Exploiting Vocal Tract Coordination using Dilated CNNs for Depression Detection in Naturalistic Environments

Zhaocheng (David) Huang, Julien Epps, Dale Joachim

Outline

- Motivation
- Related Work
  - Speech articulation affected by depression → VTC features by MIT LL
- Proposed FVTC-CNN Framework
- Dataset
  - The SH2-FS Corpus
  - The DAIC-WOZ Corpus
- Experimental Settings
- Results
- Conclusions

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 [Huang et al. 2019a, 2019b, 2020]
  - Vocal Tract Coordination (VTC) [Williamson et al., 2013, 2014]

- Deep Learning
  - towards learning speech articulation information → improved interpretability
  - exploit big data

Related Work

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

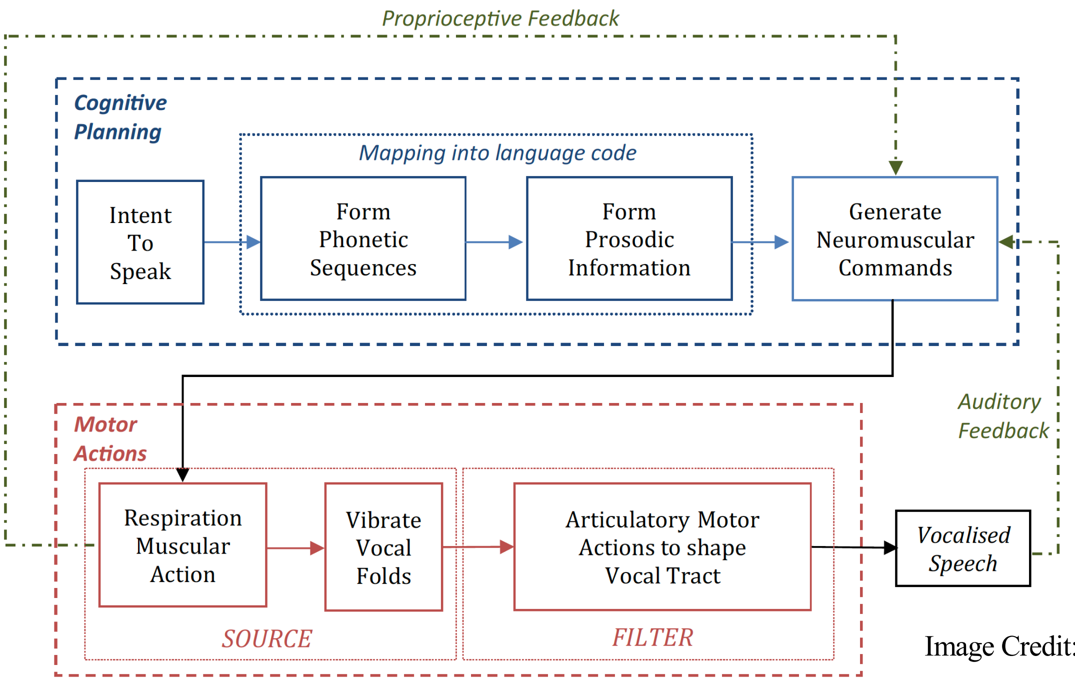


Image Credit: Cummins et al, 2015

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.

Related Work

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

- Speech Articulation-based Features
  - Vowel space area [Scherer et al., 2016]
  - Speech landmarks [Huang et al. 2019a, 2019b, 2020]
  - Vocal Tract Coordination (VTC) Features [Williamson et al., 2013, 2014]
    - Won the AVEC 2013 & 2014 Challenges on Depression Severity Prediction
    - Vocal tract parameters are less “coordinated” (correlated) for depressed speakers than for healthy speakers.

Scherer, S., 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.

