This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing

Natural Language Processing Methods for Acoustic and Landmark Event-based Features in Speech-based Depression Detection

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

***Abstract*— The processing of speech as an explicit sequence of events is common in automatic speech recognition (linguistic events), but has received relatively little attention in paralinguistic speech classification despite its potential for characterizing broad acoustic event sequences. This paper proposes a framework for analyzing speech as a sequence of acoustic events, and investigates its application to depression detection. In this framework, acoustic space regions are tokenized to 'words' representing speech events at fixed or irregular intervals. This tokenization allows the exploitation of acoustic word features using proven natural language processing methods. A key advantage of this framework is its ability to accommodate heterogeneous event types: herein we combine acoustic words and speech landmarks, which are articulation-related speech events. Another advantage is the option to fuse such heterogeneous events at various levels, including the embedding level. Evaluation of the proposed framework on both controlled laboratory-grade supervised audio recordings as well as unsupervised self-administered smartphone recordings highlight the merits of the proposed framework across both datasets, with the proposed landmark-dependent acoustic words achieving improvements in F1(depressed) of up to 15% and 13% for SH2-FS and DAIC-WOZ respectively, relative to acoustic speech baseline approaches.**

***Index Terms*—event-based features, acoustic features, speech-based depression detection, landmarks, heterogeneous speech features, paralinguistic speech processing**

I. Introduction

DEPRESSION is one of the greatest current challenges to world productivity [1], [2]. Although it has been shown that early detection of depression can dramatically increase the prevalence of individual return to normal function, effective and feasible methods for depression screening remains elusive. Speech-based technologies have however shown great potential, with key advantages being noninvasiveness and broad accessibility [1], [3], [4], [5]. These findings coupled with the increasing potential of speech-based analysis through smartphones have opened up new possibilities for cost-effective depression screening in real-life scenarios [6], [7], [8]. Smartphone audio-based applications, however, must overcome many challenges such as noisy naturalistic environments and significant variability in handset characteristics, etc [9], [10], [11].

The variability of voice quality in unsupervised environments creates significant challenges for conventional acoustic features (e.g. spectral, glottal, and prosodic features). As a result, acoustic features may experience degraded performance due to the sensitivity to environmental noise and handset channel variability, compared with performances achieved on laboratory-grade clean speech data. Recently, speech landmarks, which are events associated with speech articulation, have shown effectiveness for screening depression in such naturalistic environments [10], [12]. This paper focuses partly on speech landmarks, which occur at times of abrupt articulatory changes such as consonant closures/releases and nasal closures/releases [13].

The success of event-based speech landmarks motivates the question of whether could also be productive to treat more widely-used acoustic features (e.g. MFCCs) as 'events'. In paralinguistic speech processing, partitions of the acoustic feature space have been used previously, exemplified by the increasing popularity of Bag-of-Words (BoW) for continuous emotion prediction [14], [15] and depression classification [16], where the acoustic space of spectral features is clustered into different broad partitions of the acoustic space. Another (more intuitive) category of acoustic 'events', relating to linguistic content, is phonemes. However, interestingly, it is shown in [14], [17] that phonetic features, which also effectively partition the acoustic space, yielded encouraging performance for emotion prediction due more to the partitioning information than to the phonetic content. Despite the effectiveness of tokenizing acoustic features to produce such acoustic 'events', surprisingly few studies have gone beyond the use of Bag-of-Words (BoW) in terms of exploring system designs with sequences of event-based (symbolic) features and the natural language processing tools that can then be applied to them. This study aims to address this by 1) examining sequential patterns of acoustic 'events'; and 2) exploring effective techniques originating from natural language processing, applied to acoustic events.

In particular, we evaluate acoustic events (herein termed

Zhaocheng Huang, Julien Epps and Vidhyasaharan Sethu are with School of Electrical Engineering and Telecommunications, the University of New South Wales (UNSW Sydney), Australia (e-mail: {zhaocheng.huang, j.epps, v.sethu}@unsw.edu.au).

Dale Joachim is with Sonde Health Inc., Boston, MA, United States (e-mail: djoachim@sondehealth.com).

acoustic words) derived from spectral features extracted at a fixed frame rate (also known in the literature as bag of audio words [15]), and landmarks (herein termed landmark words). We note that conceptually such acoustic events carry more information related to speech '*content*' (i.e. it is common to find a sequence comprising entirely the same acoustic event type), whereas by contrast speech landmarks signal abrupt, significant transitions in speech articulation, and therefore contain important *boundary-related* information (analogous to edges in image processing). Acoustic words and landmark words are significantly different in another respect: the former emanates from spectral features extracted with a *fixed frame rate*, whereas the latter are *irregularly spaced in time* and only occur in response to changes in speech articulation.

Such distinct characteristics conveyed by acoustic words and landmark words are expected to be complementary. Thus, this paper also addresses the follow-up research question of whether acoustic events and speech landmark events can be effectively combined or fused to yield further improvements.

II. Related Work

A. Acoustic Events in Depression

The most widely used features to date for speech analysis, including depression detection, are arguably acoustic features extracted within fixed-length speech frames (e.g. 20 milliseconds) at regular intervals [1]. Such frame-wise features contain detailed content-based information from speech, for instance, mel frequency cepstral coefficients (MFCCs) which describe spectral properties of a speech segment. However, such information contains some redundancy as consecutive frames (e.g., monophthong) relay the same information, and may also be too detailed, since MFCCs carry significant phonetic and speaker-specific information, which can contribute unwanted variability [18], [19].

Several studies have investigated the variability of the acoustic feature space for depressed speakers [16], [18], [20]. For instance, acoustic feature spaces from either vowel-specific formants [18] or spectral features [20] have been found to exhibit less variability for depressed speakers. A more recent technique is Bag-of-Words (BoW), which clusters and divides the acoustic feature space into multiple partitions, then computes their occurrence frequencies as features. BoW features have showed improved performances over conventional MFCCs for classifying three bipolar disorders (i.e., remission, hypo-mania, and mania) in the Audio-Visual Emotion Challenge (AVEC) 2018 bipolar sub-challenge baseline [16]. These studies confirm the usefulness of the partitioned acoustic space information for depression.

In order to better understand the superior performance for acoustic BoW, the partitions of the acoustic space need special scrutiny. This paper refers to each partition of the acoustic space as an acoustic event or 'word'. In this context, we note that previous studies have not investigated n-grams for $n > 2$, i.e. two or more consecutive acoustic words, as well as deriving more meaningful and useful representations from the acoustic words by going beyond BoW [15].

B. Articulatory Events in Depression

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 [21], phonation and articulation errors, articulatory incoordination [22], disturbances in muscle tension, psychomotor retardation, phoneme rates [23], and altered speech quality and prosody [24]. Thus, articulator movements, which shape speech production, are expected to be informative for assessing changes incurred by depression.

Recently, another category of speech feature based on speech landmarks has been successfully applied to depression detection in a naturalistic environment with promising results [10], [12]. By contrast with acoustic events, speech landmark events occur irregularly, only when certain articulatory changes such as burst onset/offset occur with measurable abrupt energy level changes. It was reported in [10] that landmark bigram counts yielded promising results for classifying depressed speakers and can be adapted to different elicitation tasks by tailoring landmark choices. Moreover, performances were further boosted by applying Latent Dirichlet Allocation (LDA) to exploit latent articulatory events, however a much wider array of embedding methods could also be considered. The landmark events herein are referred to as landmark words, because they are intrinsically symbolic. The more general possibilities for exploiting landmark words, which contain both sequence-related and timing information, for health-related applications are still yet to be fully realized.

C. NLP Event Processing Techniques for Depression

Natural Language Processing (NLP) methods have evolved for mining useful information from text data and (to a lesser extent) conversational transcripts over the last few decades [25], [26]. However, they have been rarely applied to the analysis of acoustic speech, due in large part to the distinct natures of the speech signal and the text modality: 1) the former contains rapidly-varying continuous information that is usually characterized numerically, whereas the latter is usually characterized symbolically using discrete vocabularies (or words) in documents. 2) the former tends to explore altered characteristics in response to changes in either linguistic content (e.g. phonemes, words, etc.) or paralinguistic content (e.g. depression, emotion, identify, etc.), whereas the latter tends to explore the meaning of a cohort of words or individual words by finding relevant document-wise topics [27], [28], [29] or word similarities [30], [31].

Learning an effective vector representation (i.e. embedding) for words has been an active research topic in NLP. The 'embedding' refers to mapping a symbol to a numeric vector in high dimensional space. The conventional default choice is one-hot encoding, which has 1 for the position of the present word and 0 for the remaining words in a fixed dictionary. However, this representation is high dimensional, and assumes no interaction between different words. Word2Vec and GloVe, which explicitly learns semantic meaning and word similarity via supervised learning, has proved to be more effective, and is prevalent in NLP tasks.

