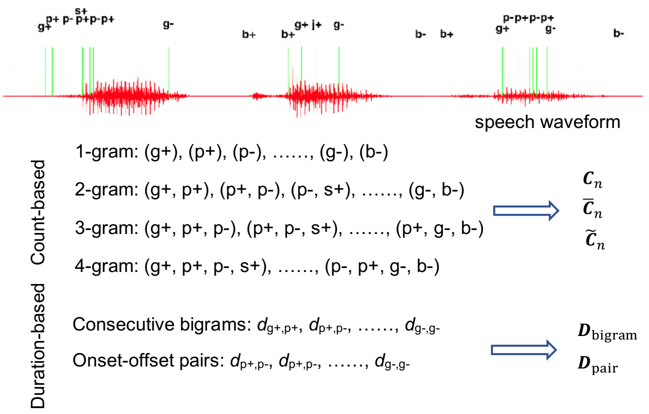

g+
p+
p-
s+
p+
p-
p+
g-
b+
b+
g+
j+
g-
b-
b+
g+
p-
p+
p-
p+
g-
b-
speech waveform
Count-based
1-gram: (g+), (p+), (p-), ……, (g-), (b-)
2-gram: (g+, p+), (p+, p-), (p-, s+), ……, (g-, b-)
3-gram: (g+, p+, p-), (p+, p-, s+), ……, (p+, g-, b-)
4-gram: (g+, p+, p-, s+), ……, (p-, p+, g-, b-)
$\boldsymbol{C}_n$
$\bar{\boldsymbol{C}}_n$
$\tilde{\boldsymbol{C}}_n$
Duration-based
Consecutive bigrams: $d_{g+,p+}$, $d_{p+,p-}$, ……, $d_{g-,g-}$
Onset-offset pairs: $d_{p+,p-}$, $d_{p+,p-}$, ……, $d_{g-,g-}$
$\boldsymbol{D}_{\text{bigram}}$
$\boldsymbol{D}_{\text{pair}}$

Fig. 1. Feature extraction of the landmark count-based and duration-based features from the word 'PaTaKa' uttered by male depressed speaker within 1.5 seconds.

The $\tilde{\boldsymbol{C}}_n$ calculates the probability from the (n-1)-gram to the last landmark. This can be extended to transition probabilities from the (n-2)-gram to the last 2-gram when $\boldsymbol{w}^{m,l} = \{w_{i,j}|w_{i,j-2} = m, w_{j-1\to j} = l\}$, and from the ($n$-3)-gram to the last 3-gram when $\boldsymbol{w}^{m,l} = \{w_{i,j}|w_{i\to j-3} = m, w_{j-2\to j} = l\}$. These (per-)'landmark normalized' counts are referred to herein as '*n-gram probabilities (1), (2), (3)*' respectively, for example in Fig. 3.

3.3 Landmark Duration Features

In addition to the proposed n-gram count-based features, this section explores the timing information embedded in landmarks further. More precisely, we propose duration-based features, which are statistics calculated from two types of bigrams, i.e. arbitrary consecutive bigrams and onset-offset pairs. Defining $t(w_i)$ as the time index of landmark w_i, the duration between two landmarks can be defined as

$$d_{i\to j} = t(w_j) - t(w_i). \tag{11}$$

Then durations associated with a certain type of consecutive bigrams or onset-offset pairs for a speech file are:

$$\boldsymbol{d}^{m,l} = \{d_{i\to j}|w_i = m, w_j = l\}_{i=1}^{j=M}$$
$$\text{s.t.} \begin{cases} j = i+1 & \text{if consecutive bigrams} \\ m = w_+, l = w_- & \text{if onset-offset pairs} \end{cases} \tag{12}$$

M again is the total landmark count for a speech file and the landmark $w \in \{g, p, s, f, v, b\}$. The condition $j = i+1$ limits the bigrams to be adjacent for the consecutive bigrams, whereas the condition $(m = w_+, l = w_-)$ restricts the bigrams to start from an onset state of a landmark and end with an offset state of the same landmark for the onset-offset pairs.

Statistics (functionals) then can be calculated from the durations $\boldsymbol{d}^{m,l}$:

$$d_k^{m,l} = |\boldsymbol{d}^{m,l}|_k \tag{13}$$

$$\boldsymbol{D}_{m,l} = \left[d_1^{m,l} \ldots \ d_k^{m,l} \ldots \ d_K^{m,l}\right]^T \tag{14}$$

where $|\cdot|_k$ represents a certain type of statistic (e.g. mean, standard deviation, percentiles, etc.), $d_k^{m,l}$ represents the kth type of statistic calculated from all durations $\boldsymbol{d}^{m,l}$ specific to each landmark bigram per file, and K is the total number of adopted statistics. Concatenating duration statistics for all possible arbitrary consecutive bigrams and all possible onset-offset pairs yields:

$$\boldsymbol{D}_{\text{bigram}} = \left[\boldsymbol{D}_{g_+,g_+}^{T} \dots \boldsymbol{D}_{m,l}^{T} \dots \boldsymbol{D}_{b_-,b_-}^{T}\right]^{T} \in \mathbb{R}^{(2L)^2 * K} \qquad (15)$$

$$\boldsymbol{D}_{\text{pair}} = \left[\boldsymbol{D}_{g_+,g_-}^{T} \dots \boldsymbol{D}_{w_+,w_-}^{T} \dots \boldsymbol{D}_{b_+,b_-}^{T}\right]^{T} \in \mathbb{R}^{L * K} \qquad (16)$$

Note that the dimensionality of $\boldsymbol{D}_{\text{bigram}}$ is much larger than that of $\boldsymbol{D}_{\text{pair}}$ ($4L$ times larger).

4 Experimental Settings

4.1 Databases

The experiments in this study were conducted on two datasets: The Distress Analysis Interview Corpus (DAIC-WOZ) dataset [6] and the SH2 dataset [9].

DAIC-WOZ is a laboratory-based dataset collected in scenarios where participants were interviewed by a virtual human agent named Ellie who asks all participants the same set of questions. The speech was recorded via the same high-quality close-talk microphones (i.e. fixed single channel) with minimum environmental background noise. Each interview produced up to 20 minutes of speech for each participant, and an accompanying binary label indicating whether the participant was depressed or healthy. The database has a large group of speakers, 189 speakers in total, which were divided into training (107 speakers), development (35 speakers) and test (47 speakers) partitions for the AVEC2016 and 2017 challenges. Further details of the DAIC-WOZ conventions can be found in [6]. The training and development partitions were adopted for training (with 3-fold cross validation) and testing respectively in this study.

SH2 is a subset of a large dataset collected by Sonde Health. It contains speech (sampled at 16kHz), along with device metadata and questionnaire data, from a general population sample in the United States under a human subject protocol reviewed and approved by an Institutional Review Board. All data were encrypted on-device and transmitted to secure cloud storage. Participants completed several voice tasks on their personal smartphones in uncontrolled natural environments, including free speech, read speech (Rainbow passage and Harvard sentences), and elicited tasks: sustained vowel “ahh” and diadochokinetic repetition. For example, participants were instructed to repeat a sentence from the Harvard Sentence database on the screen, or to freely respond for up to 30 seconds on a generic topic such as “What is the weather like outside?”.

SH2 contains around 16 hours of speech. The 5763 total audio files comprise six tasks (i.e. sustained vowel, diadochokinetic, free speech, rainbow passage, cognitive load and sentence), completed by 498 to 810 participants. The SH2 recordings were made from a wide variety of mobile device and smartphone manufacturers (28 in total). The SH2 corpus has the same training and testing partition as [9]: 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 partitions of ‘healthy’ (PHQ-9 < 10) and ‘depressed’ (PHQ-9 ≥ 10) speakers (as suggested by [51]), 122 depressed and 35 depressed speakers were respectively found in the training and test data partitions.