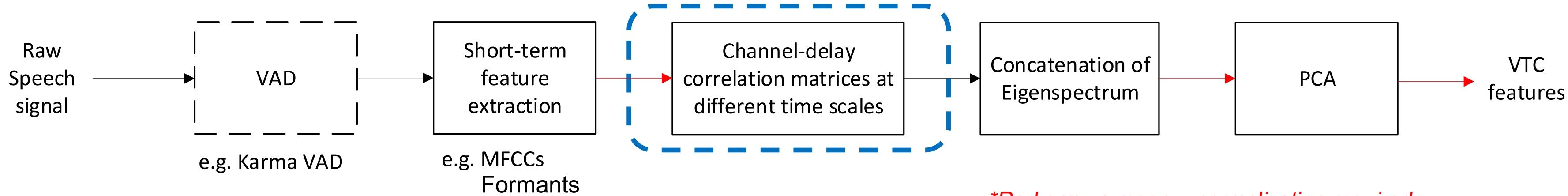
Huang, Z., J. Epps, and D. Joachim, “Investigation of Speech Landmark Patterns for Depression Detection,” *IEEE Trans. Affect. Comput. 2019, to appear*

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

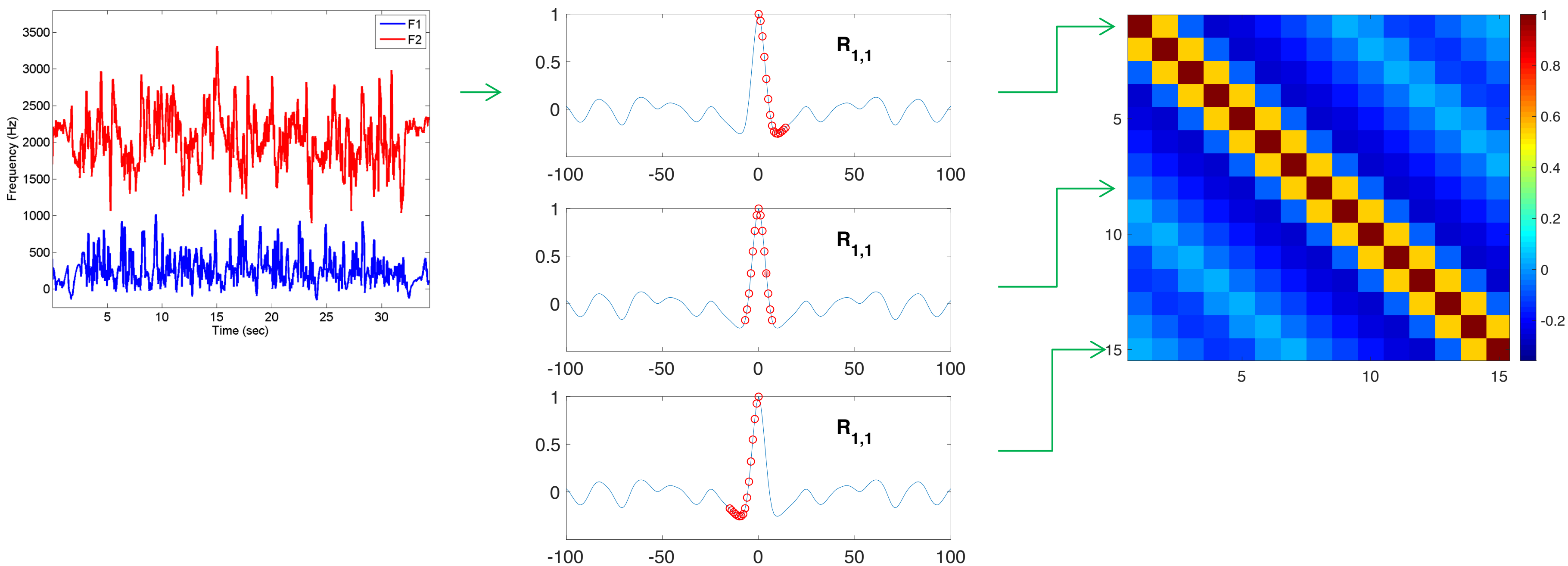
Williamson, J., T. Quatieri, and B. Helfer, “Vocal and facial biomarkers of depression based on motor incoordination and timing,” in *Proceedings of the 4th International Workshop on AVEC, ACM MM*, 2014.

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.

