Speech Landmark Bigrams for Depression Detection from Naturalistic Smartphone Speech

Zhaocheng (David) Huang, Julien Epps, Dale Joachim

Outline

- Motivation
- Related Work
  - Speech articulation affected by depression → Speech Landmarks
- Proposed Features based on landmark bigrams
  - Bigram-count
  - LDA-bigram
- Dataset
  - The SH2 Corpus
- Experimental Settings
- Results
- Conclusions
- Future Work

Motivation

- Depression is a big burden to the society.

- To date, depression detection has primarily focused on laboratory-controlled clean speech samples, which is atypical in naturalistic environments.

- Smartphones: offer huge potential in spreading depression screening, which however has some challenges.
  - environmental noise
  - various handset characteristics

- Speech Articulation → Speech Landmarks

Related Work

- Speech articulation affected by depression
  - cognitive impairment,
  - phonation and articulation errors,
  - articulatory incoordination,
  - disturbances in muscle tension, phoneme rates,
  - altered speech quality and prosody.

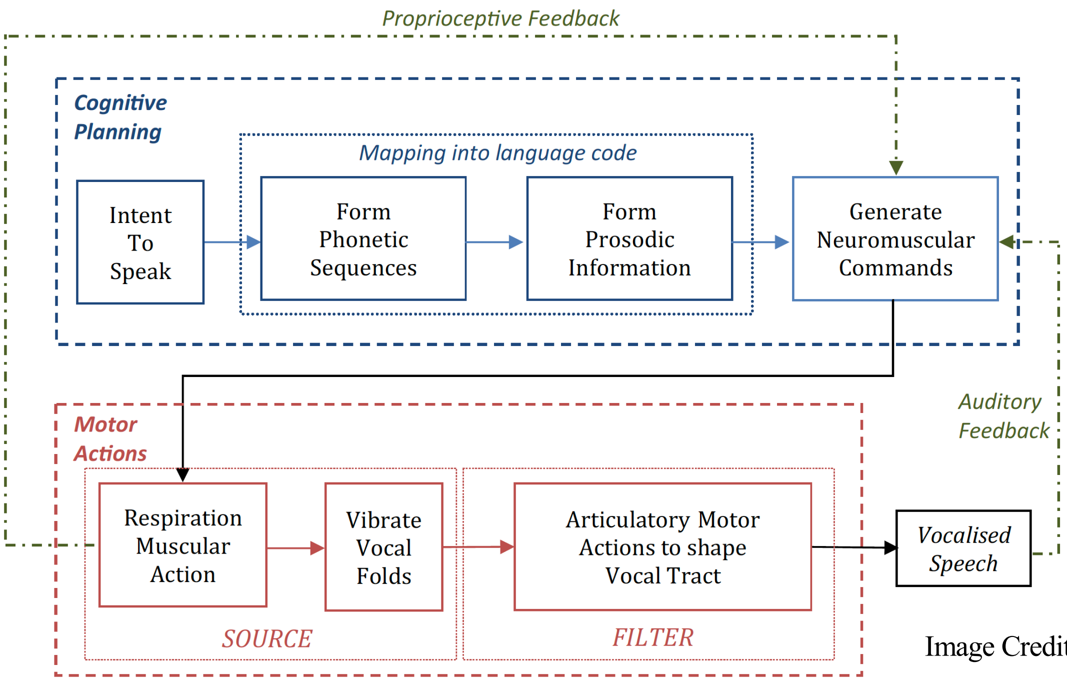


Image Credit: Cummins et al, 2015

Cummins, N., S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015.
Stevens, K. N., S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, “Implementation of a Model for Lexical Access based on Features,” in ICSLP, 1992, no. October, pp. 499–502.