TABLE 2

Summary of the number of depressed and healthy subjects (#depressed/#healthy) for the adopted datasets. Note that in DAIC-WOZ, each speaker has one recording, whereas in SH2, each speaker has multiple recordings due to the different elicitation tasks. ‘CV’ means Cross Fold Validation for parameter search.

	DAIC-WOZ		SH2 (FS)		SH2	
	Train (CV)	Test	Train (CV)	Test	Train (CV)	Test
Male	8/55 (7.48hrs)	4/12 (1.78hrs)	33/193 (1.32hrs)	13/59 (0.41hrs)	52/291 (6.06hrs)	15/92 (1.92hrs)
Female	13/31 (5.30hrs)	3/16 (3.06hrs)	41/172 (1.24hrs)	10/46 (0.32hrs)	70/282 (6.48hrs)	20/65 (1.57hrs)
Total	21/86 (12.78hrs)	7/28 (4.84hrs)	74/365 (2.56hrs)	23/105 (0.73hrs)	122/573 (12.54hrs)	35/157 (3.49hrs)

Compared with DAIC-WOZ, which has clean, long-duration recordings from a single set of recording hardware, SH2 has a larger number of speakers, shorter durations, different smartphone recording hardware characteristics, and noisy naturalistic environments. For closer comparisons with DAIC-WOZ, the Free Speech (FS) portion of SH2 was selected for the investigation of the landmark count-based features (Section 5.1 and 5.2) and duration-based features (Section 5.3).

A summary of the adopted dataset and the training-testing partitions can be seen in Table 2. The average speech utterance durations are 20.5 ± 10.2s for SH2 (FS), 9.8 ± 6.86s for SH2, and 446.9 ± 227.0s for DAIC-WOZ.

4.2 Settings

All experiments in this study adopted a linear Support Vector Machine (SVM) classifier [52], for two main reasons: 1) it allows direct comparisons with our previous studies (i.e., [9], [19]), in which conventional acoustic features and linear SVM were used; 2) SVM exhibited good generalization and consistently strong performance for depression detection in our preliminary experiments. SVM was fine-tuned through parameter sweeps of the complexity coefficient C from 10^{-5} to 10 in log space. 3-fold cross validation was performed within the training data for parameter training, mainly the complexity coefficient C in linear SVM. Within each fold, the same percentages of the positive-negative class ratio were maintained. During training, C was weighted inversely proportionally to class frequencies to handle imbalanced training data for the healthy and depressed classes, as per [9], [19]. For both datasets, the best parameter configurations selected from the 3-fold cross validation were used to retrain a model on the whole training data and then to test on the test partition. Gender normalization, which applies z-normalization to dataset subsets specific to gender, was found to be effective in [9], [19] and was used throughout the following experiments. F1 score for the depressed class, which combines recall and precision, was used as the main

TABLE 3
THE NUMBER OF N-GRAMS THAT ACTUALLY OCCUR WITHIN THE TRAINING SET OF THE TWO DATASETS USED.

	Maximum Possible	DAIC	SH2 (FS)	SH2
1-gram	12	12	12	12
2-gram	144	73	66	73
3-gram	1728	275	218	297
4-gram	20736	764	555	859

metric. Additionally, F1 (healthy), classification accuracy and confusion matrices were calculated for final results.

The landmarks were extracted using the SpeechMark® toolbox [16], a publicly available, representative landmark extraction software. Note that *n*-grams that did not occur within the training data were removed from the *n*-gram list, leading to actual feature dimensions that were much smaller than the maximum possible (Table 3). The problem of unseen context is common in natural language processing, and smoothing methods were used. However, smoothing methods (e.g. add 1 smoothing) were tried and observed to have poorer performances and cause very high dimensionality when it comes to 3-gram or 4-gram.

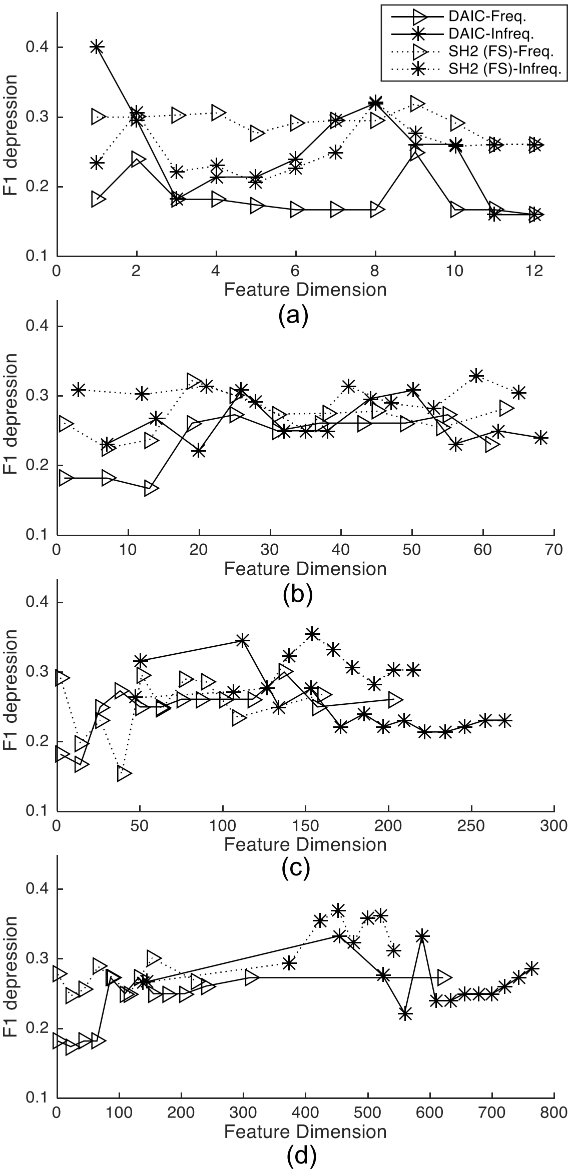
5. Results – Landmark Count-based and Duration-based features

5.1 Frequently-occurring vs. Infrequently-occurring *n*-grams

Since landmark *n*-gram features have not previously been explored and may have high dimensionality as seen in Table 3, a key question is whether it is advantageous to select particular feature subsets. If so, then it is interesting to determine whether frequently-occurring *n*-grams are more effective (since these will have more reliable count features), or whether infrequently-occurring *n*-grams are more effective (since these may be present or absent only in the depressed state).

To investigate this, *n*-grams firstly were sorted based on their occurrence frequency within the training data. Then, two sets of experiments were conducted: 1) starting with only the most frequently occurring *n*-grams for depression detection, gradually less frequently occurring *n*-grams were included until all possible *n*-grams were included. 2) Similarly, we repeated the process, but instead started from the most infrequently occurring *n*-grams. The F1 (depression) results can be seen in Fig. 2.

In Fig. 2, the infrequently occurring *N*-grams performed better than their frequently-occurring counterparts on SH2 (FS) when considering 2-gram, 3-gram and 4-gram, although this was not true for 1-gram. The benefits of infrequently occurring *n*-grams were more pronounced for the DAIC-WOZ dataset when considering 1-gram, 3-gram, 4-gram, especially for the first few points. This suggests that counts of only a few infrequently occurring landmark *n*-grams can offer good separation of the healthy and depressed classes. This is consistently observed for the two very different datasets.