As the importance of the text modality (conversation transcripts) has been recently discovered for detecting depression, there has been increasing interest in using text-based analysis approaches for useful text-based features, e.g.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing

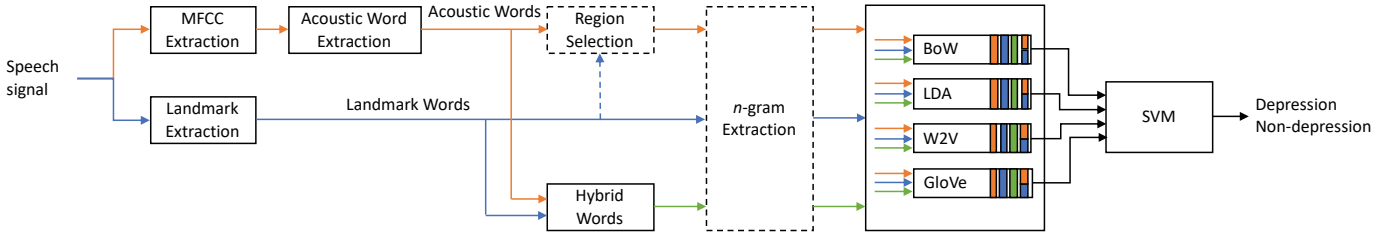


Fig. 1: General overview of proposed acoustic and landmark word system design paradigm. 'Hybrid words' can be generated from the combination of these heterogeneous words. Either acoustic events or landmark events can optionally define regions of interest from which *n*-grams are extracted. Various embedding representations can then be applied either individually to acoustic words or in combination to exploit event-based features for depression detection. 'W2V' means Word2Vec embeddings.

[32], [33] as part of conventional multimodal system designs [32], [34]. For example, Global Vector (GloVe) embedding was used very effectively to exploit the questions and answers in spoken transcripts for classifying depression on the DAIC-WOZ dataset using question-based GloVe with the ZCA whitening transformation [32]. The use of GloVe was further extended to automatic ASR transcriptions to enhance privacy-preservation in [33]. [For AVEC 2018 emotion prediction, the word2vec, bow and glove were used on text modality [35]. For AVEC 2017, depression prediction, also on text modality, an extension of the word2vec embedding the so-called paragraph vector was used [36].] However, it is important to note that to date, the use of such text-based analyses, apart from BoW, has only been limited to the text modality, and not (to our knowledge) the acoustic speech signal.

This gap is filled by the present study, wherein a number of NLP methods are used to not only exploit acoustic words and landmark words, but also fuse different types of words at different levels.

III. Methods – Acoustic and Landmark Words

A. Overview of Proposed System Design Paradigm

The proposed paradigm explored in this paper is shown in Fig. 1. Remarkably, apart from MFCC extraction and the classifier, all system components operate on symbolic rather than numerical inputs. This section defines acoustic words (AW) and landmark words (LW), as well as their sequential patterns, i.e. *n*-grams of AW or LW. The paradigm is novel in applying text-based methodologies to acoustic and landmark words. It also allows unique possibilities for combining heterogeneous event types. Herein we investigate three frameworks to fuse/combine the AW and LW: word-level fusion (i.e. creating hybrid words), embedding-level fusion, and landmark-dependent AW.

B. Acoustic Words (AW) and Landmark Words (LW)

1) Acoustic Words via Clustering (unigram)

The formulation of acoustic words starts with applying *k-means* clustering algorithm to divide the acoustic feature space (akin to vector representation, which then serve as tokens), building on top of low-level descriptors such as spectral features [15]. This leads to K distinct acoustic words or tokens, where K is the number of words/tokens manually specified.

$$S_{AW} = \{a_1, a_2, \dots a_k, \dots, a_K\} \quad (1)$$

Each symbol a_k has a d-dimensional numeric vector $\boldsymbol{x}_{a_k}$, with $k \in [1, K]$, representing the corresponding centroid within the feature space.

Given that each speech file has a set of frame-level acoustic features $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_{M_a}\}$, where M_a is the total number of frames and varies for different files, i is typically a uniform time index, frame $\boldsymbol{x}_i$ is mapped to a symbolic acoustic word a_k which has the minimum Euclidean distance between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{a_k}$ [15].

$$\boldsymbol{x}_i \rightarrow a_k = \arg\min_{a_k \forall k} \left\| \boldsymbol{x}_i - \boldsymbol{x}_{a_k} \right\| \quad (2)$$

By this, each numeric vector $\boldsymbol{x}_i$ is converted into an acoustic word $w_i^a = a_k$, leading to a sequence of acoustic words W^a, associated with each speech file.

$$W^a = \{w_1^a, \dots, w_i^a, \dots, w_{M_a}^a\}, \text{s.t.} \, w_i^a \in S_{AW} \quad (3)$$

In this way, frame-level features X are converted to frame-level acoustic words W^a. Despite previous studies on clustering the acoustic space for emotion prediction [15], [14], sentiment analysis [37], abnormal heart sound detection [38], and bipolar disorder classification [16], representing the acoustic space as a set of acoustic words allows a number of techniques from natural language processing to be applied. For example, learning sequential patterns (e.g. using *n*-grams as per in Section III-B-3), extracting latent topics (e.g. using LDA as per in Section III-C-2), and deriving meaningful word embeddings that explore the dependency (e.g. in text, apple is related to juice) or similarity (e.g. in text, apple and orange are in the category of fruits) using context information (e.g. using Word2Vec and GloVe as per in Section III-C-3 and III-C-4). More specifically, for words that occur in proximity (mostly within a certain window) in text processing, any word can be learnt from the other nearby words, which provide useful information as context. Similarly, acoustic words that occur in proximity carry knowledge about their nearby acoustic words, which can be explored by considering unique sequences of words (III-B-3) and modeling this relationship using techniques from natural language processing (III-C).

2) Landmark Words (unigram)

This study employs six landmarks: '**g**(lottis)', '**p**(eriodicity)', '**s**(onorant)', '**f**(ricative)', '**v**(oiced fricative)', and '**b**(ursts)', each of which specifies points in time at which different abrupt acoustic events occur. Each of the six landmarks is represented by onset and offset times. We therefore define a set of $L = 6$

landmarks, each with onset (+) and offset (-) states, i.e. 2*L* states in total:

$$S_{LW} = \{g_{\pm}, p_{\pm}, s_{\pm}, f_{\pm}, v_{\pm}, b_{\pm}\} \quad (4)$$

The 12 unique landmark states are referred to as landmark words (LW) in this study. Similarly to W^a, each speech file has a sequence of associated landmark words W^l:

$$W^l = \{w_1^l, \dots, w_i^l, \dots, w_{M_l}^l\}, \text{s.t.} \, w_i^l \in S_{LW} \quad (5)$$

where w_i^l is the i^{th} landmark word (i is a non-uniform time index) and M_l is the landmark count for a speech file.

This paper explores the subset of published consonantal landmarks described in Table 1, which can be extracted using the SpeechMark software [39]. Note that the non-standard p landmark provided by SpeechMark, although not part of the original landmark set and seldom described in the literature, is included as it has been shown to be a valuable differentiator of depressed/non-depressed speech [10]. The other landmarks include voiced and unvoiced fricatives (v, f) as well as bursts and sonorants (b, s).

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

It is worth noting that landmark words occur at irregular intervals in time, whereas acoustic words occur at a fixed rate basis (i.e. 10ms), as seen in Fig. 3. Conceptually, acoustic words can be considered to capture mainly '*content*'-like information (most frames will not be abrupt changes), whereas landmark words capture '*boundary*'-like information that by definition reflects changes in speech articulation.

3) AW and LW n-grams

The acoustic words and landmark words formulated from 1) and 2) above can be regarded as 'unigrams', as in speech recognition and natural language processing. A natural extension of unigrams, i.e. grouping of multiple words per time, results in *n*-grams such as 2-gram (bigram) and 3-gram (trigram), as proposed in this subsection. In order to generalize the notation of acoustic and landmark words, w_i is used to represent either w_i^a or w_i^l, M represents either M_a or M_l, and S represents either S_{AW} or S_{LW}. Accordingly, *n*-grams $w_{i\to j}$ are sequences of *n* consecutive unigrams:

$$W_n = \{w_{i\to j}\}_{i=1}^{j=M}, \text{s.t.} \, j = i + n - 1, \quad (6)$$

$$w_{i\to j} = \{w_i, w_{i+1}, \dots, w_j\}, \text{s.t.} \, i \geq 1, j \leq M, \quad (7)$$

where $w_{i\to j}$ is a sequence starting from w_i and ending with w_j, and i, 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 W^a, W^l in (3) and (5).

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

$$w^{m,l} = \{w_{i\to j} | w_{i\to j-1} = m, w_j = l\}, \text{s.t.} \, m, l \in S \quad (8)$$

where m and l denote a certain word or word sequence, in which each word belongs to the vocabulary set S $w^{m,l}$ specifies a certain *n*-gram. For instance, if $m = \{g_+, p_+, p_-\}$ and $l = \{s_+\}$, then $w^{m,l} = \{g_+, p_+, p_-, s_+\}$ in the case of landmark words. For acoustic words, if $m = \{a_1, a_1, a_2\}$ and $l = \{a_4\}$, then $w^{m,l} = \{a_1, a_1, a_2, a_4\}$.

Such sequential patterns allow additional information to be captured, such as transition information and more uniquely, specific word sequences. The use of *n*-gram has commonly yielded improved performances in a range of speech-related tasks such as speech recognition (triphones) [40] or machine translation (*n*-grams). Related work applied to sequences of acoustic words is the "bags-in-bag" method [41], proposed to incorporate context awareness for emotion prediction; however this approach does not explicitly capture transitions between different words. It is worth noting that the sequential or temporal information regarding the words can also be implicitly modelled by recurrent neural nets [42]. However, the investigation conducted here aims to explicitly examine the importance of considering sequence information.