Related Work

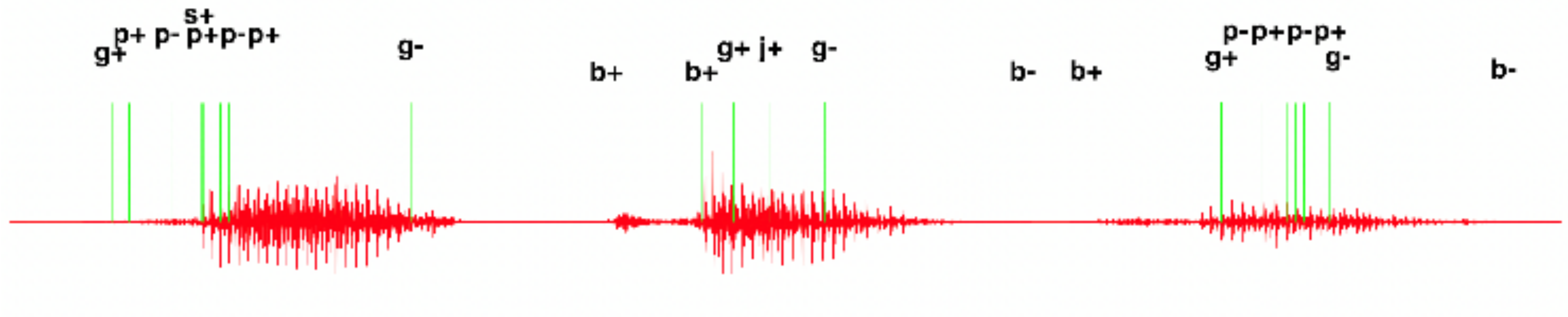
- Speech articulation affected by depression
  - cognitive impairment,
  - phonation and articulation errors,
  - articulatory incoordination,
  - disturbances in muscle tension, phoneme rates,
  - altered speech quality and prosody.

- Speech landmarks are symbols associated with speech articulation
  - Introduced by K. Stevens in 1992
  - Linguistic or lexical:
    - Speech recognition [Park 2002; Stevens et al 2002; Johnson et al 2004]
  - Paralinguistic:
    - Parkinson's disease and sleep deprivation [Ishikawa et al 2017]
    - Emotion recognition [Dai et al 2008]
    - Vocalization Age [Fell et al 2002]

Cummins, N., S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015.
Stevens, K. N., S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a Model for Lexical Access based on Features," in ICSLP, 1992, no. October, pp. 499–502.

What are speech Landmarks ?

- Symbols about articulatory changes
  - Determined based on energy changes across several frequency bands and multiple time scales

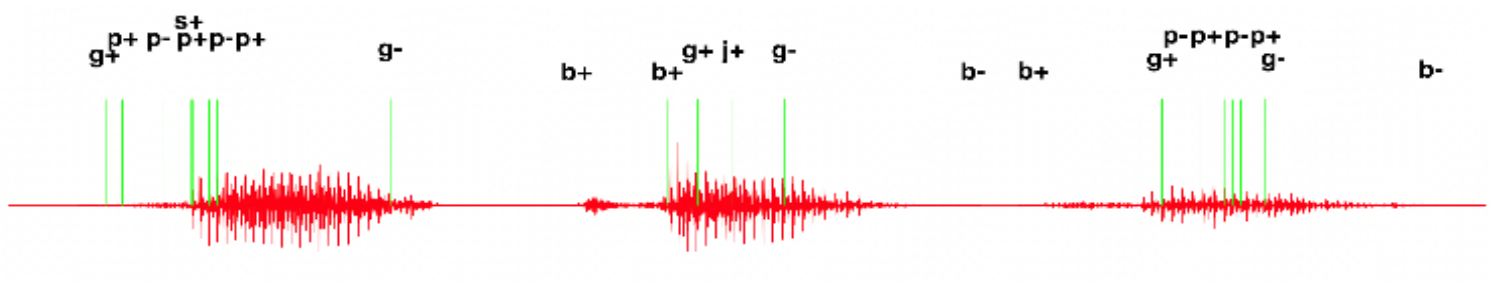


Landmark	Description
g	sustained vibration of vocal folds starts (+) or ends (−)
p	sustained periodicity begins (+) or ends (−)
s	opening (+) or closing (−) of the velopharyngeal port during a sonorant sound
f	frication onset (+) or offset (-)
v	voiced frication onset (+) or offset (-)
b	onset (+) or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