(a)

(b)

(c)

(d)

Fig. 2. Evaluation of feature selection based on occurrence frequency for *n*-gram count features, for the DAIC (solid lines) and SH2 dataset (Free Speech) (dashed lines). "Freq." (represented by "Δ") means using the most frequently occurring *n*-gram counts, whereas "Infreq." (represented by "∗") means using the most infrequently occurring *n*-gram counts. *n* was set to (a) 1 (i.e. 1-gram), (b) 2 (i.e. 2-gram), (c) 3 (i.e. 3-gram) and (d) 4 (i.e. 4-gram).

5.2 Time- and Landmark-Normalization

This subsection evaluates the proposed *n*-gram count $\boldsymbol{C}_n$, *n*-gram rate $\overline{\boldsymbol{C}}_n$, and *n*-gram probability $\widetilde{\boldsymbol{C}}_n$ on both the DAIC-WOZ and SH2 (FS) datasets, as shown in Fig. 3.

There are several interesting findings from Fig. 3. Firstly, the time and landmark normalization yielded significant improvements for all *n*-grams on DAIC-WOZ, and for 3-gram and 4-gram on SH2 (FS). This suggests that timing and transition information is useful, especially for those relatively clean and long recordings on DAIC-WOZ. On one hand, it was expected that *n*-gram counts would be effective where depressed people say less than non-depressed people. On the other hand, for variable-duration

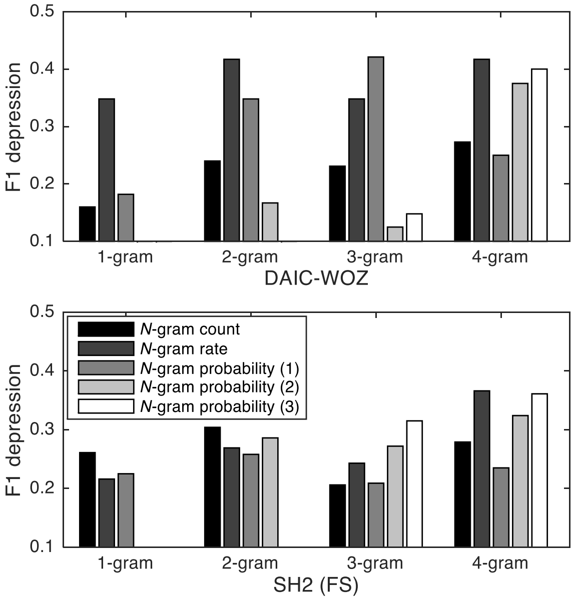


Fig. 3 Comparisons among *n*-gram count, rate, and probabilities. Three transition probabilities were considered, i.e. the transition probability from the *(n*-1*)*-gram to the last landmark, from the *(n*-2*)*-gram to the last 2-gram, from the *(n*-3*)*-gram to the last 3-gram, referred to as probability (1), probability (2) and probability (3) respectively. Note that the last bar for each 1-gram, 2-gram and 3-gram is self-transition, which will always be 1 once it occurs, regardless of its occurrence number.

utterances, it is beneficial to time-normalize them and look instead at the spread of different individual *n*-gram densities, which might spread more widely for non-depressed than depressed. The usefulness of the *n*-gram rate highlights the importance of speech articulation in detecting depression.

Secondly, *n*-gram probabilities performed better for larger *n* values, i.e. 3-gram and 4-gram than 1-gram and 2-gram, on both DAIC-WOZ and SH2 (FS). One possible reason is that transition probabilities for 3-gram and 4-gram is provide a fuller picture of the diversity of transitions than for 1-gram and 2-gram due to a larger set of possible transitions.

Overall, *n*-grams with $n > 1$ always provided better depression detection than unigrams, which shows the importance of modelling landmark *patterns* rather than just landmark densities. Furthermore, the DAIC-WOZ corpus tends to benefit more from using larger *n* than SH2 (FS), which may be due to the difference in speech duration between both datasets: DAIC-WOZ has longer-duration files that produce more sequences, whereas the shorter-duration SH2 (FS) files produce fewer sequences, especially for larger *n*. It is also worth noting that 3- and 4-gram of speech landmarks could contain linguistic content at a high level; however, it did not seem to aid system performance and it is expected that the 3- or 4-gram of speech landmarks primarily contains sequential information regarding speech articulation.

Even though *n*-grams with $n > 2$ sometimes gave improved detection accuracy than for $n = 2$ (Fig. 3), $n = 2$ might be recommended as the most parsimonious choice.

5.3 Consecutive vs. Onset-Offset Pairs

The *n*-gram rate demonstrates great effectiveness for detecting depression (Fig. 3), suggesting that it is significant to consider timing information from a sequence of landmarks. This, therefore, motivates our further investigations, in which timing information was explicitly exploited by considering durations of landmark sequences, mainly bigram (i.e. n=2) in this study. In particular, two forms of bigrams were investigated, i.e. arbitrary consecutive bigrams (15) and onset-offset pairs (16). Consecutive bigrams represent two adjacent bigrams that have no landmarks in between, whilst onset-offset pairs represent the closest pairs of the same landmark following the specific "onset → offset" pattern, as illustrated in Fig. 1.

To start with, statistical analysis was used to evaluate and quantify the significance of differences in durations of consecutive bigrams and onset-offset pairs between depressed speakers and healthy speakers. The landmark duration-based features were then used to detect depression for both DAIC-WOZ and SH2 (FS).

5.3.1 Statistical Analysis

This subsection statistically evaluates the usefulness of the durations of consecutive bigrams and onset-offset pairs. To be more specific, durations between healthy and depressed speakers were compared for both consecutive bigrams (shown in Table 4) and onset-offset pairs (shown in Table 5). Herein, the Mann-Whitney *U* test [53], a non-parametric statistical test approach, was utilized over parametric approaches (e.g. *t*-test) to avoid an assumption of Gaussian distributions over durations. It was observed that the duration distributions are skewed towards short lengths, and thus the assumption of normality is inappropriate for this context. Other non-parametric significance tests such as Wilcoxon signed-rank test and Kruskal-Wallis *H* test were also trialed, and similar results were obtained.

For consecutive bigrams (Table 4), 19 out of 73 bigrams showed $p < .05$ for DAIC-WOZ (mostly evaluated on 10,000 to 100,000 duration samples), whilst 9 out of 66 bigrams exhibited $p < .05$ for SH2 (FS) (mostly evaluated on 1,000 to 10,000 duration samples). Statistical analysis results in Table 5 show significant differences in durations of all onset-offset pairs on DAIC-WOZ (evaluated on from 1,000 to 150,000 duration samples). Significance was also found for all onset-offset pairs on the SH2 (FS) (evaluated on 800 to 25,000 duration samples), except for v+ → v-.

Table 4 lists pairs of consecutive landmarks with the most strongly significant discriminative properties. Among these, the durations of (g+, p+) and (s+, p-) are notable for their discriminative strength in both datasets.

The (g+, p+) pair represents consecutive transitions between onset of glottal activity and sustained periodic signal of significant strength. The *p* landmark provided by the SpeechMark software denotes regions where the fundamental frequency is between 70 and 350 Hz. As such, this (g+, p+) consecutive pair duration represents the settling time between air flow and proper fundamental frequency periodicity. This duration difference between depressed and non-depressed classes might suggest unusual control or function of the cricothyroid muscle, or

TABLE 4
P VALUES FROM MANN-WHITNEY U TEST ON TRAINING DATA PARTITIONS, COMPARING THE DISTRIBUTIONS OF CONSECUTIVE BIGRAM DURATIONS BETWEEN DEPRESSED AND HEALTHY SPEAKERS. THE TOP 10 MOST SIGNIFICANT/DISCRIMINATIVE CONSECUTIVE BIGRAMS ARE PRESENTED.