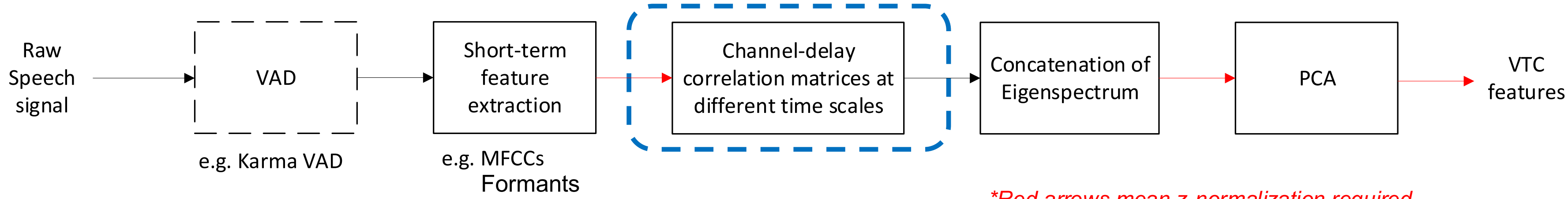
What are VTC features? [Williamson et al., 2013]



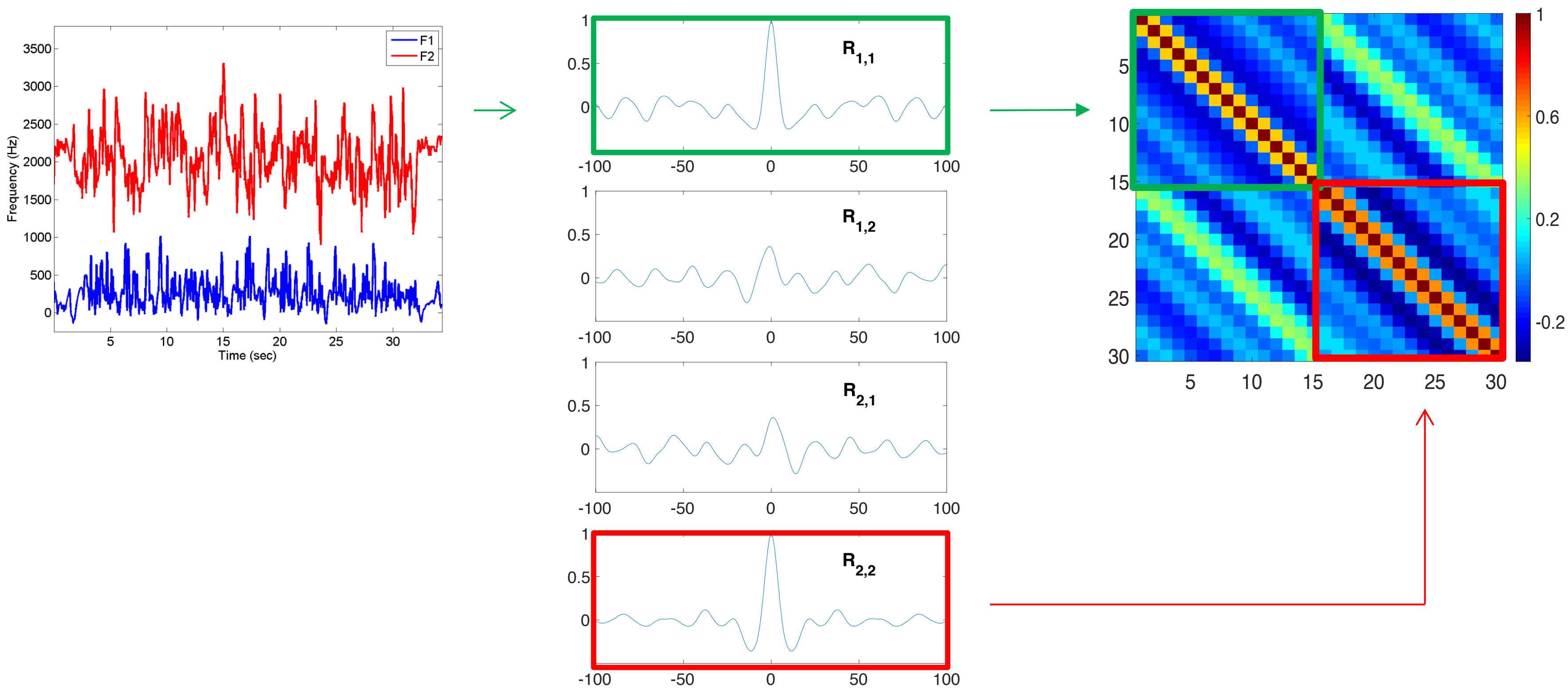
**Red arrows mean z-normalization required*