Landmarks → Landmark Bigrams

Landmark Bigrams: (g+, p+), (p+, p-), (p-, s+), (s+, p+) …… (g-, b-)

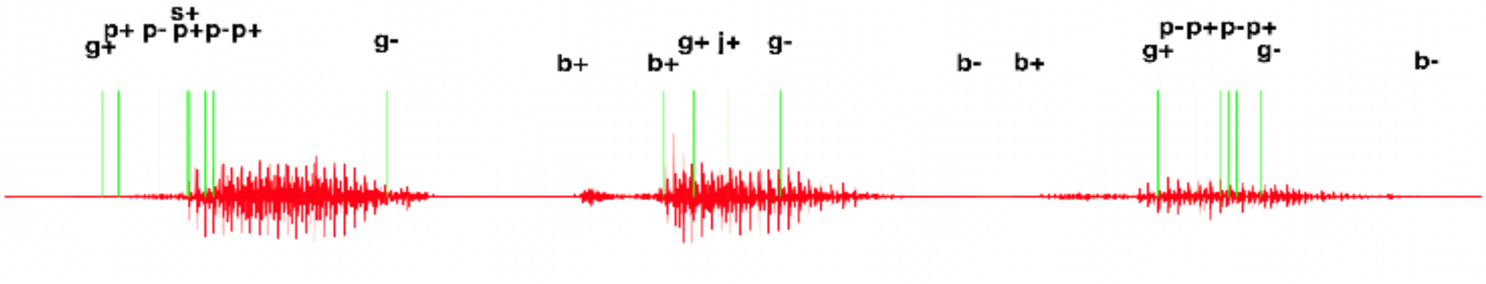
Landmarks: (g+), (p+), (p-), (s+), (p+) …… (p-), (p+), (g-), (b-)



- More complex patterns about speech articulation
- Transitions from one landmark to another – richer information

Proposed Landmark Features – Bigram-count

Landmark Bigrams: (g+, p+), (p+, p-), (p-, s+), (s+, p+) …… (g-, b-)



- Count how many times each bigram occurs
- Concatenate all counts

$$\boldsymbol{c} = [c^{g+,g+}, \ldots, c^{i,j}, \ldots, c^{b-,b-}]$$

Proposed Landmark Features – LDA-bigram

- Latent Dirichlet Allocation (LDA)
  - LDA for text → latent topic modelling
  - Why LDA for landmark bigrams? → latent articulatory events

Text

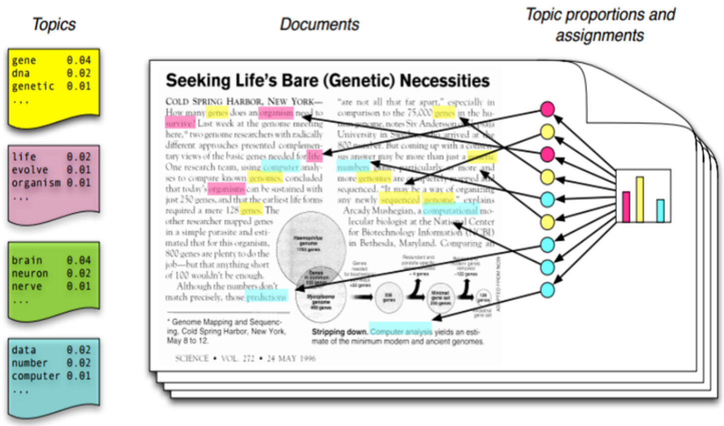
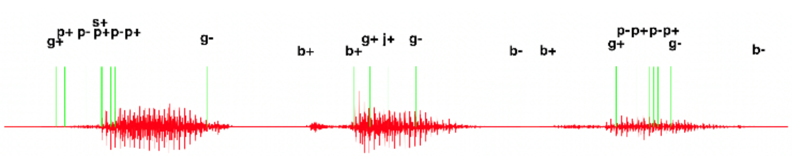