C. Framework for Extraction and Aggregation of Token-based Features

1) BoAW and BoLW

The BoW features for acoustic words and landmark words are referred to as bag-of-acoustic-words (BoAW) and bag-of-landmark-words (BoLW). Given a sequence of *n*-grams (including unigram) for a speech file, the counts for all unique *n*-gram sequences are

$$\boldsymbol{x}_n^{\text{BoW}} = \frac{\#(W_n)}{M} = \{\frac{\#(w^{m,l})}{M}\}_{\forall m \in S, l \in S} \quad (9)$$

where $\#(\cdot)$ is the counting operation applied to the whole speech recording. That is, all the possible unique *n*-grams $w^{m,l}$ (including unigram) are counted, and then divided by the total number of words M to yield the BoW features $\boldsymbol{x}_n^{\text{BoW}}$. The BoAW $\boldsymbol{x}_n^{\text{BoAW}}$ and BoLW $\boldsymbol{x}_n^{\text{BoLW}}$ are extracted using (9) using W_n, M, S, m, l specific to acoustic words and landmark words respectively. The feature dimension depends on the number of unique words occurring within the whole training data.

2) Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) has been widely used for topic modelling in NLP since its introduction in 2003 [27], [43], [29]. In topic modelling, LDA generates a representation of latent topics (e.g. sports, travel, etc.) given text documents consisting of words (vocabularies). This idea motivated our previous work [10], in which landmark bigrams (i.e. of two consecutive landmarks) were treated as 'words' to effectively exploit meaningful latent articulatory events using LDA for improved performance in depression classification in real-world environments.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing

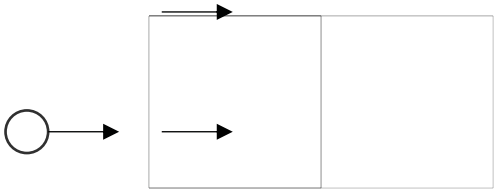


Fig. 2: The graphical model for applying LDA to depressed speech. There are *D* speech files ('documents'), *M* acoustic/landmark words ('words'), and K_T latent articulatory events ('topics'). $\boldsymbol{w}_{d,i}$ is the *i*th bigram in the *d*th speech file. The latent variables $\boldsymbol{z}_{d,i}$, $\boldsymbol{\beta}_k$, $\boldsymbol{\theta}_d$ are estimated from training, controlled by hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$.

In Fig. 2, parameters β_k and θ_d follow a Dirichlet distribution, describing respectively probability distributions over *M* words for the k^{th} event and over K_T events for the d^{th} speech file. Sampling from θ_d gives $z_{d,i} = k$, and subsequently $\beta_{z_{d,i}=k}$, which specifies the probability of generating $w_{d,i}$. Both $z_{d,i}$ and $w_{d,i}$ follow a multinomial distribution. Overall, $z_{d,i}$, β_k and θ_d describe relationship for *word-events-speech*, which is similar to *word-topic-document* in topic modelling.

The training of LDA was done via variational Bayesian inference to approximate the 'true' posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \eta)$. After the LDA model is trained, given a new audio file d^*, which has words, the resultant features are θ_{d^*}, that give probabilities of K_T latent events. Detailed derivation of LDA for depression detection is described in [10]. Since θ_{d^*} follows a Dirichlet distribution, this gives probabilities for all K_T latent events:

3) Word2Vec

Unlike BoW, Word2Vec can explore semantic meaning and word similarity. Word2Vec leverages the relationship between 'target' and 'context': given a set of *T*+1 words $\{w_{i-T/2} \dots w_{i-1}, w_i, w_{i+1} \dots w_{i+T/2}\}$, the middle word w_i becomes the 'target', whereas the remaining words become the 'context' words $\boldsymbol{w}_{-i} = \{w_{i-T/2} \dots w_{i-1}, w_{i+1} \dots w_{i+T/2}\}$. The task of Word2Vec training is to train a feed-forward neural network that either 1) takes the target words as input to predict the context words or 2) takes the context words to predict the target words. This study adopted the latter approach.

Let *V*-dimensional $\tilde{x}_w$ represent a one-hot vector for word w, and *V* is the total number of words. Using a feed-forward neural network with one hidden layer that has *N* linear neurons, the whole network can be characterized by a $V \times N$ matrix $\boldsymbol{W}$ and a $N \times V$ matrix $\boldsymbol{W}'$. $\boldsymbol{W}$ maps the input to hidden layer, and $\boldsymbol{W}'$ maps the hidden layer to the output. Accordingly, every single word w has an input vector v_w and output vector v'_w.

The posterior probability for the target word w_i given context words $\boldsymbol{w}_{-i}$ is [44]:

$$p(w_i|\boldsymbol{w}_{-i}) = \frac{\exp({v'_{\boldsymbol{w}_i}}^T v_{\boldsymbol{w}_{-i}})}{\sum_{j=1}^{V} \exp({v'_{\boldsymbol{w}_j}}^T v_{\boldsymbol{w}_{-i}})} \tag{12}$$

where the input vector $v'_{\boldsymbol{w}_i}$ for multiple context words $\boldsymbol{w}_{-i}$ is an average of input vectors of all individual context words. The Word2Vec training updates weights in $\boldsymbol{W}$ and $\boldsymbol{W}'$ to maximize the probabilities on the training set using backpropagation and gradient ascent (for maximization). After training, an embedding e_i for each word w_i can be obtained from $\boldsymbol{W}$, which the embedding matrix. The file-level embedding is then an average of embeddings of individual words.

—

It is our hypothesis that there exist more meaningful sets of embeddings that Word2Vec can learn from the proposed acoustic words and landmark words. Moreover, according to (11), if the context input vector $v_{\boldsymbol{w}_{-i}}$ is similar to the target output vector $v'_{\boldsymbol{w}_i}$, their inner product will be large and in turn the posterior probability becomes large. This is how the Word2Vec model enforces word similarity, which can be beneficial for embedding learning for depressed speech.

4) GloVe

Like Word2Vec, GloVe also learns embeddings per word. One significant difference between Word2Vec and GloVe is that GloVe looks at global word-word co-occurrence counts across the training set, whereas Word2Vec tries to capture co-occurrence within one window per time. The original intuition was that the words that frequently appear together should have similar vector representations, allowing fast training and scalability to large corpora.

Let *P* be the co-occurrence matrix where P_{ij} is how often (probability) word w_i appears in the context of word w_j. Identically, we have input vector v_w and output vector v'_w for the word w as in (11). The loss function that guides the training process is as follows:

$$\theta^* = \arg\min_{\theta^*}\left(\frac{1}{2}\sum_{v_{w_i}, v'_{w_j} \in \mathcal{V}} f(P_{ij})({v_{w_i}}^T v'_{w_j} - \log i\right. \tag{14}$$

where $f(P_{ij})$ is a weighting function to enforce a few conditions such as $f(0) = 0$, i.e. not overly weighting rarely and frequently occurring words [31].

After training, a GloVe embedding e_i for each word w_i can be obtained, as for Word2Vec, and the file-level GloVe features are an averaged of all the embeddings across the file.

—

Note that in the above methods, BoW and LDA provides one feature vector per file, whereas Word2Vec and GloVe provide one feature vector per word, which was then averaged per file.

IV. Methods: Proposed Hybrid Systems

The tokenization of the acoustic space offers unique opportunities for combining with event features like landmarks. Here, we propose three approaches for hybrid systems that contain information from both AW and LW: hybrid words (HW), hybrid embeddings (HE), and landmark-dependent acoustic words (LD-AW).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing

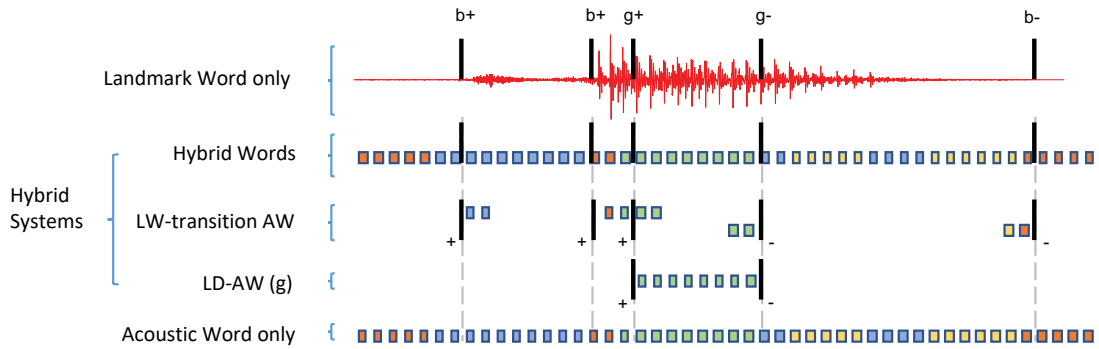


Fig. 3: Illustration of proposed framework for combining acoustic and landmark words. A sample waveform of about 0.5 seconds is shown (top row), from which acoustic and landmark words are extracted and combined to create hybrid words, LW-transition AW and LD-AW specific to landmark g. Note that the colours in LW-transition AW and LD-AW should be different for acoustic words only, since the acoustic space of the selected spectral features will be re-tokenized per landmark. LW-transtition AW, which is not investigated in this study, presents an additional possibility to combine acoustic and landmark words motivated by the findings in the literation that abrupt acoustic speech changes are particularly informative. In LW-transition AW, *N* frames of acoustic features both after onsets and before offsets are selected for clustering the acoustic space for acoustic words.