DAIC-WOZ		SH2 (FS)	
bigram	*p*-value	bigram	*p*-value
(s-, p-)	7.90×10^{-26}	**(s+, p-)**	2.70×10^{-3}
(g+, p+)	3.66×10^{-22}	(v-, p-)	3.59×10^{-3}
(g-, b-)	2.03×10^{-19}	**(g+, p+)**	1.01×10^{-2}
(b-, b+)	5.85×10^{-14}	(p+, p-)	1.13×10^{-2}
(g-, b+)	2.52×10^{-7}	(b+, b+)	1.87×10^{-2}
(p-, g-)	1.88×10^{-4}	(p+, s-)	2.20×10^{-2}
(s+, p-)	3.36×10^{-4}	(s+, v+)	3.35×10^{-2}
(b+, b-)	6.27×10^{-4}	(p+, s+)	3.71×10^{-2}
(b+, f+)	7.12×10^{-4}	(b-, b-)	4.53×10^{-2}
(v+, p-)	8.59×10^{-4}	(b+, b-)	5.15×10^{-2}

TABLE 5
P VALUES FROM MANN-WHITNEY U TEST ON TRAINING DATA PARTITIONS, COMPARING THE DISTRIBUTIONS OF ONSET-OFFSET PAIR DURATIONS BETWEEN DEPRESSED AND HEALTHY SPEAKERS.

	DAIC-WOZ	SH2 (FS)
p+ → p-	7.65×10^{-85}	4.29×10^{-8}
g+ → g-	3.74×10^{-41}	9.53×10^{-4}
s+ → s-	1.06×10^{-8}	4.86×10^{-3}
f+ → f-	2.34×10^{-5}	1.19×10^{-2}
v+ → v-	3.34×10^{-3}	1.56×10^{-1}
b+ → b-	6.80×10^{-2}	1.52×10^{-2}

improper subglottal pressure control [54]. This observation is aligned with findings that vocal fold vibrations become increasingly irregular due to higher tension and greater emotional stress [55], [56], hence the reported prominence of jitter features in spectral-based features. Our measurements between g+ and p+ in individuals with depression (in both datasets) show shorter average durations possibly emanating from over tensioning of vocal folds.

Durations between consecutive landmarks s+ and p- are also shown in Table 4 to be important discriminators in separating depression classes. The s+ (sonorant) landmark, measured by relative broadband power surges, denotes the release of a nasal or [40] and infers [+consonantal] and [+sonorant] articulator-free features. The fluctuations in duration between such closure and the end of periodicity (p-) once again suggest unusual control of either air flow or vocal fold tension release. Longer durations are observed on both data sets (SH2 and DAIC) for depressed individuals in transitioning between the sonorant (s+) release and the end of periodicity (p-). The visualization of speech waveform and spectrogram for (g+, p+) and (s+, p-) is shown in Fig. 4. The durations of (s+, p-) and (g+, p+) in Fig. 4 have been visually compared to those of a healthy speaker, showing that the depressed speaker has shorter durations for (s+, p-) and longer durations for (g+, p+).

Variations between periodic (p) onset/offset pairs durations shown on Table 5, as well as their glottal enclosures (g) appear most indicative of depressed and non-depressed populations. This may be due to differing onset/offset transitions between classes and supports the hypothesis of glottal fold control being a major discriminator of speech from depressed individuals [57]. These glottal onset/offset features are particularly salient due to the nature of landmarks and could complement other spectrum-based depression indicators.

5.3.2 Depression Detection using Duration-based Features

Motivated by the preceding statistical test results, duration-based features were evaluated in this subsection.

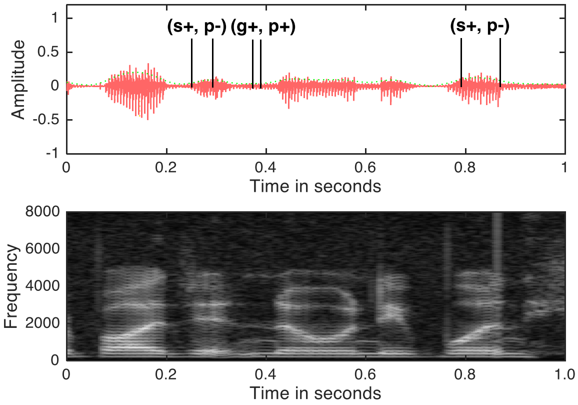


Fig. 4. Speech waveform of one second and its spectrogram from a depressed speaker (ID=321) in the DAIC-WOZ dataset, containing the consecutive bigrams (g+, p+) and (s+, p-), which were found to be significantly different between depressed and healthy speakers on both DAIC-WOZ and SH2 (FS).

More specifically, statistics were calculated from the durations of consecutive bigrams (i.e. $\boldsymbol{D}_{\mathrm{bigram}}$) and onset-offset pairs (i.e. $\boldsymbol{D}_{\mathrm{pair}}$) for each audio file, and used as features to classify depression and non-depression on both the DAIC-WOZ and the SH2 (FS) datasets.

17 functionals were trialed: mean, standard deviation, median, minimum, maximum, 10%, 20%, 30%, 40%, 60%, 70%, 80%, 90% percentiles, 40%-60% percentile range, 20%-80% percentile range, skewness, kurtosis. With these percentiles, the whole distribution of the durations can be roughly characterized into low (10%, 20%, 30% percentiles), medium (40%, 60% percentiles and median), and high (70%, 80%, 90% percentiles) regions, as well as the range within medium, and between the low and high regions. It is worth noting that the dimensionality of the onset-offset pairs (6 pairs × K) is much less than those of consecutive bigrams (66 bigrams × K).

It is expected that not all of the trialed functionals are useful, and more functionals might result in overfitting of training data due to larger dimensions. Thus, we selected up to 5 best functionals, each of which was searched individually in a way that the chosen functionals were progressively included in subsequent searches. 5 functionals were adopted since no improvements were observed with a higher number of functionals.

Table 6 summarizes the results in F1 (depression) and accuracy of landmark duration-based features for consecutive bigrams and onset-offset pairs on DAIC-WOZ

TABLE 6
DEPRESSION DETECTION USING DURATION-BASED FEATURES FOR ARBITRARY CONSECUTIVE BIGRAMS AND ONSET-OFFSET PAIRS ON THE DAIC DATASET AND SH2 (FREE SPEECH) CORPUS. “ALL STATS” MEANS USING ALL 17 FUNCTIONALS AS FEATURES, WHEREAS THE “BEST X” MEANS SELECTING BEST X STATISTICS OUT OF 17 STATISTICS FOR DEPRESSION CLASSIFICATION.

		DAIC-WOZ			SH2 (Free Speech)		
		F1 (D)	Accuracy	Chosen Stats.	F1 (D)	Accuracy	Chosen Stats.
Consecutive bigrams	All 17 stats.	0.364	60.0%	all	0.218	66.4%	all
	Best 1 stat.	0.571	74.3%	q70	0.302	53.1%	kt
	Best 2 stats.	0.737	85.7%	q70, mean	0.303	64.1%	kt, mean
	Best 3 stats.	0.588	80.0%	q70, mean, q30	0.300	67.2%	kt, mean, q60
	Best 4 stats.	0.667	80.0%	q70, mean, q30, q60	0.300	67.2%	kt, mean, q60, q70
	Best 5 stats.	0.667	80.0%	q70, mean, q30, q60, median	0.295	66.4%	kt, mean, q60, q70, median
Onset-offset pairs	All 17 stats.	0.211	57.1%	all	0.341	57.8%	all
	Best 1 stat.	0.526	74.3%	mean	0.328	64.8%	kt
	Best 2 stats.	0.526	74.3%	mean, q20-80	0.357	57.8%	kt, q40-60
	Best 3 stats.	0.600	77.1%	mean, q20-80, q70	0.370	60.2%	kt, q40-60, q40
	Best 4 stats.	0.600	77.1%	mean, q20-80, q70, q80	0.424	61.7%	kt, q40-6, q40, q90
	Best 5 stats.	0.571	74.3%	mean, q20-80, q70, q80, q30	0.414	60.2%	kt, q40-60, q40, q90, median

'kt'→ kurtosis, 'q30'→ 30% percentile, 'q40'→ 40% percentile, 'q60'→ 60% percentile,'q70'→ 70% percentile, 'q80'→ 80% percentile, 'q90'→ 90% percentile, 'q40-60'→ the range between 40%-60% percentile, 'q20-08'→ the range between 20%-80% percentile.

and SH2 (FS). Compared with all 17 functionals, using a few can achieve significantly higher results.