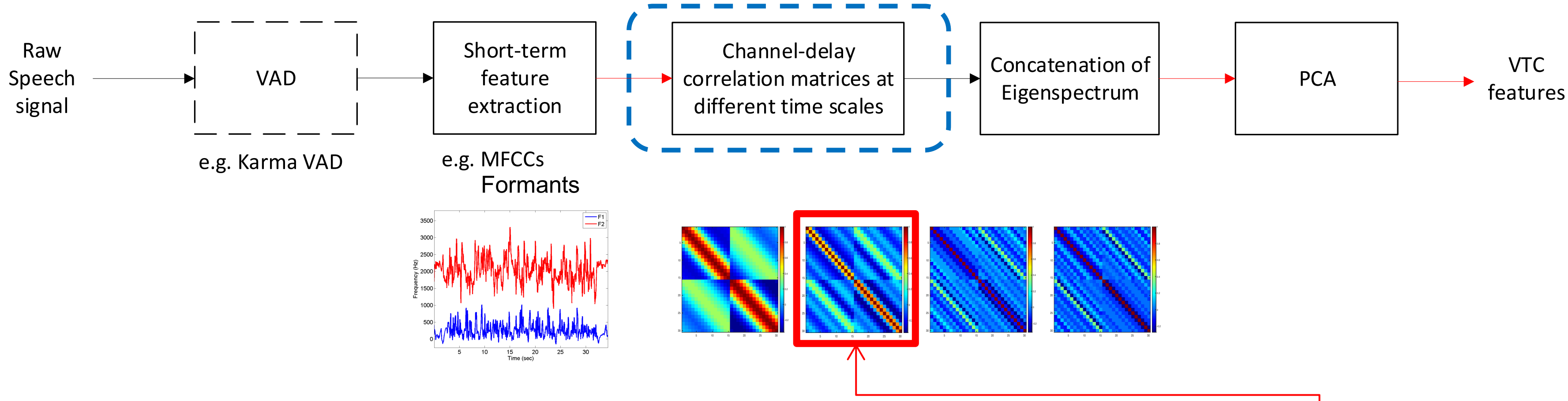
What are VTC features?



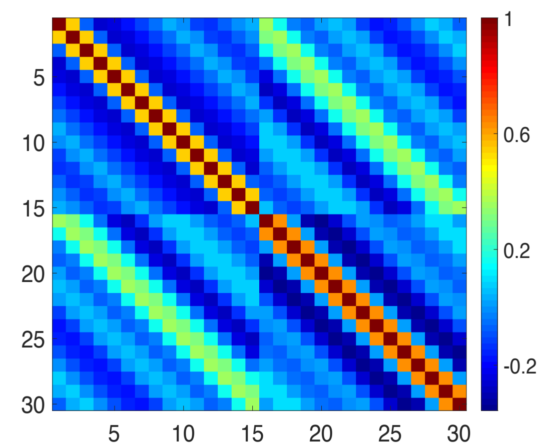
**Red arrows mean z-normalization required*



What are VTC features?



- Why VTC features work?
  - A knowledge-driven elegant framework for capturing vocal tract coordination using delay correlations of feature contours.

