The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Speech Recognition with no Speech or with Noisy Speech

Gautam Krishna Co Tran Jianguo Yu Ahmed Tewfik

The University of Texas at Austin
University of Aizu
ICASSP 2019

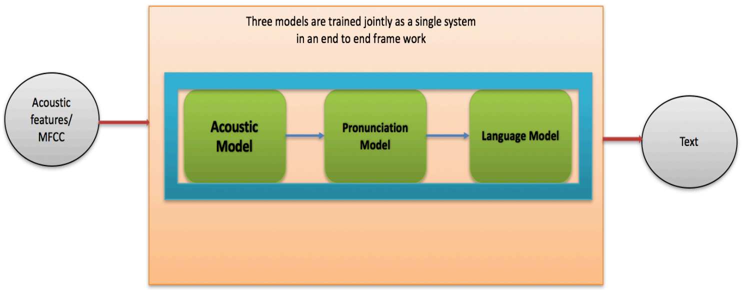
THE UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING

WNCG

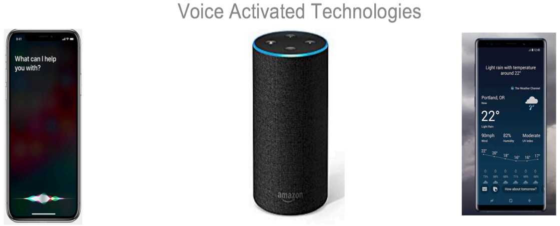
Outline:

- Introduction
- Model
- Results

Challenges for Robust ASR



Voice Activated Technologies

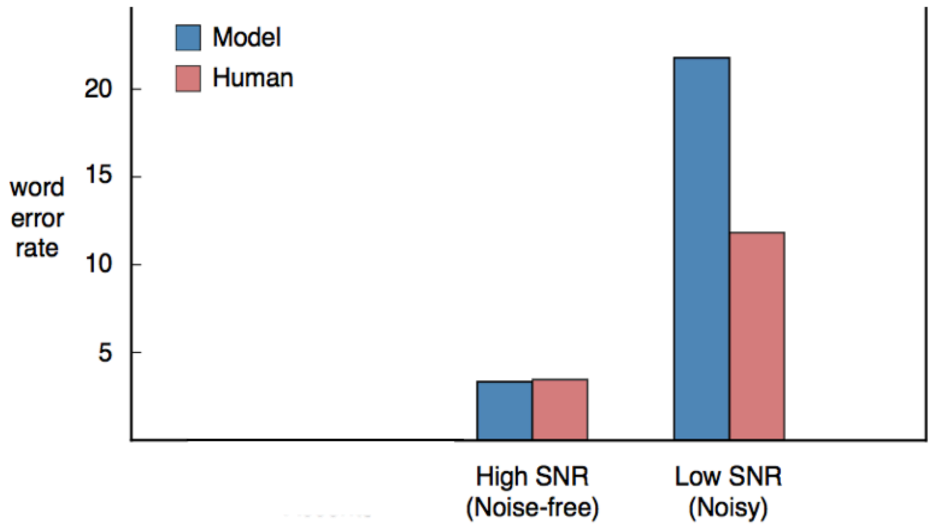


IPhone Siri
Source: Apple

Amazon Alexa
Source:Amazon

Samsung Bixby
Source:Samsung

Background Nosie, People with speaking difficulties



Source: Awni Hannun, Stanford, Speech Recognition is not solved Blog, Baidu AI Lab

Can EEG be used to improve Speech Recognition Performance?