A. Proposed Hybrid Words (HW)

This approach considers potentially heterogeneous combinations of acoustic and landmark words, i.e. K+12 possible words in total.

$$S_{HW} = \{a_1, a_2, \dots, a_K, g_{\pm}, p_{\pm}, s_{\pm}, f_{\pm}, v_{\pm}, b_{\pm}\} \quad (16)$$

Similarly to (3) and (5), a sequence of hybrid words for each speech file is

$$W^h = \{w_1, \dots, w_i, \dots, w_{M_a+M_l}\}, \text{s.t.} \, w_i \in S_{HW} \quad (17)$$

where acoustic and landmark words are aligned in time (as shown in Fig. 3). M_a, M_l are the total number of acoustic and landmark words per file. The hybrid word sequences W^h from the training data were used to train LDA, Word2Vec and Glove. Based on the trained model, $\boldsymbol{x}_{HW}^{\text{LDA}}$, $\boldsymbol{x}_{HW}^{\text{W2V}}$, $\boldsymbol{x}_{HW}^{\text{GloVe}}$ can be calculated using (10), (15), (17) respectively, alongside $\boldsymbol{x}_{HW}^{\text{BoW}}$ using (9) which requires no training.

It is expected that acoustic words (which occur at fixed intervals and capture content-like information) and landmark words (which occur on an irregular time basis and capture boundary-like information) may be complementary at the word level. Moreover, hybrid words allow interactions between acoustic and landmark words to be learnt, for instance, for Word2Vec, a model is trained to predict a target landmark word given context acoustic words (or mixed words).

It is worth noting that the choice of K in acoustic words will have an impact on the relative occurrence percentages for AW and LW given that the total number of acoustic frames is fixed: a large K will produce a relatively smaller number of occurrences for each word, whereas smaller K will lead to each word occurring more frequently. In turn, the former case leads to landmark words outnumbering acoustic words, whereas the latter causes landmark words to be outnumbered by acoustic words.

B. Proposed Hybrid Embeddings (HE)

Another type of hybrid system fuses embedding representations of individual words (i.e. AW and LW), which is referred to as hybrid embedding (HE) or embedding-level fusion, as shown in Fig. 1. To be more specific, for example, BoAW features $\boldsymbol{x}_{AW}^{\text{BoW}}$ and BoLW features $\boldsymbol{x}_{LW}^{\text{BoW}}$ are concatenated and then input to Linear SVM for training and testing. Such concatenation can be similarly applied to other embedding approaches, that is LDA, Word2Vec, and GloVe.

$$\begin{aligned} \boldsymbol{x}_{HE}^{\text{BoW}} &= \begin{bmatrix} \boldsymbol{x}_{AW}^{\text{BoW}} \\ \boldsymbol{x}_{LW}^{\text{BoW}} \end{bmatrix}, \boldsymbol{x}_{HE}^{\text{LDA}} = \begin{bmatrix} \boldsymbol{x}_{AW}^{\text{LDA}} \\ \boldsymbol{x}_{LW}^{\text{LDA}} \end{bmatrix}, \\ \boldsymbol{x}_{HE}^{\text{W2V}} &= \begin{bmatrix} \boldsymbol{x}_{AW}^{\text{W2V}} \\ \boldsymbol{x}_{LW}^{\text{W2V}} \end{bmatrix}, \boldsymbol{x}_{HE}^{\text{GloVe}} = \begin{bmatrix} \boldsymbol{x}_{AW}^{\text{GloVe}} \\ \boldsymbol{x}_{LW}^{\text{GloVe}} \end{bmatrix} \end{aligned} \quad (18)$$

Hybrid embedding offers an advantage over embedding of individual words (i.e. AW or LW) or hybrid words by combining unique information encoded in the AW and LW using (19). As such, it can harness the complementary properties between acoustic and landmark words at the embedding level. It is not uncommon in the literature that such intermediate-level fusion (i.e. after word-level and before model-level) is a feasible way to combine different modalities with improved robustness [3], [45], [46]. For instance, in [45], BoW features were extracted separately from audio and video modalities, and then concatenated before Principle Component Analysis (PCA) for dimensionality reduction. This yielded better results than fusing SVM scores or binary outputs for binary classification of depression. A similar approach can be seen in [46], which however concatenated video BoW features with KL-mean from audio (which resembles BoW features) for depression prediction.

Despite the fact that concatenation of BoW features from different modalities has been studied in the aforementioned literature, this fusion approach is relatively naïve. Other embedding techniques, i.e. LDA, Word2Vec, and GloVe, which have found to produce consistently more effective embedding representations in NLP, might capture more unique information associated with different modalities (i.e. acoustic words and landmark words in this study) than BoW.

C. Proposed Landmark-Dependent Acoustic Words (LD-AW)

The third type of hybrid system applies region selection of acoustic words based on landmarks, i.e. acoustic words are only considered within the onset and offset states of a landmark, on a per-landmark basis. This approach is referred to as Landmark-Dependent Acoustic Words (LD-AW). For instance, as shown in Fig. 3, for LD-AW (g), spectral (MFCC) features were selected between g+ (i.e. vocal fold vibration starts) and g- (vocal fold vibration ends), and the selected spectral features were tokenized into acoustic words via clustering. By this, the selected features are related to activities of the vocal fold, and therefore deriving acoustic words from the g-specific spectral features enables each word to contain information about acoustic space specific to speech articulation, in the above case, vocal fold vibration.

Given that each speech file has a set of frame-level acoustic features $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_{M_a}\}$, we define $t(w_i^{l+})$ and $t(w_i^{l-})$ as the time indexes of an onset landmark w_i^{l+} and its corresponding offset landmark w_i^{l-}, where $l \in \{g, p, s, f, v, b\}$. Similarly, we define $t(\boldsymbol{x}_i)$ as the time index for the i^{th} frame of acoustic features. Hence, $\{t(w_i^{l+}), t(w_i^{l-})\}$ specifies a l-dependent region. Each speech file contains M_j (which varies for each file) landmark-dependent regions, and accordingly we define the j^{th} region as $\{t_j(w_i^{l+}),\ t_j(w_i^{l-})\}$. LD-AW operates by selecting spectral features within the l-specific regions $\{t_j(w_i^{l+}),\ t_j(w_i^{l-})\}$ as follows:

$$X^l = \left\{\boldsymbol{x}_{i,j}^l\right\}_{j=1}^{M_j}, \text{s.t.}\, t_j(w_i^{l+}) \le\ t(\boldsymbol{x}_{i,j}) \le\ t_j\left(w_i^{l-}\right) \quad (19)$$

Concatenating the M_j regions in time gives:

$$X^l = \left\{\boldsymbol{x}_1^{a,l}, \dots, \boldsymbol{x}_i^{a,l}, \dots, \boldsymbol{x}_{M_{LD}}^{a,l}\right\} \quad (20)$$

where M_{LD} is the total number of selected frames specific to landmark $l \in \{g, p, s, f, v, b\}$.

Following the same scheme as in Section III-B-1, the landmark-dependent acoustic words can be obtained by tokenizing the acoustic space via clustering, leading to

$$\boldsymbol{x}_i^{a,l} \rightarrow a_k^l = \arg\min_{a_k^l\ \forall k} \left\| \boldsymbol{x}_i^{a,l} - \boldsymbol{x}_{a_k^l} \right\| \quad (21)$$

where $\boldsymbol{x}_{a_k^l}$ represents a numeric vector for the k-th l-dependent words, and l-dependent spectral features $\boldsymbol{x}_i^{a,l}$ are converted into an l-dependent acoustic word $w_i^{a,l} = a_k^l$, leading to a sequence of acoustic words W^a, associated with each speech file.

$$W^{a,l} = \left\{w_1^{a,l}, \dots, w_i^{a,l}, \dots, w_{M_{LD}}^{a,l}\right\}, \text{s.t.}\, w_i^{a,l} \in S_{AW} \quad (22)$$

Further, embeddings for the landmark-dependent acoustic words, i.e., $\boldsymbol{x}_{LD-AW(l)}^{\text{LDA}}, \boldsymbol{x}_{LD-AW(l)}^{\text{W2V}}, \boldsymbol{x}_{LD-AW(l)}^{\text{GloVe}}$ can be calculated using (10), (15), (17) respectively, alongside $\boldsymbol{x}_{LD-AW\,(l)}^{\text{BoW}}$ using (9) which requires no training.

One merit of LD-AW is that speech articulation information is integrated into embedding representations in a manner that isolates some of the unwanted phonetic variability, which might aid system performances for depression detection.

V. Experimental Settings

A. Databases

This study employed two corpora, selected for comparison with literature and for evaluation across multiple smartphone types in naturalistic environments respectively: The Distress Analysis Interview Corpus (DAIC-WOZ) corpus [6] and the SH2 corpus [9]. DAIC-WOZ is a laboratory-based dataset collected in scenarios where participants were interviewed by a virtual human agent who asks all participants the same set of questions. All 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 per participant, and an accompanying binary label indicating whether the participant was depressed or healthy. The database has 189 speakers in total, 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 [47]. 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 recordings (sampled at 44.1kHz) collected in unsupervised naturalistic environments on participants' own smartphones, as well as PHQ-9 scores from self-administered questionnaires, which measure depression severity. Participants completed several voice tasks including timed free speech, read speech (Rainbow passage and Harvard sentences), and elicited tasks: sustained vowel "ahh" and diadochokinetic repetition. Details of the SH2 corpus were provided in [9]. This study adopted the free speech voice partition, referred to as SH2-FS, due to its similarity to the voice samples from DAIC-WOZ. SH2-FS has the same training and testing partition as [9]: 440 files (436 speakers) for training and 130 files (128 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 [48]), 74 depressed and 23 depressed speakers were respectively found in the training and test data partitions.

The DAIC-WOZ corpus is among the largest publicly available depression corpora to date and has been widely used, whereas SH2-FS contains a larger number of speakers compared to DAIC-WOZ and other existing depression datasets within the research community. Compared with DAIC-WOZ, which has clean, long-duration recordings from a single set of recording hardware, SH2-FS has a larger number of speakers, shorter durations, different smartphone recording hardware characteristics, and noisy naturalistic environments. The average speech utterance durations are 446.9 ± 227.0s for DAIC-WOZ and 20.5 ± 10.2s for SH2-FS. The fact that the two datasets are very different not only contributes insights to the research community into depression detection in a naturalistic environment, but also validates the feasibility of the proposed methods in both scenarios.