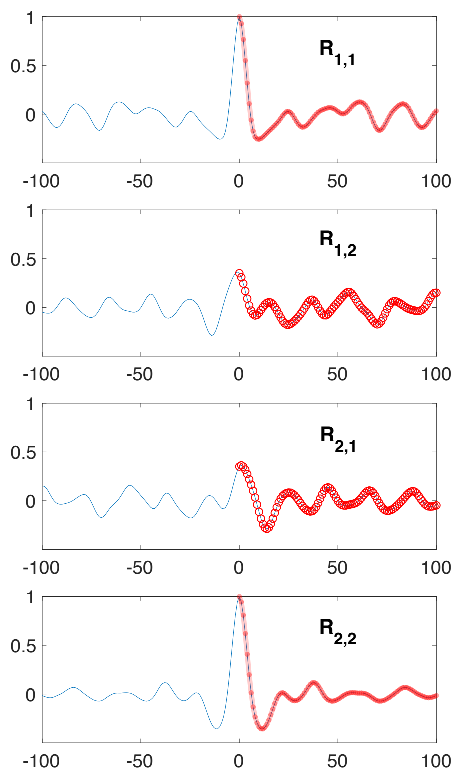


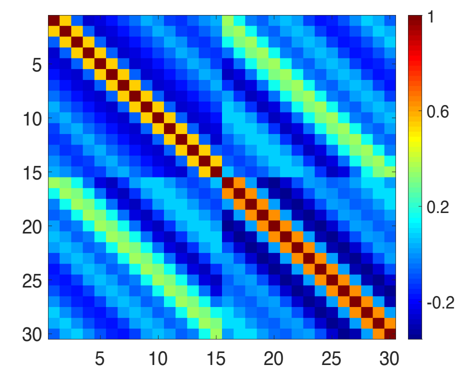
- Any Limitations → YES!
  - Repeated sampling
  - Discontinuities
  - Eigenvalues + PCA may not be the most effective way to decompose useful information.

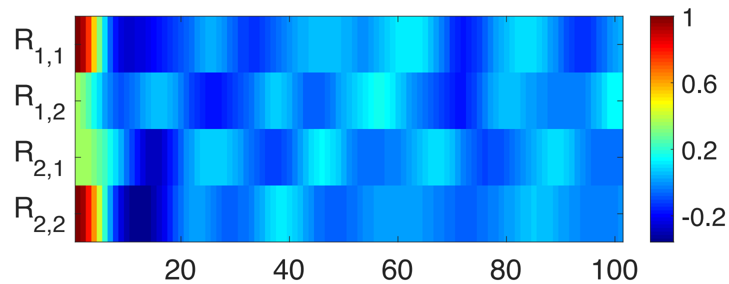
Proposed FVTC-CNN Framework

Any Solutions→ YES!

- Repeated sampling → Use all correlations once.
- Discontinuities → Learn each block individually
- Eigenvalues + PCA may not be the most effective way to decompose useful information. → Deep learning





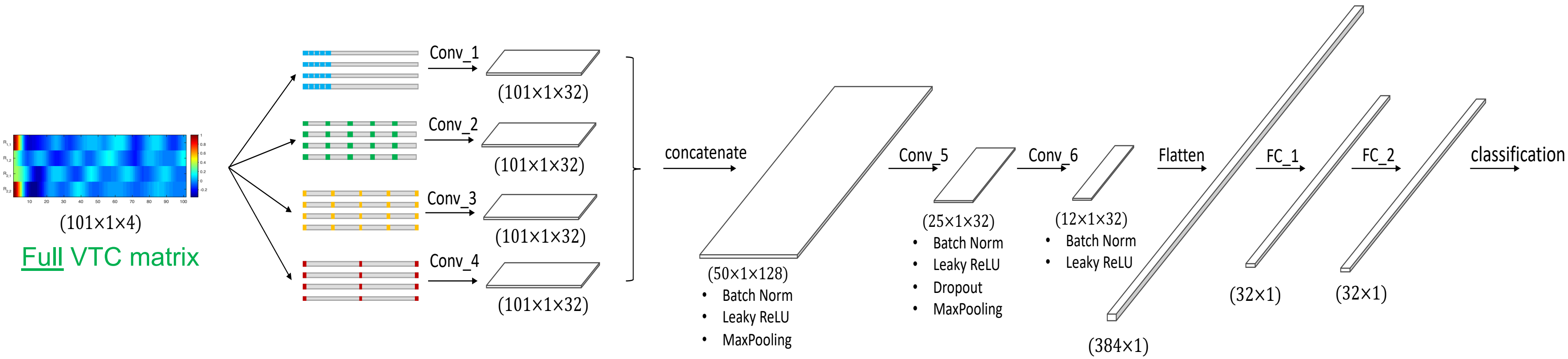


<u>Full</u> VTC matrix

Proposed FVTC-CNN Framework

Any Solutions→ YES!