Image Credit: Blei 2010

Document → topics (e.g. sports)
topic → words (e.g. football)

<u>LDA gives a vector of probabilities for latent topics in document.</u>

Speech



Speech → articulation (e.g. vocal fold)
articulation → bigrams (e.g. "g+,g-")

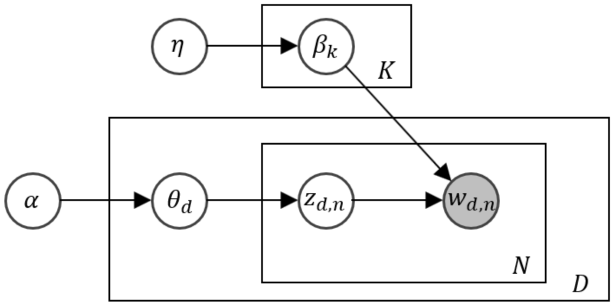
<u>LDA gives a vector of probabilities for latent articulatory events in speech.</u>

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

Proposed Landmark Features – LDA-bigram

- LDA-bigram
  - N bigram, K events, D speech files.
    - $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha) = \{\theta_{d,1}, \dots, \theta_{d,k}, \dots, \theta_{d,K}\}$, $\sum_{i=1}^{K} \theta_{d,i} = 1$ → *speech-articulation*
    - $\beta_k \sim \mathrm{Dir}(\eta) = \{\beta_{k,1}, \dots, \beta_{k,n}, \dots, \beta_{k,N}\}$, $\sum_{n=1}^{N} \beta_{k,n} = 1$ → *articulation-bigram*
    - $w_{d,n} \sim \mathrm{Multi}(\beta_{z_{d,n}=k})$ → *speech-articulation-bigram*
  - Overall, $z_{d,n}$, β_k, and θ_d together describe relationships for *speech-articulation-bigram*, similar to *document-topic-word in topic modelling*

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \alpha, \eta)$$



- Variational Bayesian Inference

$$q(\beta_k) \sim \mathrm{Dir}(\lambda_k), q(\boldsymbol{\theta}_d) \sim \mathrm{Dir}(\gamma_d), q(z_{d,n} = k) \sim \mathrm{Multi}(\phi_{d,n}^k)$$

  - Training

$$\phi_{d,n}^k \propto \mathbb{E}_{q(\theta_d)}[\log \theta_{d,k}] + \mathbb{E}_{q(\beta_k)}[\log \beta_{k,w_{d,n}}]$$

$$\gamma_{d,k} = \alpha + \textstyle\sum_w c_{d,w}\, \phi_{d,n}^k \,,\ \lambda_{k,w} = \eta + \textstyle\sum_d c_{d,w}\, \phi_{d,n}^k$$