Several cross-corpus consistencies can be seen from Table 6: 1) for consecutive bigrams, mean, 60% percentile and 70% percentile are among the most useful functionals for DAIC-WOZ and SH2 (FS). 2) for onset-offset pairs, it is found that larger percentiles (i.e. 70%, 80%, and 90% percentiles), as well as the percentile ranges tend to be more useful than other percentiles for both datasets. 3) F1 (depression) scores plateau at 4-best functionals, suggesting that including more functionals did not seem to aim detection of depression.

Besides the cross-corpus consistencies, there are unique patterns within each corpus. Interestingly, consecutive bigrams achieve much better performances than onset-offset pairs for DAIC-WOZ, whereas this is the other way around for SH2 (FS). The reason may be two-fold: 1) DAIC-WOZ has a larger number of statistically significant bigrams than SH2 (FS) (19 vs 9 in Table 4); 2) DAIC-WOZ has longer speech durations than SH2 (FS), which result in more reliable estimates of the distributions of speech duration.

Taken together, these findings reinforce the usefulness and effectiveness of duration-based features, especially for large percentiles, in detecting depression.

6 RESULTS – TASK-SPECIFIC ANALYSIS (SH2)

It was found in [19] that landmark bigram counts calculated from tailored landmarks were effective in exploiting various elicitation tasks for detecting depression. This implies the importance and merit of task-wise analysis of landmarks, since distinct articulatory aspects elicited by the tasks can be captured by selecting the best landmarks for each task to produce count and duration features. To this end, besides free speech, which have been investigated on DAIC-WOZ and SH2(FS) in previous sections, this section evaluates the proposed count-based and duration-based features for other five tasks on the SH2 corpus, namely diadochokinetic (‘Word’), rainbow passage (‘Passage’), cognitive load (‘CL’), sustained vowels (‘Utterance’), and Harvard sentences (‘Sentence’).

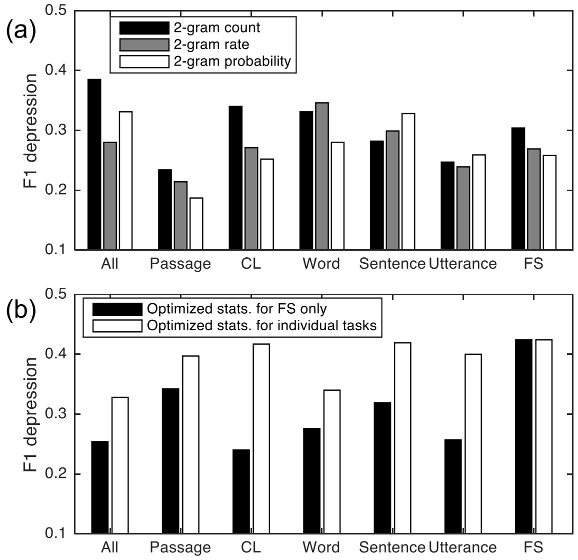


Fig. 5. Evaluation of the proposed (a) count-based and (b) duration-based features for various elicitation tasks on the SH2 corpus. The upper plot is comparison of the proposed count-based features, and the lower plot used four best four statistics (the black bar) and up to three task-specific statistics (the white bar) of onset-offset pairs. The tailored statistics were ['mean', 'q60'] for Passage, ['std', 'q10'] for Cognitive Load, ['max', 'skew'] for Diadochokinetic (‘Word’), ['std', 'q90', 'q20-80'] for Harvard Sentence, ['mean', 'q70', 'q40-60'] for Sustained Vowel, and ['skew'] for the whole SH2 dataset.

According to Fig. 3, n in (3) was set to 2, leading to 2-gram count, rate and probability. For duration, onset-offset pairs with four statistics optimized for free speech (i.e. q40, q40-60, q90, and kurtosis) were evaluated (Table 6). However, using the same statistics for different tasks might

TABLE 7
Optimized and fused results for DAIC-WOZ. Within the brackets {} is feature dimensionality.

	F1(D)	F1(H)	Acc.	Conf. Mat.
Audio (eGeMAPS)	0.29	0.82	71.4%	$\begin{bmatrix}23 & 5\\ 5 & 2\end{bmatrix}$
AVEC 2016 Baseline (A) [12]	0.41	0.58	51.4%	$\begin{bmatrix}12 & 16\\ 1 & 6\end{bmatrix}$
AVEC 2016 Baseline (A+V) [12]	0.58	0.86	77.1%	$\begin{bmatrix}22 & 6\\ 2 & 5\end{bmatrix}$
Ensemble (A+V+T+G) [58]	0.62	0.91	-	-
Audio (A) [59]	0.59	0.87	80.0%	$\begin{bmatrix}23 & 5\\ 2 & 5\end{bmatrix}$
Video+FeatS (V) [60]	0.63	0.89	82.9%	$\begin{bmatrix}24 & 4\\ 2 & 5\end{bmatrix}$
Winner (A+V+T+G) [61]	0.86	0.97	94.3%	$\begin{bmatrix}27 & 1\\ 1 & 6\end{bmatrix}$
2-gram rate (C) {73}	0.42	0.70	60.0%	$\begin{bmatrix}16 & 12\\ 2 & 5\end{bmatrix}$
Duration (D)[1] {332}	**0.86**	**0.97**	**94.3%**	$\begin{bmatrix}\mathbf{27} & \mathbf{1}\\ \mathbf{1} & \mathbf{6}\end{bmatrix}$

A-Audio, V-Video, T-Text, G-Gender Information, FeatS – Feature Selection.

TABLE 8
Optimized and fused results for SH2 (FS) and SH2. Note that results for 'Lmk. Bigram' from [16] requires optimal landmark choices, whereas results in this study used all the landmarks. Within the brackets {} is feature dimensionality.