- Repeated sampling → Use all correlations once.
- Discontinuities → Learn each block individually
- Eigenvalues + PCA may not be the most effective way to decompose useful information. → Deep learning



Dilated Convolutional Neural Nets

- A number of additional advantages:
  - Capable of capturing changes at different time scales
  - CNNs learnt for discrimination, whereas PCA learns based on variance.
  - Handle high dimensionality
  - Scalability on larger datasets

Datasets

- The SH2-FS corpus [Huang et al., 2018]
  - Naturalistic: a variety of noises (e.g. office, restaurant, background TV noise, etc.); 23 device manufacturers; short durations.
    - Averaged recording duration: 20.5 ± 10.2s
  - Self-assessed Patient Health Questionnaire (PHQ-9)
    - Healthy: [0, 9] vs. Depressed: [10, 27]
    - There are 438 speakers (74 are depressed) for training and 128 speakers (23 are depressed) for testing.

- The DAIC-WOZ corpus [Gratch et al., 2014]
  - Laboratory-based: clean, single channel, long duration.
    - Averaged recording duration: 446.9 ± 227.0s
    - There are 107 speakers (21 are depressed) for training and 35 (7 are depressed) for testing.

Binary Classification Problem

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.
Gratch, J., R. Artstein, G. Lucas, G. Stratou, S. Scherere, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. (Skip) Rizzo, et al., "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2014, pp. 3123–3128.

Experimental Settings

- Datasets: SH2-FS and DAIC-WOZ
- Low-level Descriptors for VTC and FVTC-CNN
  - 3 Formants
  - 13 Spectral Centroid Frequencies (SCF)
  - 16 MFCCs
  - 16 delta MFCCs
- Hyperparameters for dilated CNNs:
  - Adam optimizer
  - Learning rates (optimization): {1e-3, 3e-4, 1e-4, 3e-5, 1e-5, 1e-6}
  - Dropout rate (regularization): {0.2, 0.3, 0.4}
  - Early stopping based on F1 scores up to 200 epochs
  - Class weights were empirically determined to deal with class imbalance.
- Performance Metric
  - F1 score (depression) (chance=0.264 for SH2-FS and 0.286 for DAIC-WOZ), Accuracy, Unweighted Average Recall (UAR).

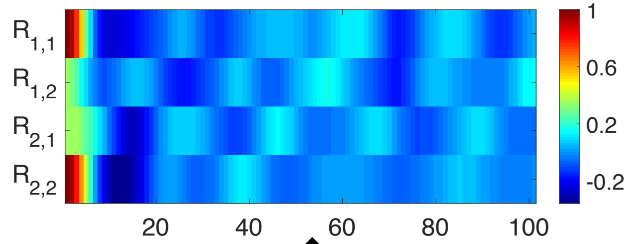
Experimental Results

- How well the proposed FVTC-CNN framework perform?
  - Grid search was done 3 times, and the best F1 scores were selected.
  - Strong results well above chance-level.
  - MFCCs, followed by Formants produced the best results.

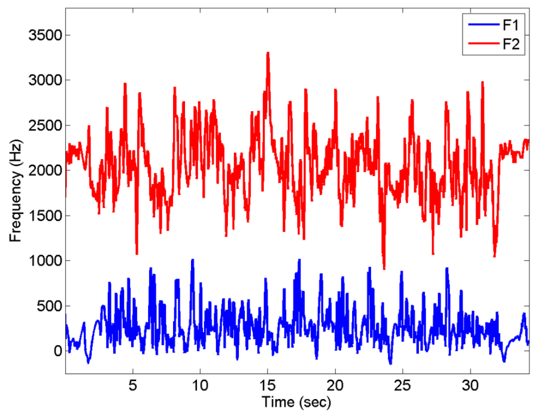
		SH2-FS		DAIC-WOZ	
		F1 (D)	Accuracy	F1 (D)	Accuracy
Chance-level		0.264	---	0.286	---
Proposed FTVC-CNN	Formants	0.373	71.1%	0.615	85.7%
	SCF	0.386	60.2%	0.500	82.9%
	MFCCs	**0.468**	**80.5%**	**0.700**	**82.9%**
	dMFCCs	0.366	64.8%	0.588	80.0%