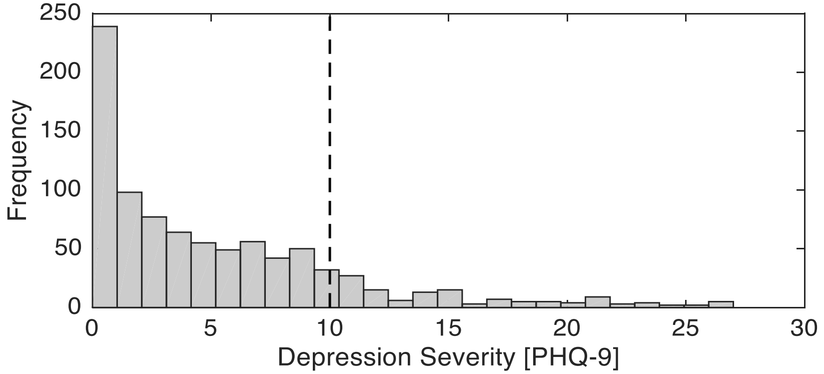
  - Testing: for a new speech file d^*

$$\gamma_{d^*,k} = \alpha + \textstyle\sum_w c_{d^*,w}\, \phi_n^k \,,\ \theta_{d^*} \sim \mathrm{Dirichlet}(\gamma_{d^*,1}, \dots, \gamma_{d^*,K})$$

θ_{d^*} gives a vector of probabilities for all latent articulatory events in each speech file

Dataset – the SH2 corpus [Huang et al, 2018]

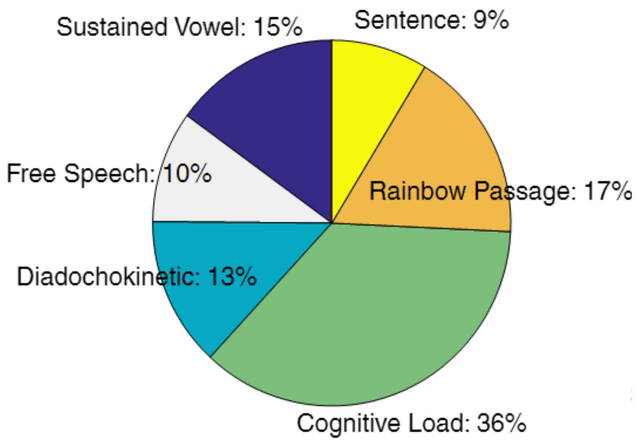
- The SH2 corpus
  - Naturalistic: a variety of noises (e.g. office, restaurant, background TV noise, etc.); 28 device manufacturers.
  - 16 hours of speech; 887 speakers (450 males); 5937 voice recordings (sampled at 44.1kHz).
  - Six elicitation tasks
  - self-assessed Patient Health Questionnaire (PHQ-9)
    - Healthy: [0, 9]
    - Depressed: [10, 27]
    - There are 695 speakers (122 are depressed) for training and 192 speakers (35 are depressed) for testing.



Huang, Z., 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.

Dataset – the SH2 corpus [Huang et al, 2018]

- Elicitation Tasks
  - Cognitive Load
    - Stroop test
  - Free Speech
    - Free response to questions like "what is the weather like outside"
  - Rainbow Passage
    - "When the sunlight strikes raindrops in the air, … with little or no green or blue"
  - Harvard Sentence
    - "The birch canoe slid on the smooth planks.", etc.
  - Sustained Vowel
    - "ahh.."
  - Diadochokinetic
    - "PaTaKa"