In this study, the aforementioned training-testing partitions are the same as per [45], [52], [53] for DAIC-WOZ and as per [9], [10] for SH2-FS for comparison purposes. It is worth noting that existing statistical test methods are not feasible in the case

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing

of fixed training and testing set, since algorithm was executed once in each system. Although, according to [49], "*For algorithms that can be executed only once, McNemar's test is the only test with acceptable Type I error*.", McNemar's test was implemented in this study but observed to be inappropriate and impractical, since neither can it evaluate whether one algorithm is more or less accurate than the other, nor it does not account for the class imbalance in both datasets.

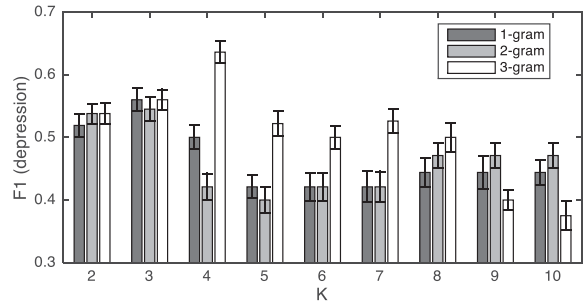
B. Settings

All experiments in this study adopted a linear Support Vector Machine (SVM) classifier [50], whose complexity coefficient C was tuned from 10^{-5} to 10 in log space in a 3-fold cross validation scheme within the training data. 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], [10]. 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 evaluated on the test partition. The F1 score for the depressed class, which combines recall and precision, was used as the main metric. Additionally, F1 (healthy), classification accuracy and confusion matrices were calculated for final results.

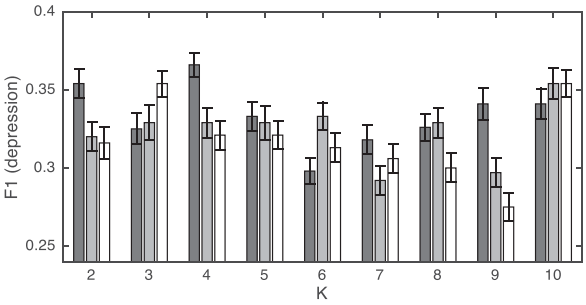
For the landmark words and hybrid words, gender normalization [9], [10], which applies z-normalization to dataset subsets specific to gender, was used to reduce the gender dependency of the landmark occurrence as reported in [51]. That is, the training data and testing data specific to gender were normalized by subtracting the mean and divided by the standard deviation, based on normalization coefficients learnt from the training data. Dimensionality for LDA, Word2Vec and GloVe was optimized in terms of F1 scores from among {4, 8, 12, 16, 20, 24, 28, 32, 36, 40}, unless stated. The implementation of LDA used scikit-learn [52], and the implementation of Word2Vec and GloVe used gensim [53]. For short-term acoustic features, we used 16 MFCCs and 16 ΔMFCCs. The *k*-means clustering algorithm was used for acoustic words [15], [14]. For Word2Vec and GloVe, the window size in which the context words are considered was set to 10 words. We used the CBOW algorithm in Word2Vec without negative sampling, since the vocabulary size is small. For GloVe, we used 30 epochs and a learning rate of 0.05 throughout all the experiments. The word-level embeddings from Word2Vec and GloVe were averaged across each speech utterance as shown in (15) and (17). Weighting using the term frequency – inverse document frequency (TF-IDF) was also trialed for Word2Vec and GloVe, but improvements were not significant.

VI. Results: Acoustic Words only and Landmark Words only

This section investigates the performance of acoustic words (AW) and landmark words (LW) separately, seeking answers to the following questions: 1) what are appropriate characterizations of the acoustic space and acoustic word sequences, determined by *K(-words)* and *n(-gram)* respectively?; 2) how well do text-based approaches perform in exploiting the tokenized acoustic space?; and 3) how well do word sequences, determined by *n(-gram)* and text-based methods perform on the *landmark* words?



(a) DAIC-WOZ



(b) SH2-FS

Fig. 4: F1 (depression) scores and their 95% confidence intervals for the DAIC-WOZ and SH2-FS datasets under different *K* values for 1-gram, 2-gram and 3-gram using BoW features. It was found that in general *K*=4 provided better results, suggesting that a general partitioning of the acoustic space is helpful. No embedding techniques were used here.

A. Acoustic Words – Characterization and Embedding

In tokenizing the acoustic space into acoustic words, two parameters are important, i.e. the number of clusters/centroids K and the number of consecutive tokens considered n. K refers to the number of clusters used to partition the feature acoustic space, whereas n refers to the number of sequential clusters used. For instance, larger K means more clusters, which provide detailed partitions of the acoustic space, and larger n means more sequential information. K and n together characterize the acoustic space in the form of clusters, as well as their transitions, each of which could be affected by depression. F1 (depression) scores and their 95% confidence intervals are shown in Fig. 4 as a function of various choices of K and n for BoW (i.e. with no embeddings). In order to generate the 95% confidence intervals on a fixed test set, we randomly sampled around 2/3 of the test predictions (that is 24 out of 35 speakers for DAIC-WOZ and 86 out of 128 speakers for SH2-FS) and evaluated their performances, which was repeated 100 times. The results suggest that K words of 3 or 4 are effective for both datasets and most choices of n. This small value of K means that the AWs are likely to characterize very broad categories such as voiced, unvoiced and silent speech.

This experiment was then repeated using LDA, Word2Vec or GloVe, to understand the effect of embeddings, with results shown in Fig. 5. From these it can be noted that embeddings provided a significant and consistent improvement for both datasets. Overall, Fig. 4 and 5 convey the insights that smaller K and $n \geq 2$ (acoustic word sequences) are beneficial.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing

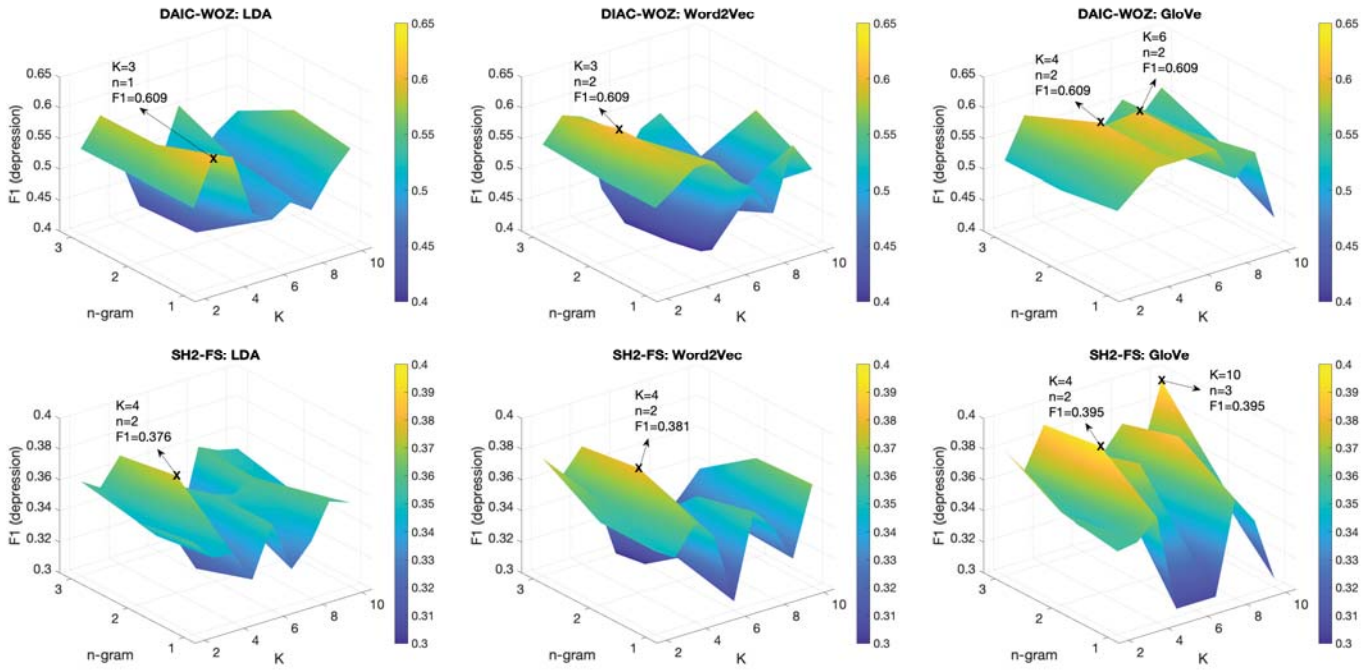


Fig. 5: Comparison of the optimal *K* and *n* with different embedding methods (i.e. LDA, Word2Vec, and GloVe) across DAIC-WOZ and SH2-FS. For SH2-FS, the combination of *K*=4 and *n*=2 consistently produced optimal performances, which were better than BoW features, suggesting better exploitation of embedding methods than simple count-based features.

TABLE 2
COMPARISON OF SELF-TRANSITIONS AND NON-SELF-TRANSITIONS F1 SCORES FOR AW N-GRAM, USING DIFFERENT EMBEDDING METHODS.