		F1(D)	F1(H)	Acc.	Conf. Mat.
SH2 (FS)	Acoustic (eGeMAPS)	0.323	0.784	67.2%	$\begin{bmatrix}76 & 29\\ 13 & 10\end{bmatrix}$
	Acoustic [9]	0.333	0.739	62.5%	$\begin{bmatrix}68 & 37\\ 11 & 12\end{bmatrix}$
	Lmk. Bigram [19] (C)	0.353	0.678	57.0%	$\begin{bmatrix}58 & 47\\ 8 & 15\end{bmatrix}$
	4-gram rate (C) {555}	0.366	0.757	64.8%	$\begin{bmatrix}70 & 35\\ 10 & 13\end{bmatrix}$
	Duration (D)[2] {30}	0.474	0.778	68.8%	$\begin{bmatrix}70 & 35\\ 5 & 18\end{bmatrix}$
	C+D – equal weights	0.417	0.772	67.2%	$\begin{bmatrix}71 & 34\\ 8 & 15\end{bmatrix}$
	C+D – optimal weights	**0.533**	**0.807**	**72.7%**	$\begin{bmatrix}\mathbf{73} & \mathbf{32}\\ \mathbf{3} & \mathbf{20}\end{bmatrix}$
	Acoustic+C+D	0.459	0.78	68.8%	$\begin{bmatrix}71 & 34\\ 6 & 17\end{bmatrix}$
SH2	Acoustic (eGeMAPS)	0.267	0.667	54.2%	$\begin{bmatrix}88 & 69\\ 19 & 16\end{bmatrix}$
	Acoustic [9]	0.396	0.799	69.8%	$\begin{bmatrix}115 & 42\\ 16 & 19\end{bmatrix}$
	Lmk. Bigram [19] (C)	0.433	0.808	71.4%	$\begin{bmatrix}116 & 41\\ 14 & 21\end{bmatrix}$
	2-gram count (C) {73}	0.385	0.756	65.1%	$\begin{bmatrix}104 & 53\\ 14 & 21\end{bmatrix}$
	Duration (D)[3] {318}	0.394	0.683	58.3%	$\begin{bmatrix}86 & 71\\ 9 & 26\end{bmatrix}$
	C+D – equal weights	0.418	0.766	66.7%	$\begin{bmatrix}105 & 52\\ 12 & 23\end{bmatrix}$
	C+D – optimal weights	0.432	0.769	67.2%	$\begin{bmatrix}105 & 52\\ 11 & 24\end{bmatrix}$
	Acoustic+C+D	**0.466**	**0.804**	**71.4%**	$\begin{bmatrix}\mathbf{113} & \mathbf{44}\\ \mathbf{11} & \mathbf{24}\end{bmatrix}$

be suboptimal, and therefore a search of up to three best statistics was conducted to tailor discriminative statistics for each elicitation task. The F1 (depression) scores for both count-based and duration-based features can be found in Fig. 5.

In Fig. 5(a), the 2-gram counts outperformed the 2-gram rate and probability for the tasks Passage, CL, FS, whereas for Word and Sentence, the 2-gram rate provided marginal improvements over 2-gram counts. Moreover, it was found that using all the tasks performed better than individual tasks.

Some observations can be made about the relative performance of the normalization approaches between the different task types in Fig. 5(a). For the read (Harvard) sentences task, where the same landmark counts would be expected across all speakers, 2-gram rate and 2-gram probability were more discriminative, as would be expected. For time-sensitive tasks such as Word (speakers must say "PaTaKa" as many times as possible in five seconds), the 2-gram rate was most informative, as expected. For the CL (Stroop) task, response accuracy may be a key factor, and for this the raw 2-gram count was more effective than normalized 2-gram features. The differences between the rainbow passage and Harvard sentences tasks very likely reflect the higher lexical difficulty and number of unfamiliar words in the rainbow passage.

In Fig. 5(b), task-specific results using statistics optimized for FS were compared with those using tailored statistics for individual tasks. The tailored statistics consistently yielded improved performance in F1 scores over those using the same statistics. This is not surprising, since the best for FS is not necessarily optimal for other tasks. The chosen statistics allow task-specific adjustability and interpretability as to articulatory events. For instance, variability of durations is beneficial for the CL task, because depressed speakers tend to have more variability due to cognitive impairment. Also, for most tasks, high-percentile regions of the duration distributions such as q60, q70, q90, q20-80, max, were consistently found important, exemplified by the Sentence and Utterance tasks. This suggests that depressed speakers tend to produce longer durations in general.

Overall, differently from [19], where landmark choices were optimized, experimental results suggest that the different articulatory aspects elicited by different tasks can be captured using different statistics of onset-offset durations.

7 Optimized and Fused System Results

This section presents optimized results for landmark count-based and duration-based features on the DAIC-WOZ, SH2 (FS) and SH2 corpora, in comparison with published results in the literature. A widely used reference feature set, eGeMAPS [62], was also examined for comparison. The 88-dimensional eGeMAPS features were extracted per audio file. Furthermore, we examined two fusion schemes at decision levels to study whether the proposed count and duration features are complementary, and moreover, whether the proposed landmark-based features are complementary to acoustic features. The first fusion scheme linearly combines

[1] Note that this performance can also be achieved by other combinations of five statistics such as 1) std, q10, q20, q90, kurtosis; 2) std, q10, q40, q90, kurtosis; 3) std, q20, q80, q90, kurtosis.

[2] The chosen statistics were median, maximum, skewness, q90, and q40-60, calculated from onset-offset pairs.

[3] The chosen statistics were mean, minimum, q70, q20-80, and kurtosis, calculated from the consecutive bigrams.

SVM scores (distance to the optimum hyperplane of the trained model) before calculating the 'sign' for binary decisions. The second fusion scheme fuses duration, count, and acoustic features via majority voting of the binary decisions from each individual system for each speech file. F1 for depression (D), F1 for healthy(H), accuracy and Confusion matrix (Conf. Mat.) are summarized in Table 7 for DAIC-WOZ, and Table 8 for SH2 (FS) and SH2.

In Table 7, the chosen optimized system for count was 2-gram rate (i.e. with time normalization), which outperformed the AVEC 2016 audio baseline, 0.56 vs 0.50 in mean F1 scores. The chosen optimized system for duration was 5 statistics calculated from consecutive bigrams, i.e. mean, standard deviation, q70, q90, and kurtosis. This feature set achieved the state-of-the-art result, 94.3% accuracy and 0.92 mean F1 score, outperforming the published results in literature, and matching those of the AVEC 2016 depression challenge winning submission [61]. However, note that the authors in [61] employed gender information (by optimizing features and models per gender), PHQ-8 sub-symptom scores (metadata), and multiple modalities: audio, video, text, and emotional cues. Furthermore, the achievement of the landmark duration features is significant, because exploiting the audio modality for DAIC has been more challenging than other modalities, and the text modality has dominated the performances on this particular dataset [63], [58], [60]. It is also worth noting that the duration feature set achieved 100% accuracy on the training partition, correctly classifying 21 depressed and 86 healthy speakers. As a result, fusion was not found to yield further gains.

Table 8 summarizes the optimized and fused results on SH2 (FS) and SH2. The baselines are published results using acoustic features [9] and landmark bigram counts (with tailored landmark choices) [19]. The acoustic features in [9] adopted 8 functionals (i.e. mean, std, median, q20, q80, q20-80, skewness, and kurtosis) of the 38-dimensional IS2010 low level descriptors [64]. For SH2(FS), the duration system used the best 5 statistics of the onset-offset pairs, namely median, max, skewness, 90% percentile, and the 40-60% percentile range, whereas the count system used 4-gram rate, according to Fig. 3(a). Both the duration and count systems improved upon previous systems that use either acoustic or bigram counts, especially for the duration system (0.474 in F1). The significance of duration features on SH2(FS) concurs with DAIC-WOZ, suggesting the effectiveness of landmark durations for depression detection using free speech. Fusing the two systems using a linear combination of SVM scores (the weights were 0.8 and 0.2 for the duration and count systems respectively) yielded further significant gains, achieving 0.533 F1 scores and 72.7% accuracy.

However, for SH2, the count and duration systems achieved 0.385 and 0.394 in F1 scores for depression respectively, outperformed by the baselines, which were 0.396 and 0.433. The reason why the duration system did not perform as well as on DAIC-WOZ and SH2(FS) was observed to be that SH2 comprises six different tasks, each of which has distinct sets of statistically significant consecutive bigrams or onset-offset pairs. These task-wise significances were undermined by merging different tasks and adopting the same bigrams or onset-offset pairs to train a single classifier. Despite this, fusion of the count system (whose SVM scores were weighted by 0.4) and the duration system (whose SVM scores were weighted by 0.6) aided system performance, achieving 0.432 in F1, which is on par with [19]. Even better performances can be attained by fusing with the acoustic baseline system, achieving 0.466 in F1 scores.