UNSW SYDNEY

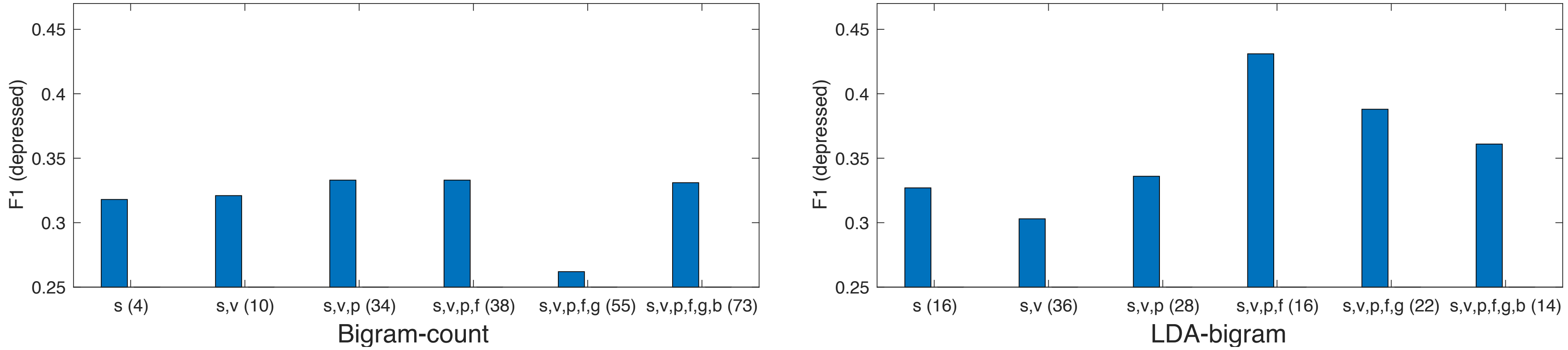
Experimental Settings

- The SH2 corpus
- Classification Model:
  - Linear SVM, with optimized C value from 3-fold cross validation within the training set.
- Performance Metric
  - F1 score (depression) (chance=0.267), Unweighted Average Recall (UAR), Accuracy, Confusion Matrix.
- Speech Landmarks were extracted using the SpeechMark toolkit [Boyce et al 2012].
- LDA-bigram
  - The LDA #topic was optimized from 2 to 40, unless specified.
    - i.e. number of latent articulatory events.

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

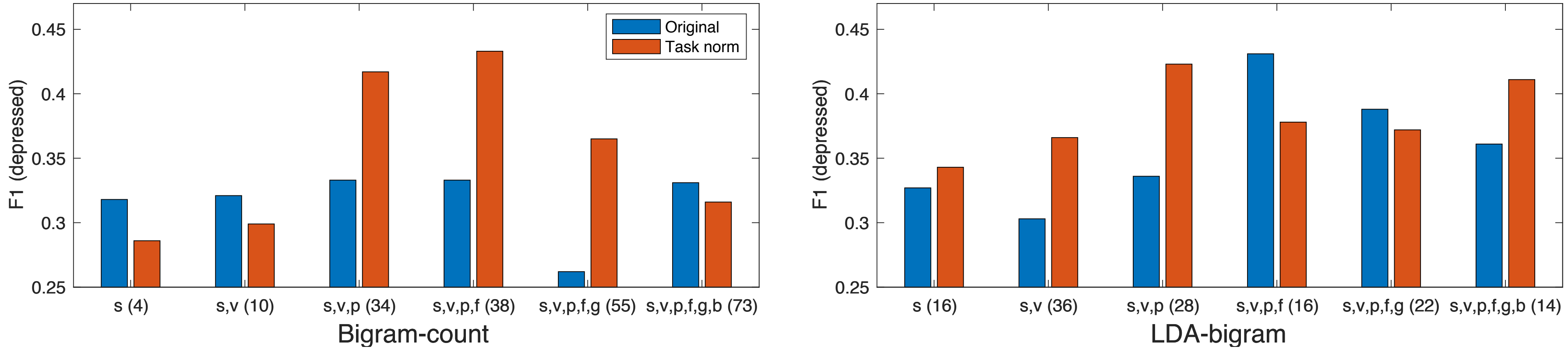
Experimental Results

- How well the proposed features perform?
  - Landmarks were added one-by-one for choosing effective bigrams
  - LDA-bigram tends to be better
  - It is important to tailor landmark choices