	DAIC-WOZ			SH2-FS		
	All	Self	Non-self	All	Self	Non-self
BoW	0.421	0.571	0.400	0.329	0.329	0.264
LDA	0.500	0.600	0.526	0.376	0.333	0.306
Word2Vec	0.583	0.609	0.519	0.381	0.381	0.288
GloVe	0.560	0.609	0.522	0.395	0.360	0.381

Accordingly, for all experiments on AWs in later sections except where otherwise indicated, based on Fig. 4 and Fig. 5, *K* and *n* was set to 4 and 2 respectively, for both datasets.

Sequences of AWs (i.e. an *n*-gram for $n \geq 2$) consist of self-transitions (i.e. repeated AWs) and non-self-transitions (i.e. differing AWs). Self-transitions occur more frequently than non-self-transitions (88% vs. 12% on DAIC-WOZ and 87% vs. 13% on SH2-FS when *K*=4). This raises questions about the relative contribution of each transition type to depression detection. In order to address this question, depression detection performances were compared using self-transition and non-self-transition components. More precisely, *n*-grams that only consist of same acoustic words (i.e. $w_i = w_{i+1} = \cdots = w_j$ in (7)) were termed as self-transitions, whereas the remaining n-grams that consist of different acoustic words were termed as non-self-transitions. The sequence of either self-transitions or non-self-transitions (W_n) was used to extract BoW (9), LDA (10), Word2Vec (15) and GloVe (17) for depression classification, compared with using both two types of transitions, as shown in Table 2.

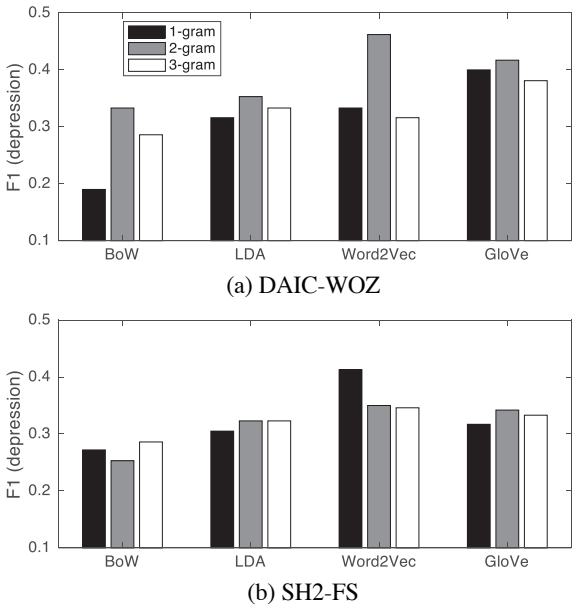
An interesting cross-corpus consistency from Table 2 is that self-transitions were found to be more important than non-self-transitions for both DAIC-WOZ and SH2-FS (except GloVe). This suggests that a potential differentiating factor between depression and non-depression is whether a person's speech is broadly stable in the acoustic space or not, since self-transitions of acoustic words mean that MFCCs from consecutive speech frames tend to stay within the same region of the acoustic space (i.e. same acoustic word) instead of moving to other regions. Another observation is that, for DAIC-WOZ, self-transitions alone yielded much better performances in F1 (depression) than using all components or non-self-transitions only, whereas on SH2-FS, all components performed the best. This could be attributed to the fact that DAIC-WOZ contains cleaner and longer speech than SH2-FS; noise and channel variability might force the aforementioned useful self-transitions to non-self-transitions among different acoustic words, and shorter durations might be unable to provide sufficient information. However, in the former case, it is worth noting that a broad partitioning of the acoustic space (i.e. smaller *K* for acoustic words) is more likely to be robust to noise than a specific one (i.e. larger *K* for acoustic words).

B. Landmark Words – Embedding

Similarly to Section VI-A, this subsection examines the effectiveness of embedding methods in exploiting landmark words for depression detection under various choices of *n* as shown in Fig. 6. Gender normalization [10] was applied to consider gender dependency of landmark occurrence [10], [51].

In Fig. 6, Word2Vec and GloVe outperformed BoW and LDA, with BoW always performing the worst. This again underlines our hypothesis for landmark words that embedding methods like Word2Vec and Glove which learn word similarities in context can better exploit the symbolic landmark words to aid depression detection. Comparing performances in Fig. 4 and 5, landmark words performed weaker than acoustic words on DAIC-WOZ with 0.462 vs. 0.609, yet relatively

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing



(a) DAIC-WOZ

(b) SH2-FS

Fig. 6: Comparison landmark words using different text-based methods.

stronger on SH2-FS with 0.413 vs. 0.395. This may be due to different speech recording environments: acoustic words are built from spectral features, which are more likely to be affected by various handset characteristics on SH2-FS, compared with DAIC-WOZ that contains recordings collected from clean environments. Landmarks, by contrast, may be less susceptible to unwanted variability from diverse handsets and noisy backgrounds.

Moreover, except for the case of Word2Vec + 1-gram, adopting $n \geq 2$ can always produce improvements over $n = 1$, confirming, again, our hypothesis (here for landmark words) that the landmark word sequence is helpful. To this end, we set $n = 2$ for landmark words in all the following experiments, unless specified.

VII. Results: Hybrid Landmark-Acoustic Features

The effectiveness of embedding approaches and the usefulness of considering word sequences in exploiting two types of words, i.e. AW and LW have been shown in Section VI. This section investigates three proposed hybrid approaches to fuse or combine AW and LW: 1) hybrid words (HW) (IV-A); 2) hybrid embeddings (HE) (IV-B); and 3) landmark-dependent acoustic words (LD-AW) (IV-C). Such hybrid frameworks are expected to further aid depression detection performances compared to individual systems (i.e. either AW or LW) due to the different nature of AW and LW: AW relates to 'content' and acoustic space, whereas LW relates to 'boundaries' and speech articulation. One the other hand, AW and LW are similar in that they are both symbolic, which allows the hybrid (combined) system to be investigated using embedding techniques.

A. Fusion of Acoustic Words and Landmark Words

To begin with, the three proposed hybrid approaches were examined, compared with AW-only and with LW-only using BoW, LDA, Word2Vec, and GloVe, as shown in Table 3. The

TABLE 3
F1 (DEPRESSION) SCORES FOR COMPARING VARIOUS FUSION METHODS AGAINST AW-ONLY AND LW-ONLY SYSTEMS.

		Features			
		BoW	LDA	W2V	GloVe
DAIC-WOZ	AW-only	0.421	0.500	0.583	**0.609**
	LW-only	0.333	0.353	0.462	0.417
	HW	0.455	0.519	**0.609**	0.600
	HE	0.320	0.500	0.400	0.375
	LD-AW Fusion	**0.545**	**0.522**	0.560	0.483
SH2-FS	AW-only	0.329	0.376	**0.381**	**0.395**
	LW-only	0.253	0.323	0.350	0.340
	HW	0.253	0.357	0.368	0.375
	HE	**0.359**	**0.417**	0.342	0.324
	LD-AW Fusion	0.333	0.314	0.347	0.317

three hybrid systems include HW (word-level fusion), HE (embedding-level fusion), and LD-AW Fusion (fusion of individual landmark-dependent AW). More specifically, for LD-AW Fusion, SVM binary outputs from individual LD-AW systems were fused via majority voting to predict the final output. In LD-AW Fusion, landmarks 'v' and 'f' were omitted for SH2-FS, and only 'v' was omitted for DAIC-WOZ due to their scarcity. For individual LD-AW systems, dimensionality for LDA topic number, Word2Vec, and GloVe was optimized from among {4, 8, 12, 16, 20, 24, 28, 32, 36, 40} as in other systems. That is, for instance in LDA, we fixed the number of topics to 4 for all individual LD-AW systems, which were later fused. In all systems in Table 3, K was set to 4 for acoustic words and n was set to 2 as per previous experiments. The use of 2-gram allows the interaction between acoustic words and landmark words to be learnt. Gender normalization was applied for HE.

In Table 3, for the DAIC-WOZ corpus, HW produced better results than AW-only and LW-only, except slight performance degradation using GloVe. This suggests that it is beneficial to use the heterogeneous hybrid words, which not only preserve the uniqueness of acoustic and landmark words, but also learn interactions (transitions) between them. However, this does not hold true on SH2-FS, where HW performed more poorly than AW-only, which is presumably due to the weak performance of LW-only on SH2-FS.

As for HE, when BoW and LDA were used, results on SH2-FS demonstrate that it is more effective than HW, achieving 0.417 in F1 score, which outperformed AW-only and LW-only by a large margin. However, HE was less performant than AW-only and LW-only for Word2Vec and GloVe on SH2-FS and all cases on DAIC-WOZ. Overall, the comparisons between HW and HE suggest that in clean speech (DAIC-WOZ), hybrid words are more efficient in exploiting the heterogeneity between two kinds of words, whereas in noisy speech (SH2-FS), fusion at the intermediate level could be helpful. This is sensible as intermediate level representation (e.g. LDA) of acoustic words might be more robust to the channel variability and noise on SH2-FS than acoustic words.

As for LD-AW Fusion, compared with AW-only and LW-only, LD-AW fusion systems only showed improved performances for BoW and LDA on DAIC-WOZ, and did

relatively poorly in other cases. We observed that this was due to fusion of all individual LD-AW systems, some of which are less performant and therefore not necessarily helpful to incorporate.

Overall, fusion or combination of AW and LW can often produce improvements over individual AW-only or LW-only systems, although this is not guaranteed across all settings. Furthermore, note that AW-only has *K* and *n* optimized (Fig. 4 and Fig. 5), and it is expected that an optimal *K* that allows a balance between AW and LW in the proposed hybrid systems could yield further improvements (such optimization of *K* and *n* was shown in Table 5).