Taken together, the proposed landmark-based features have shown state-of-the-art performance, achieving mean F1 scores of 0.92, 0.77, 0.64 for DAIC-WOZ, SH2 (FS), and SH2 respectively. These results outperformed published results as well as the reference eGeMAPS feature set on three different datasets. Moreover, Table 7 and Table 8 convey several important insights: 1) the landmark count and especially duration features are effective for depression detection on DAIC-WOZ and SH2(FS), two dramatically different datasets; 2) the count features and durations features are complementary; 3) the proposed landmark-based features are complementary to acoustic systems; 4) the choices of consecutive bigrams and onset-offset pairs need to be tailored for elicitation tasks.

The difference in performance between DAIC-WOZ and SH2 is also notable and could be due to a few reasons: 1) DAIC-WOZ has clean speech collected from a high-quality microphone, whereas SH2 contains noisy speech collected in the wild from smartphones with diverse channel characteristics. As a result, depression detection on SH2 is more challenging, since noise and channel diversity could undermine depression-related features; 2) DAIC-WOZ has longer speech durations than SH2 (FS), which result in more reliable estimates of the distributions of speech duration. Despite the difference, the mean F1 of 0.77 and 0.64 are among the best performance in the literature on SH2 (FS) and SH2.

It is acknowledged that a potential limitation of this study is the mere use of linear SVM as the back-end classifier. For this, we further examined the best proposed duration-based landmark features (Table 7 and 8) on all three datasets using two additional classifiers that consider non-linearity, i.e. a random forest classifier and a neural network classifier. However, the two classifiers were outperformed by linear SVM, which is primarily due to the duration-based features being optimized for linear SVM. Although it is expected that in general, other classifiers could produce improved performances by optimizing the choices of statistics, this is beyond the focus of the paper (i.e. the proposed landmark-based features), and therefore considered as future work.

8 Conclusion

The massive and growing societal burden imposed by depression necessitates automatic screening of depression via human voice, a non-invasive, readily accessible behavioral signal, for early detection and treatment. The smartphone represents a key opportunity: it has become a major tool for daily tasks and can reach a very wide spectrum of users. To cope with the challenges of finding effective depression-related features, especially for degraded recording conditions and diverse smartphones, herein we proposed two novel, effective sets of features based on speech landmarks, which delivered promising results on two dramatically different datasets - DAIC-WOZ (clean speech collected from a high-quality microphone) and SH2 (noisy speech collected in the wild from smartphones). Speech landmarks are time markers indicating important abrupt changes in speech articulation, so features developed from the speech landmarks can exploit useful articulatory information for depression detection. It is therefore expected that distributions of landmark-based features can capture a wide range of information related to articulatory malfunctions affected by depression, and could still function relatively effectively with absence of a few not always occurring symptoms such as psychomotor retardation.

For *n*-gram count-based features, two normalization methods were proposed to capture different information from the raw counts – time normalization and landmark normalization – and these *n*-grams were investigated across different values of *n*. There were three main findings: 1) modelling landmark *patterns* (i.e. *n*>1) is more useful than unigrams; 2) rarely occurring landmarks are more important than their frequently occurring counterparts; 3) the usefulness of timing information in speech articulation was highlighted by the strong performances that *n*-gram rates produced.

As for duration-based features, durations of consecutive bigrams and onset-offset pairs were found to be statistically significant for depression characterization across two datasets. Statistics of durations achieved promising results, coupled with a number of interesting findings: 1) higher percentiles of duration distributions tend to be more discriminative between depressed and healthy speakers, which was consistent across two drastically different datasets - DAIC-WOZ and SH2(FS); 2) speech recording length matters, because the distribution for durations can be more fully characterized from longer durations, exemplified by comparisons of DAIC and SH2 (FS); 3) four statistics are sufficient to encode critical information from the durations to achieve good results; 4) durations of consecutive bigrams were more effective than onset-offset pairs on a clean dataset (DAIC-WOZ), which however was reversed for SH2(FS), where the onset-offset pairs achieved better results. Further to the above findings, the landmark framework is inherently interpretable, revealing insights into what aspects of speech articulation were affected by depression.

Last but not least, when systems were optimized for the proposed count-based (by selecting the best *n* and the best normalization method) and duration-based features (by selecting the best five statistics), state-of-the-art performances were achieved on both DAIC-WOZ and SH2(FS). With duration features alone and a linear SVM classifier, 94.3% accuracy was achieved, better than nearly all published results that nonetheless adopted features based on multiple modalities such as audio, video, text with additional consideration of gender information, emotional cues and PHQ8 sub-symptoms. The strong performance of duration features also extended to SH2 (FS), a dataset composed of speech collected in noisy environments and various handsets via smartphones, giving 0.474 in F1 score, compared with 0.333 and 0.323 using acoustic features. Further, fusion of the optimized count and duration systems, and fusion of optimized count, duration, and acoustic systems reveal an important finding that not only are the landmark-based systems complementary to the acoustic systems (which was expected), but also the count and duration systems are complementary to each other.

Overall, the proposed novel landmark-based features are very promising for depression detection on not only speech collected in clean environment and a single channel, but also for speech collected in the wild (noisy environments and distinct handsets) from smartphones. Note that this kind of approach is not limited to only depression detection; potentially other affective computing problems (e.g. emotion, other mental disorders, etc.), and more broadly, speaker traits associated with speech articulation can be well targeted with the proposed features.

Besides the findings, the landmark framework opens up a wide range of exciting possibilities for future work. First, the landmark processing framework is event-based, which is akin to text analysis; each landmark can be viewed as one "word". For instance, Latent Dirichlet Allocation was used effectively to derive useful latent articulatory events from bigram counts [19]. Hence, further application of text analysis methods would extract more useful features. Moreover, event-based features like these might be productive to explore for other affective computing applications, and event-based features (conventionally used little perhaps because they haven't mapped straightforwardly to numerical feature vectors), across *all* modality types might be effectively harnessed using the methods introduced here. Second, these features can be viewed as pseudo-linguistic; they do contain linguistic information, but only at a very high level. It is expected that they might contain enough linguistic information to detect the key articulation events and transition types, but not so much as to make the features language-specific. The language dependency of the landmark-based features is interesting and could be examined on new languages in the future. Third, landmark-based features can be easily adapted to each elicitation task to aid depression detection, so further investigations into designing task-specific features are worth exploration, which might also shed light on how articulatory aspects are best elicited by different tasks. Fourth, the effect of recording duration and background noise can be examined for further understanding and improvement in performance. Fifth, more sophisticated back-end classifiers can be examined to exploit proposed landmark-based features for improved performances.

Acknowledgment

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

REFERENCES

[1] N. Cummins, 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.

[2] A. Pampouchidou *et al.*, “Automatic Assessment of Depression Based on Visual Cues: A Systematic Review,” *IEEE Transactions on Affective Computing*, pp. 1–27, 2017.

[3] S. Alghowinem *et al.*, “Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors,” *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 1–1, 2016.

[4] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, “Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2018.

[5] J. Walker *et al.*, “The Prevalence of Depression in General Hospital Inpatients: A Systematic Review and Meta-Analysis of Interview Based Studies,” *Psychological Medicine*, 2018.

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

[7] D. Ben-Zeev, 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.

[8] J. F. Cohn, 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.

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

[10] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. M. Provost, “The PRIORI emotion dataset: Linking mood to emotion detected in-the-wild,” *INTERSPEECH*, vol. 2018-Septe, pp. 1903–1907, 2018.