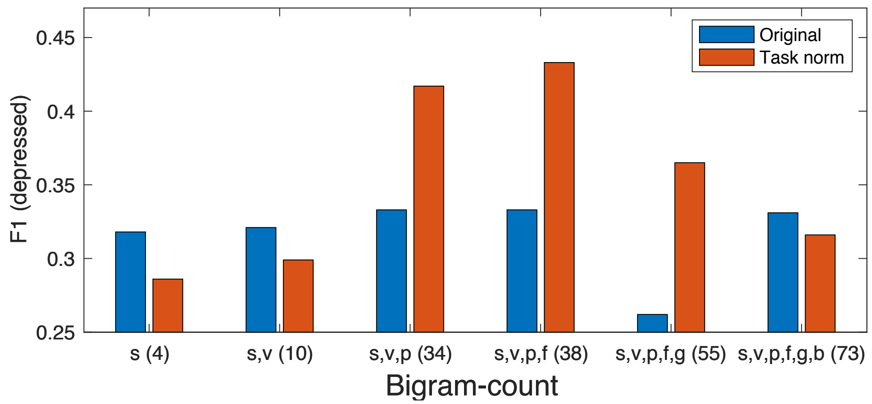
Experimental Results

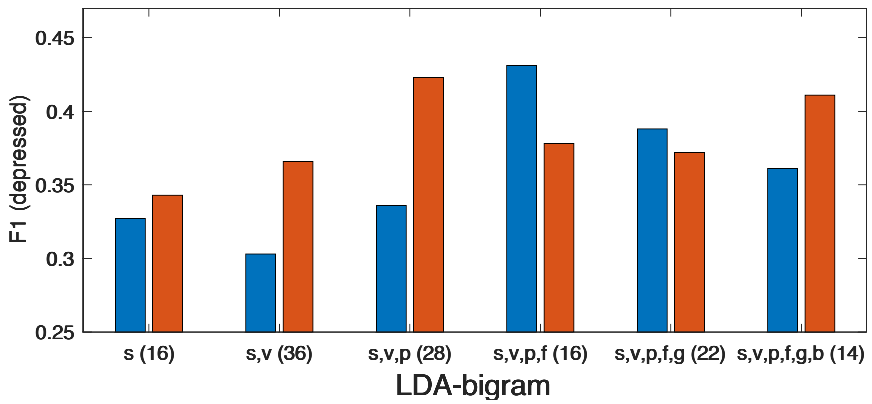
- How well the proposed features perform?
  - Landmarks were added one-by-one for choosing effective bigrams
  - LDA-bigram tends to be better
  - It is important to tailor landmark choices
- Remove task dependency
  - Task norm: z-normalization specific to each task.
  - Landmarks are specific to different elicitation tasks



Experimental Results

- How well the proposed features perform?
  - Landmarks were added one-by-one for choosing effective bigrams
  - LDA-bigram tends to be better
  - It is important to tailor landmark choices
- Remove task dependency
  - Task norm: z-normalization specific to each task.
  - Landmarks are specific to different elicitation tasks
- How about optimizing landmark choices for each elicitation task?