Experimental Results

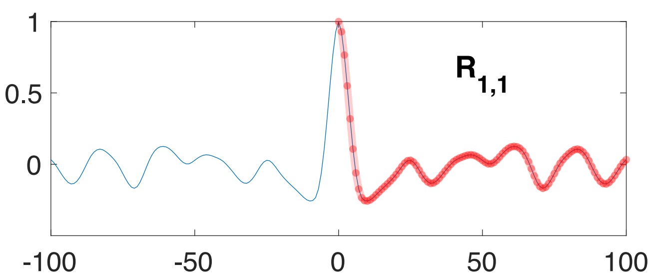
- Two important parameters in FVTC-CNN?
  - How many correlation points are needed? → *D*
  - How long the speech file needs to be? → *T*



Full VTC matrix



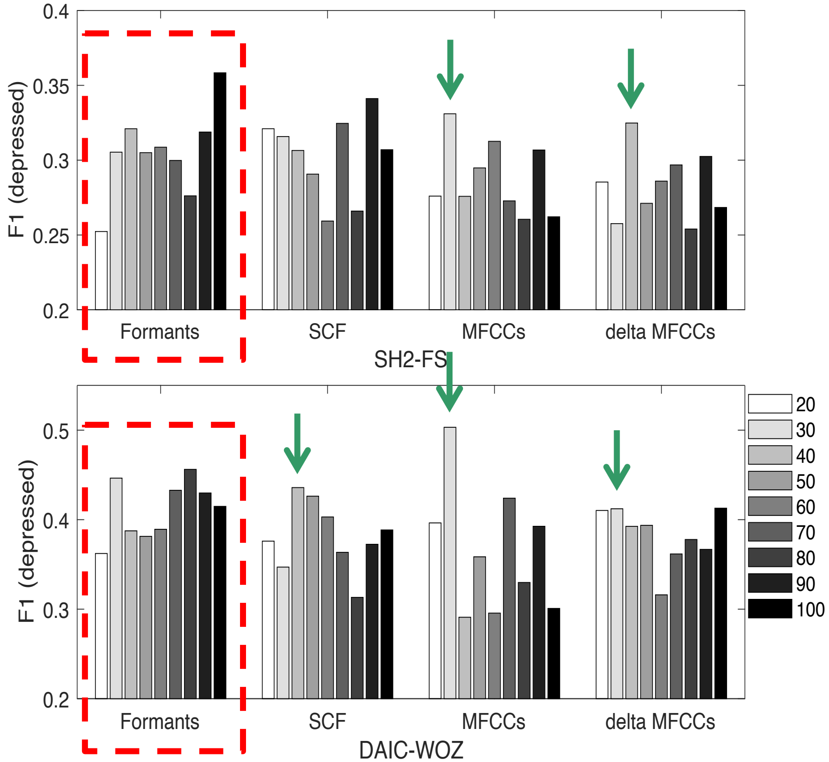
T=35 secs



D=100

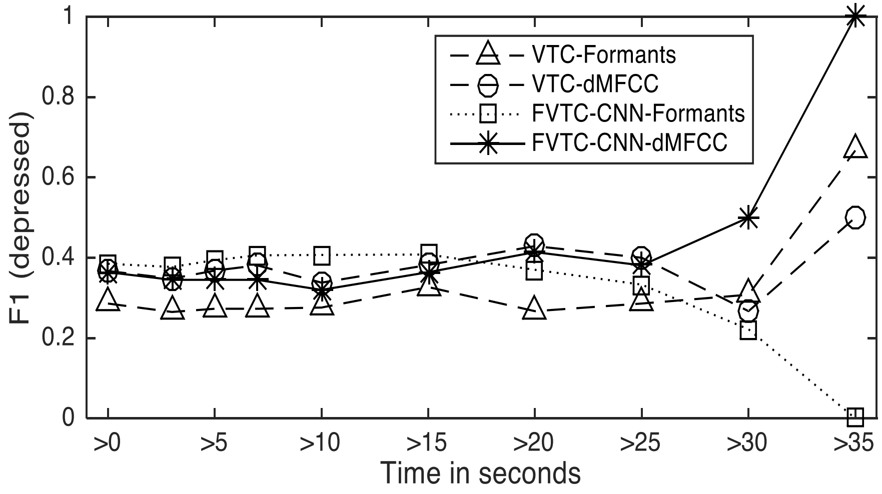
Experimental Results

- How many correlation points are needed? → *D*
  - Learning rate and dropout rate were fixed. Performances averaged across 6 runs.
  - $D \in \{20, 30, 40, 50, 60, 70, 80, 90, 100\}$
- For formants, more correlations are better.
- For others, D=30 or 40 is sufficient.



Experimental Results

- How long the speech file needs to be? → T
  - Learning rate and dropout rate were fixed. Performances averaged across 6 runs.
  - $D = 100, T \in \{> 0, > 5, > 10, \ldots, > 35\}$ seconds
- It is beneficial to have longer speech recordings.



Experimental Results

- Comparison with existing results/approaches?
  - Grid search: each system configuration was repeated 30 times for averaged results for statistical stability.
  - For DAIC, FVTC-CNN outperformed VTC, 0.64 vs. 0.55 in Mean F1 scores.
  - For SH2-FS, FVTC-CNN performed on par with or better than VTC
  - Challenge observed and to be addressed : inconsistency for different initializations in FTVC-CNN.

		SH2-FS			DAIC-WOZ		
		Mean F1	Accuracy	UAR	Mean F1	Accuracy	UAR
Chance-level		0.44	--	0.5	0.45	--	0.5
Acoustic Baselines	eGeMAPS [28]	**0.56**	**67.2%**	**0.579**	0.56	71.4%	0.554