B. Landmark-Dependent Acoustic Words (LD-AW)

In the previous experiment, LD-AW fusion has been evaluated. This subsection investigates LD-AW further by 1) examining the performance for LD-AW specific to individual landmarks; and 2) apart from using all components, examining the self-transitions and non-self-transition components, which were found to be important for acoustic words (Table 2). Only the Word2Vec approach was evaluated herein due to its strong performance reported in previous sections. Also, landmarks 'v' and 'f' were omitted for SH2-FS, and only 'v' was omitted for DAIC-WOZ due to their scarcity. As per previous experiments, *K* was set to 4 and *n* was set to 2 for LD-AW.

An advantage of LD-AW is incorporation of articulatory information. For example, LD-AW (g), which is built from spectral features during vocal fold vibrations (see landmark definition of Table 1), can be treated as articulation-based acoustic events specific to vocal fold activities. Similarly, LD-AW (p) is acoustic events based on periodic voicing, LD-AW (s) is related to nasal activities or sonorant-based, LD-AW (b) is burst-based events, and LD-AW (f) is frication-based events. Since each LD-AW carries unique articulatory information, it is reasonable to expect improvements when fusing individual LD-AW systems (which individually suffer from less (unwanted) phonetic variability than AW) to exploit both useful articulatory and acoustic space information. Furthermore, LD-AW is more interpretable than HW and HE, since each landmark reflects certain articulatory activities (Table 1).

There are several observations to note in Table 4. First, comparing LD-AW to AW-only, individual LD-AW systems are more effective than AW-only on DAIC-WOZ when either self-transitions or non-self-transitions are considered; in both cases, 0.632 in F1 was achieved by LD-AW (b) with self-transitions and LD-AW (g) with non-self-transitions. However, the trend that self-transitions are more useful than non-self-transitions concluded from Table 2 does not hold in Table 4. This could be because LD-AW is affected by both articulation (landmark words) and transition (acoustic words). In other words, LD-AW may be indicative of more specific information, i.e., whether a person's speech is stable (self-transitions) or unstable (non-self-transitions), dependent on certain articulatory activities (e.g. vocal fold vibration). For SH2-FS, LD-AW generally performed more poorly than AW-only. Despite this, good performances can still be seen when non-self-transition components were used for LD-AW specific to landmarks p, s, and f.

TABLE 4
F1 (DEPRESSION) COMPARISON OF LD-AW SYSTEMS AND FUSION. WORD2VEC WAS USED TO LEARN THE EMBEDDING REPRESENTATION FROM LD-AWS.

	DAIC-WOZ			SH2-FS		
	All	Self	Non-self	All	Self	Non-self
AW-only	0.583	0.609	0.519	**0.381**	**0.368**	0.366
LD-AW (g)	0.545	0.545	0.632	0.341	0.361	0.228
LD-AW (p)	0.519	0.500	0.500	0.349	0.338	0.361
LD-AW (s)	0.538	0.545	0.364	0.295	0.269	0.352
LD-AW (f)	0.500	0.522	0.545	-	-	-
LD-AW (b)	0.583	**0.632**	0.364	0.315	0.330	0.371
Fusion: All[1]	0.560	0.545	0.588	0.347	0.350	0.312
Fusion: Best[2]	**0.636**	0.556	**0.667**	0.349	0.338	**0.400**

Second, considering self-transitions vs. non-self-transitions, it was consistently found that for DAIC-WOZ, self-transitions were more effective than either all components or non-self-transition components (except LD-AW (b)). However, this was the opposite in the case of the landmark-dependent acoustic words for SH2-FS, where three out of four cases showed improved performances for non-self-transitions.

Third, fusion of LD-AW systems was conducted to combine useful information from different LD-AW systems, which yielded notable improvements over individual LD-AW systems, achieving 0.667 and 0.400 in F1 scores for DAIC-WOZ and SH2-FS, and interestingly when non-self-transitions were used. The results also suggest that it is beneficial to select LD-AW systems before fusion since some LD-AW might be less useful than others. For instance, fusion of the top two performant LD-AW for (p) and (b) provided the largest gain on SH2-FS.

Taken together, the proposed fusion techniques, i.e. HE, HW and LD-AW, have demonstrated efficacy of exploiting articulatory dysfunction in the form of tokens for improved recognition of depressed speech on two different datasets. This is in line with the finding in the literature that articulatory dysfunction is an important aspect associated with depressed speech (discussed in Section II-B). Not only do the token-based events (i.e. acoustic words and landmark words) offer a new way to process speech, but also they offer an effective way to generate articulation-based acoustic features (the acoustic words can be built from any acoustic feature sets instead of MFCCs used in this study).

VIII. RESULTS: OPTIMIZED AND FUSED SYSTEMS

This section presents optimized results for the proposed token-based (i.e. AW and LW) features and three hybrid approaches on the DAIC-WOZ and SH2 (FS) corpora, in comparison with published results in the literature. F1 for depressed (D), F1 for healthy (H), accuracy and Confusion matrix (Conf. Mat.) are

[1] All five LD-AW systems were fused on DAIC-WOZ, whereas all four LD-AW systems were fused on SH2-FS.

[2] For DAIC-WOZ, LD-AW (g), LD-AW (f), and LD-AW (b) were fused, whereas for SH2-FS, LD-AW (p) and LD-AW (b) were fused, both using majority voting.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2949419, IEEE Journal of Selected Topics in Signal Processing

summarized in Table 5.

Unlike previous experiments where K was fixed to 4 and n was fixed to 2, in Table 5, for AW, LW, HW, HE, LD-AW, we optimized the choice of $K \in \{2,3,4,5,6,7,8,9,10\}$ and $n \in \{1,2,3\}$ to examine how best performances the proposed methods can achieve on both corpora. Respectively for DAIC-WOZ and SH2-FS, the optimized systems were chosen as follows

- AW: (BoW, K=4, n=3) and (GloVe, K=4, n=2);
- LW: both (Word2Vec, n=2);
- HW: (GloVe, K=4, n=1) and (Word2Vec, K=9, n=2);
- HE: (GloVe, K=3, n=1) and (LDA, K=4, n=2);
- LD-AW: (LDA, K=4, n=2, landmark p, non-self-transitions) and (Word2Vec, K=6, n=3, landmark p, non-self-transitions);
- LD-AW Fusion: (Word2Vec, K=4, n=2, landmark g, f, b, non-self-transitions) and (GloVe, K=4, n=2, landmark g, b, non-self-transitions).

The summary of optimized results presented in Table 5 shows that token-based methods outperform conventional frame-based acoustic features as reported in the literature (i.e. [45], [52], [53] on DAIC-WOZ and [9], [10] on SH2-FS). Furthermore, the fusion of AW and LW, two different token-based features, can yield improved results using all three proposed hybrid systems, with the best results being 0.667 and 0.417 in F1(depressed) for the DAIC-WOZ and SH2-FS corpora, which are 13% and 15% relative improvements compared with their respective acoustic baselines. Regardless of the technique considered, performance on the SH2-FS dataset was lower than that of DAIC-WOZ. This was due to a number of reasons, including in particular the unsupervised nature of both audio recording and the PHQ-9 questionnaire, background noise and different hardware characteristics.

TABLE 5
A SUMMARY OF THE OPTIMIZED AND FUSED PERFORMANCES, COMPARED WITH PUBLISHED RESULTS.

		F1(D)	F1(H)	Acc.	Conf. Mat.
DAIC-WOZ	AVEC 2016 (A) [47]	0.41	0.58	51.4%	$\begin{bmatrix}12 & 16\\ 1 & 6\end{bmatrix}$
	AVEC 2016 (A+V) [47]	0.58	0.86	77.1%	$\begin{bmatrix}22 & 6\\ 2 & 5\end{bmatrix}$
	DepAudioNet (A) [54]	0.52	0.70	65.7%	$\begin{bmatrix}15 & 13\\ 0 & 7\end{bmatrix}$
	Audio + Gender [55]	0.59	0.87	-	-
	AW	0.636	0.833	77.1%	$\begin{bmatrix}20 & 8\\ 0 & 7\end{bmatrix}$
	LW	0.462	0.682	60.0%	$\begin{bmatrix}15 & 13\\ 1 & 6\end{bmatrix}$
	HW	0.609	0.809	74.3%	$\begin{bmatrix}19 & 9\\ 0 & 7\end{bmatrix}$
	HE	0.609	0.809	74.3%	$\begin{bmatrix}19 & 9\\ 0 & 7\end{bmatrix}$
	LD-AW	0.667	0.857	80.0%	$\begin{bmatrix}21 & 7\\ 0 & 7\end{bmatrix}$
	LD-AW Fusion	0.667	0.885	82.9%	$\begin{bmatrix}23 & 5\\ 1 & 6\end{bmatrix}$
SH2-FS	Acoustic [9]	0.333	0.739	62.5%	$\begin{bmatrix}68 & 37\\ 11 & 12\end{bmatrix}$
	Lmk. Bigram [10]	0.353	0.678	57.0%	$\begin{bmatrix}58 & 47\\ 8 & 15\end{bmatrix}$
	Lmk. Bigram + LDA [10]	0.362	0.555	47.7%	$\begin{bmatrix}42 & 63\\ 4 & 19\end{bmatrix}$
	AW	0.395	0.720	61.7%	$\begin{bmatrix}63 & 42\\ 7 & 16\end{bmatrix}$
	LW	0.413	0.808	71.1%	$\begin{bmatrix}78 & 27\\ 10 & 13\end{bmatrix}$
	HW	0.375	0.792	68.8%	$\begin{bmatrix}76 & 29\\ 11 & 12\end{bmatrix}$
	HE	0.417	0.772	67.2%	$\begin{bmatrix}76 & 29\\ 11 & 12\end{bmatrix}$
	LD-AW	0.413	0.808	71.1%	$\begin{bmatrix}78 & 27\\ 10 & 13\end{bmatrix}$
	LD-AW Fusion	0.407	0.822	72.7%	$\begin{bmatrix}81 & 24\\ 11 & 12\end{bmatrix}$

IX. CONCLUSION

An investigation of a paradigm for heterogeneous token-based systems for the detection of depression from speech has been presented. The tokens, representing both broad acoustic regions and abrupt changes were independently and jointly evaluated in combination with various embedding techniques that are novel in this context.