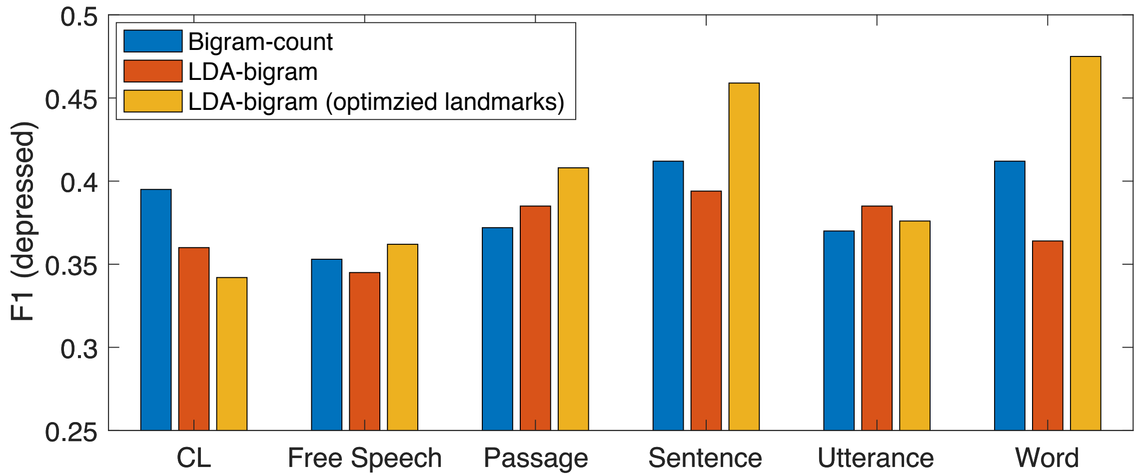




Experimental Results

- Landmark bigram features optimized for elicitation tasks
  - Bigram-count with tailored landmark choices
  - LDA-bigram with the same landmark choices as bigram-count
  - LDA-bigram, #topic =4, tailored landmark choices
  - It is beneficial to optimize landmark choices for both Bigram-count (1st column) and LDA-bigram (3rd column) within each task.

- How about fusing individual elicitation tasks together?



Experimental Results

- Fusion of elicitation tasks.
  - Majority voting of binary outputs from individual tasks
  - The proposed features based on landmark bigrams are effective, compared with acoustic baseline.

		F1 (D)	Accuracy	UAR	Confusion Matrix
	Baseline [Huang et al 2018]: Acoustic features	0.422	72.9%	0.657	$\begin{bmatrix} 121 & 36 \\ 16 & 19 \end{bmatrix}$
Same landmarks across all tasks	Bigram-count#	0.433	71.4%	0.669	$\begin{bmatrix} 116 & 41 \\ 14 & 21 \end{bmatrix}$
	LDA-bigram#	0.431	65.6%	0.679	$\begin{bmatrix} 101 & 56 \\ 10 & 25 \end{bmatrix}$
Tailored landmarks for each task	Bigram-count*	0.506	78.7%	0.714	$\begin{bmatrix} 130 & 27 \\ 14 & 21 \end{bmatrix}$
	LDA-bigram*	0.549	78.7%	0.758	$\begin{bmatrix} 126 & 31 \\ 10 & 25 \end{bmatrix}$

Huang, Z., 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.

Experimental Results

- Fusion of elicitation tasks.
  - Majority voting of binary outputs from individual tasks
  - The proposed features based on landmark bigrams are effective, compared with acoustic baseline.
  - Performances were significantly improved when fusing individual tasks with tailored landmarks.

		F1 (D)	Accuracy	UAR	Confusion Matrix
	Baseline [Huang et al 2018]: Acoustic features	0.422	72.9%	0.657	$\begin{bmatrix} 121 & 36 \\ 16 & 19 \end{bmatrix}$
Same landmarks across all tasks	Bigram-count#	0.433	71.4%	0.669	$\begin{bmatrix} 116 & 41 \\ 14 & 21 \end{bmatrix}$
	LDA-bigram#	0.431	65.6%	0.679	$\begin{bmatrix} 101 & 56 \\ 10 & 25 \end{bmatrix}$
Tailored landmarks for each task	Bigram-count*	0.506	78.7%	0.714	$\begin{bmatrix} 130 & 27 \\ 14 & 21 \end{bmatrix}$
	LDA-bigram*	0.549	78.7%	0.758	$\begin{bmatrix} 126 & 31 \\ 10 & 25 \end{bmatrix}$

Huang, Z., 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.

Conclusions

- Two novel sets of features based on speech landmark bigrams for depression detection under naturalistic environment
  - Bigram-count
  - LDA-bigram

- Novel paradigm with potential
  - Robustness & interpretability

- Significances:
  - First study to apply landmark bigrams for depression detection, which is promising.
  - Large number of speakers (887 in total)
  - No gap between PHQ-9 in determining the healthy and depressed.

Future Work

- A new paradigm in processing speech in symbols.
  - In-depth analysis and interpretability
  - Symbolic → NLP methods
  - We looked at count, how about duration (timing)?

- Applicable to other health disorders
  - Alzheimer's disease
  - Parkinson disease
  - Bipolar disease
  - Dementia
  - Vocal disorders
    - Dysarthria, Dysphonia, Laryngitis, etc.

THANK YOU

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