[11] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, “Detecting depression: A comparison between spontaneous and read speech,” in *ICASSP*, 2013, pp. 7547–7551.

[12] M. Valstar, 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.

[13] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, “Vocal biomarkers of depression based on motor incoordination,” in *Proceedings of the 4th ACM International Workshop on AVEC, ACM MM*, 2013, pp. 41–47.

[14] 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,” *ICASSP*, pp. 970–974, 2014.

[15] Z. S. Syed and D. Marshall, “Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation,” no. 2, pp. 37–43, 2017.

[16] J. R. Parker, *Algorithms for Image Processing and Computer Vision (2nd edition)*, no. 1. New York, NY, USA: John Wiley & Sons, Inc., 2011.

[17] Y. Yang, C. Fairbairn, and J. F. Cohn, “Detecting depression severity from vocal prosody,” *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2013.

[18] A. C. Trevino, 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.

[19] Z. Huang, J. Epps, and D. Joachim, “Speech Landmark Bigrams for Depression Detection from Naturalistic Smartphone Speech,” in *ICASSP*, p. 5856-5860, 2019.

[20] A. Michael *et al.*, “Emotional bias and inhibitory control processes in mania and depression,” *Psychological Medicine*, vol. 29, no. 6, pp. 1307–1321, 1999.

[21] J. Williamson, 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.

[22] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, “Voice acoustical measurement of the severity of major depression,” *Brain and Cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[23] H. Ellgring and K. R. Scherer, “Vocal indicators of mood change in depression,” *Journal of Nonverbal Behavior*, vol. 20, no. 2, pp. 83–110, 1996.

[24] J. D. Teasdale, S. J. Fogarty, and J. M. G. Williams, “Speech rate as a measure of short-term variation in depression,” *British Journal of Social and Clinical Psychology*, vol. 19, no. 3, pp. 271–278, 1980.

[25] 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,” *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[26] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP – A Collaborative Voice Analysis Repository for Speech Technologies,” in *IEEE ICASSP*, 2014, pp. 960–964.

[27] S. Scherer, L. P. Morency, J. Gratch, and J. Pestian, “Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis,” *ICASSP*, pp. 4789–4793, 2015.

[28] S. Scherer, G. M. Lucas, J. Gratch, A. Rizzo, and L. P. Morency, “Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews,” *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2016.

[29] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, “Analysis of acoustic space variability in speech affected by depression,” *Speech Communication*, vol. 75, pp. 27–49, 2015.

[30] J. R. Williamson and B. Helfer, “Segment-Dependent Dynamics in Predicting Parkinson’ s Disease,” in *INTERSPEECH*, 2015, no. September.

[31] B. Stasak 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.

[32] F. Gravenhorst *et al.*, “Mobile phones as medical devices in mental disorder treatment: an overview,” *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 335–353, 2015.

[33] J. Gideon, E. M. Provost, and M. Mclnnis, “Mood State Prediction from Speech of Varying Acoustic Quality for Individuals with Bipolar Disorder,” *Pediatr Neurol*, vol. 52, no. 6, pp. 566–584, 2016.

[34] F. Or, J. Torous, and J.-P. Onnela, "High potential but limited evidence: Using voice data from smartphones to monitor and diagnose mood disorders.," *Psychiatric Rehabilitation Journal*, vol. 40, no. 3, pp. 320–324, 2017.

[35] J. Slifka, K. N. Stevens, S. Manuel, and S. Shattuck-Hufnagel, *A Landmark-based Model of Speech Perception: History and Recent Developments*. 2004.

[36] C. Park, "Consonant Landmark Detection for Speech Recognition," *PhD Thesis, MIT, USA*, 2008.

[37] K. N. Stevens, "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.

[38] M. Hasegawa-johnson *et al.*, "Landmark-Based Speech Recognition : Report of the 2004 Johns Hopkins Summer Workshop," in *ICASSP*, 2005, pp. 213–216.

[39] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a Model for Lexical Access based on Features," in *ICSLP*, 1992, pp. 499–502.

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

[41] R. Jakobson, C. G. Fant, and M. Halle, "Preliminaries to Speech Analysis: the Distinctive Features and their Correlates," *Cambridge, MA: MIT Press*, 1952.

[42] K. Stevens and S. Keyser, "Retrieving distinctive features and segments from variable acoustic cues," *The Journal of the Acoustical Society of America*, vol. 108, no. 5, pp. 2627–2628, 2000.

[43] and S. J. K. Clements, George N., "Cv phonology: a generative theory of the syllabe," *Linguistic Inquiry Monographs Cambridge, Mass*, pp. 1–191, 1983.

[44] J. J. McCarthy, "Feature geometry and dependency: A review," *Phonetica*, vol. 45, no. 2–4, pp. 84–108, 1988.

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

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

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

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

[49] K. Ishikawa, 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. 441–447, 2017.

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

[51] K. Kroenke, 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.

[52] R.-E. Fan, 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.

[53] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, pp. 50–60, 1947.

[54] J. J. Jiang and C. Tao, "The minimum glottal airflow to initiate vocal fold oscillation," *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 2873–2881, 2007.

[55] S. E. Silverman, D. M. Wilkes, R. G. Shiavi, A. Ozdas, and M. K. Silverman, "Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[56] K. R. Scherer, "Vocal affect expression: a review and a model for future research.," *Psychological bulletin*, vol. 99, no. 2, pp. 143–65, Mar. 1986.

[57] E. Moore, M. A. Clements, J. W. 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, 2008.

[58] A. Pampouchidou *et al.*, "Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text," *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 27–34, 2016.

[59] Z. Huang *et al.*, "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.

[60] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, 2016.

[61] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision Tree Based Depression Classification from Audio Video and Language Information," *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 89–96, 2016.

[62] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[63] J. R. Williamson *et al.*, "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.

[64] B. Schuller *et al.*, "The INTERSPEECH 2010 Paralinguistic Challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

Zhaocheng Huang (S'15, M'18) received his B.Sc. degree from Harbin Engineering University, China in 2013, and Ph.D. degree from The University of New South Wales (UNSW Sydney), Australia in 2018. He is currently a Post-Doctoral Fellow at UNSW Sydney, working on automatic assessment of mental disorders from smartphone speech. His research interests include speech signal processing, machine learning, affective computing with a focus on speech-based emotion (change) recognition and depression detection. Dr. Huang is a member of the IEEE Signal Processing Society and ISCA.

Julien Epps (M’97) received the B.E. and Ph.D. degrees from the University of New South Wales, Sydney, Australia, in 1997 and 2001, respectively. He was a Post-Doctoral Fellow at the University of New South Wales. From 2002 to 2004, he was a Senior Research Engineer with Motorola Labs, where he was engaged on speech recognition. From 2004 to 2006, he was a Senior Researcher with National ICT Australia, Sydney. He then joined the UNSW School of Electrical Engineering and Telecommunications, New South Wales, Australia, in 2007, as a Senior Lecturer, and is currently a Professor. He is also a Contributed Researcher at Data61, CSIRO, Australia. Dr. Epps has authored or co-authored around 200 publications and serves as an Associate Editor for *IEEE Trans. Affective Computing*. His current research interests include characterization, modelling, and classification of mental state from behavioral signals, such as speech, eye activity and head movement.

Dale Joachim received the M.S. and Ph.D. degrees in electrical engineering from Michigan State University, East Lansing, in 1994 and 1998, respectively. He previously served as technical staff at MIT Lincoln Laboratory, visiting faculty at MIT Media Lab, assistant professor at Tulane University as well as principal investigator of research and development projects at BAE Systems. He is currently the director of speech science at Sonde Health, leading the discovery of health related vocal bio-markers.