	EMO IS10 [23] / COVAREP [32]	0.54 [23]	62.5%	0.585	0.50 [32]	51.4%	0.643
VTC	FMT	0.49	57.0%	0.534	0.55 [14]	--	--
	SCF	0.45	59.0%	0.459	--	--	--
	MFCC	0.48	65.0%	0.475	--	--	--
	dMFCC	0.49	52.0%	0.620	0.45 [14]	--	--
Proposed FVC-CNNs	FMT	0.49	59.2%	0.571	**0.64**	**73.5%**	**0.656**
	SCF	0.46	55.0%	0.565	0.60	69.6%	0.646
	MFCC	0.46	55.6%	0.547	0.62	74.8%	0.633
	dMFCC	0.49	57.9%	0.565	0.57	75.2%	0.595

Conclusions

- An effective deep learning solution (i.e. **FVTC-CNN**) to exploit vocal tract coordination for depression classification in both clean and naturalistic environments.

- The proposed **FVTC-CNN** framework brings the existing promise of the VTC concept into a deep learning paradigm, where
  - VTC's limitations were effectively addressed
    - Repeated sampling, discontinuities, and decomposition/learning method.
  - Configurations can be easily tuned
  - Can benefit from big data (i.e. scalability)
  - Latest deep learning approaches can be applied
  - Explicit discriminative learning for depression detection
  - etc…

- Future Work:
  - FVTC-CNN v2: further refine of the proposed framework.
  - Domain adaptation to bridge the gap for cross-corpus generalizability.

THANK YOU

zhaocheng.huang@unsw.edu.au

Acknowledgements:

- The authors would like to thank Thomas Quatieri and James Williamson for their insights and helpful discussions on the VTC features.
- This work was supported by Australian Research Council Linkage Project LP160101360.
- Julien Epps is also partly supported by Data61, CSIRO, Australia.