This paper makes several contributions to the detection of depression and very likely other health disorders that affect speech production. Firstly, the extraction of acoustic words allows for a different kind of analysis for paralinguistic speech processing in the proposed paradigm, one which explicitly takes account of the sequence of acoustic/articulatory changes or events in speech. Secondly, the paradigm also very naturally allows for straightforward combinations and fusion of broad acoustic tokens and articulatory event-based features, which have been less commonly used in speech classification applications, and in this paper a few of a potentially large number of interesting possibilities for combination were explored. The proposed fusion framework exploits the heterogeneous information from both acoustic words and landmark words, which are different in many aspects: 1) fixed time rates vs. irregular time intervals; 2) "content" vs. "boundary" of speech; 3) acoustic space information vs. articulatory information. A key success of the paradigm is the novel application of a wide range of attractive and well-understood tools from text analysis. Another interesting aspect of the paradigm is the very compact and privacy-preserving nature of the tokens, which could be communicated between a thin client and server extremely efficiently with minimal possibility of reconstructing the original speech.

Evaluation of various system configurations from within the paradigm showed that a small value of K (the number of broad acoustic regions) and a setting of n to 2 or 3 (bigrams or trigrams) are good choices, confirming the importance of sequence information (i.e. by contrast with n = 1). Results also showed that embedding using both Word2Vec and GloVe yields a significant improvement over the more straightforward BoW approach. Treating self-transitions and non-self-transitions in acoustic words separately was found to be advantageous, which suggests that there may be an advantage in processing speech 'content' and speech 'boundary' information differently.

Among the various system configurations proposed herein, landmark-dependent acoustic words were the most promising and effective way to combine information from heterogeneous feature types, for both datasets tested. This landmark-dependent approach is essentially a hybrid system that uses landmarks as region selection methods to derive information specific to one type of articulation at a time. AWs and LWs carry different information; the AW partitions the acoustic space into one

token per frame, whereas the LWs denote the abrupt changes in speech articulation. The hybrid combination of the AWs and LWs therefore allows different information to be exploited, more precisely, incorporation of articulatory dysfunction into conventional acoustic features.

Importantly, systems designed within this new paradigm using only linear SVM rivaled or exceeded state-of-the-art results on two very different depressed speech datasets, including one comprising recordings from diverse smartphones in noisy, naturalistic settings. The proposed methods and framework are significant in the way that they work for two drastically different datasets (i.e. clean vs. noisy, long duration vs. short duration, single channel vs. multiple channels for DAIC-WOZ and SH2-FS), reinforcing the feasibility and effectiveness of the proposed methods.

The proposed paradigm has rich possibilities for future work. There is strong potential for integration of AWs and LWs with text transcripts, which is very natural given that the acoustic information is already represented symbolically, where these transcripts are available (e.g. DAIC-WOZ has given very strong depression detection results using text only [34], [32], [56], [57]). The features proposed, in particular those associated with articulatory transitions, may be more robust to noise than conventional acoustic features, and this possibility warrants deeper investigation. Finally, in contrast with the simple linear SVM classifier used throughout this paper, a range of more sophisticated back-ends can be trialed, to increase detection accuracy further.

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 Commun.*, vol. 71, pp. 10–49, Jul. 2015.

[2] J. Walker *et al.*, "The Prevalence of Depression in General Hospital Inpatients: A Systematic Review and Meta-Analysis of Interview Based Studies," *Psychol. Med.*, 2018.

[3] 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.

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

[5] H. Ellgring and K. R. Scherer, "Vocal indicators of mood change in depression," *J. Nonverbal Behav.*, vol. 20, no. 2, pp. 83–110, 1996.

[6] 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," *Psychiatr. Rehabil. J.*, vol. 38, no. 3, pp. 218–226, 2015.

[7] F. Gravenhorst *et al.*, "Mobile phones as medical devices in mental disorder treatment: an overview," *Pers. Ubiquitous Comput.*, vol. 19, no. 2, pp. 335–353, 2015.

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

[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] Z. Huang, J. Epps, and D. Joachim, "Speech Landmark Bigrams for Depression Detection from Naturalistic Smartphone Speech," in *ICASSP*, 2019, pp. 5856–5860.

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

[12] Z. Huang, J. Epps, and D. Joachim, "Investigation of Speech Landmark Patterns for Depression Detection," *IEEE Trans. Affect. Comput.*, 2019.

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

[14] Z. Huang and J. Epps, "An Investigation of Partition-based and Phonetically-aware Acoustic Features for Continuous Emotion Prediction from Speech," *IEEE Trans. Affect. Comput.*, 2018.

[15] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *INTERSPEECH*, 2016, pp. 495–499.

[16] F. Ringeval *et al.*, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," *Proc. 2018 Audio/Visual Emot. Chall. Work.*, pp. 3–13, 2018.

[17] Z. Huang and J. Epps, "A PLLR and Multi-stage Staircase Regression Framework for Speech-based Emotion Prediction," in *ICASSP*, 2017, pp. 5145–5149.

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

[19] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models - Analysis and normalisation," in *ICASSP*, 2013, pp. 7522–7526.

[20] 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 Trans. Affect. Comput.*, vol. 7, no. 1, pp. 59–73, 2016.

[21] A. Michael *et al.*, "Emotional bias and inhibitory control processes in mania and depression," *Psychol. Med.*, vol. 29, no. 6, pp. 1307–1321, 1999.

[22] 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.

[23] A. C. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, p. 42, 2011.

[24] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 142–150, 2013.

[25] J. H. Martin and D. Jurafsky, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2008.

[26] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," *J. ofMachine Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[28] M. Hoffman, D. Blei, and F. Bach, "Online Learning for Latent Dirichlet

Allocation," in *NIPS*, 2010, pp. 1–9.

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

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *arXiv*, 2013, pp. 1–12.

[31] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[32] J. R. Williamson *et al.*, "Detecting Depression using Vocal, Facial and Semantic Communication Cues," *Proc. 6th Int. Work. Audio/Visual Emot. Chall. - AVEC '16*, pp. 11–18, 2016.

[33] P. Lopez-Otero, L. Docio-Fernandez, A. Abad, and C. Garcia-Mateo, "Depression detection using automatic transcriptions of de-identified speech," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, pp. 3157–3161, 2017.

[34] 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," *Proc. 6th Int. Work. Audio/Visual Emot. Chall. - AVEC '16*, pp. 89–96, 2016.

[35] J. Huang *et al.*, "Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks," *Avec*, vol. 18, pp. 57–64, 2018.

[36] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal Measurement of Depression Using Deep Learning Models," pp. 53–59, 2017.

[37] N. Cummins, S. Amiriparian, S. Ottl, M. Gerczuk, and M. Schmitt, "MULTIMODAL BAG-OF-WORDS FOR CROSS DOMAINS SENTIMENT ANALYSIS Chair of Embedded Intelligence for Health Care and Wellbeing , University of Augsburg , Germany Machine Intelligence & Signal Processing Group , Technische Universit ¨ at M ¨ Chair of Complex an," pp. 4954–4958, 2018.

[38] S. Amiriparian, M. Schmitt, N. Cummins, K. Qian, F. Dong, and B. Schuller, "Deep Unsupervised Representation Learning for Abnormal Heart Sound Classification," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2018-July, pp. 4776–4779, 2018.

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

[40] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[41] J. Han, Z. Zhang, M. Schmitt, Z. Ren, F. Ringeval, and B. Schuller, "Bags in Bag: Generating context-aware bags for tracking emotions from speech," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. September, pp. 3082–3086, 2018.

[42] A. Vaswani *et al.*, "Attention Is All You Need," no. Nips, 2017.

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

[44] X. Rong, "word2vec Parameter Learning Explained," pp. 1–21, 2014.

[45] J. Joshi, R. Goecke, and A. D. Michael, "Multimodal Assistive Technologies for Depression Diagnosis and Monitoring," pp. 1–18.

[46] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals," in *Proceedings of the 3rd ACM international workshop on AVEC, ACM MM*, 2013, pp. 11–20.

[47] 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.

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

[49] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[50] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[51] K. Ishikawa, J. MacAuslan, and S. Boyce, "Toward clinical application of landmark-based speech analysis: Landmark expression in normal adult speech," *J. Acoust. Soc. Am.*, vol. 142, no. 5, pp. 441–447, 2017.

[52] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825--2830, 2011.

[53] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS*, 2010, pp. 45–50.

[54] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An Efficient Deep Model for Audio based Depression Classificatio," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, 2016, pp. 35–42.

[55] A. Pampouchidou *et al.*, "Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text," *Proc. 6th Int. Work. Audio/Visual Emot. Chall. - AVEC '16*, pp. 27–34, 2016.

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

[57] T. Dang *et al.*, "Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017," in *ACM Multimedia, AVEC '17*, 2017, pp. 27–35.