Insight into using EEG

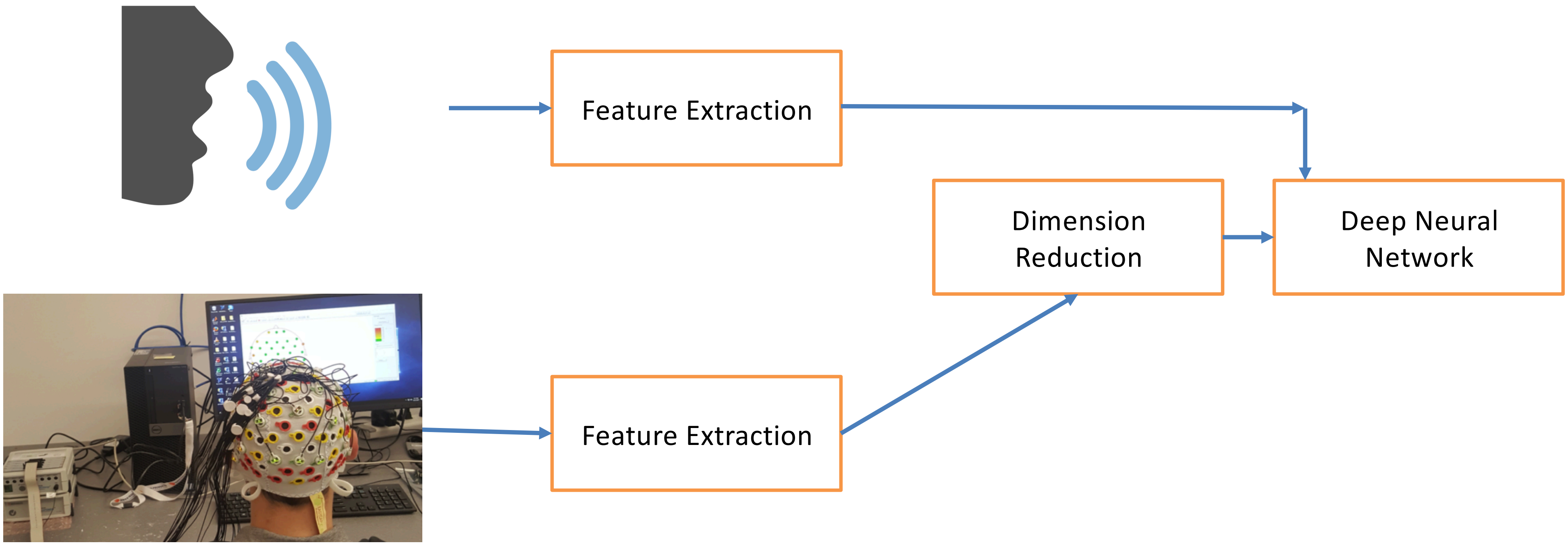
Source: medicalxpress.com

Unique property of human auditory cortex to reject environmental artifacts

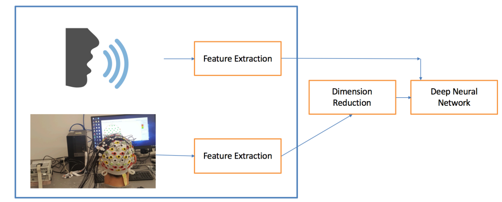
→ EEG + Acoustic features to improve robustness of ASR

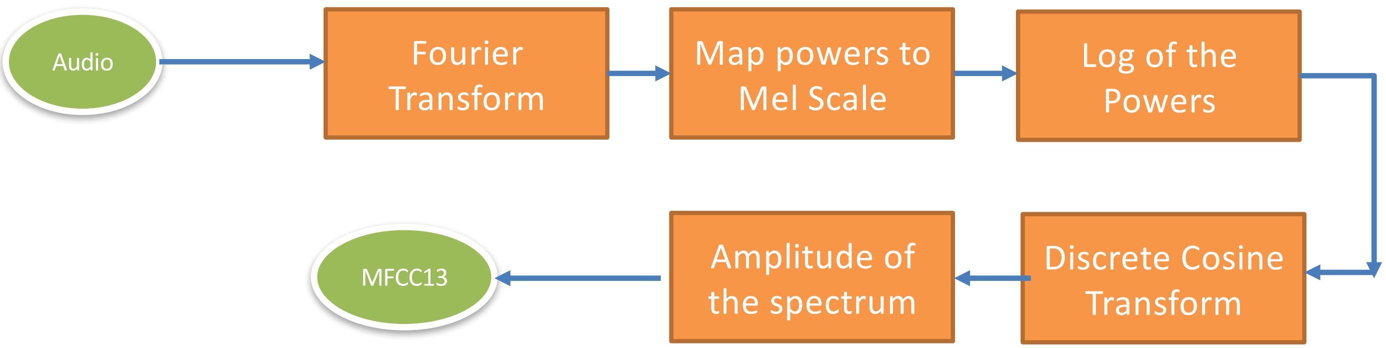
→ Improving Technology Accessibility for people with speaking disability, difficulties

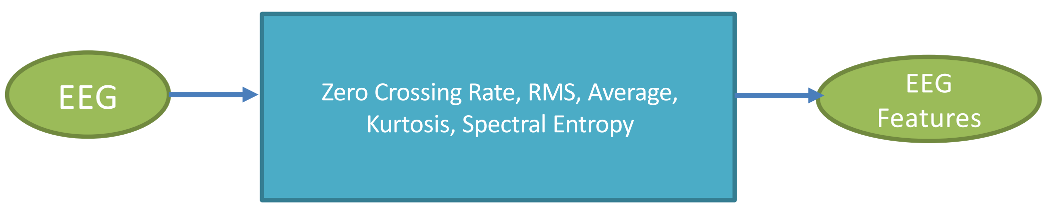
OVERVIEW


Feature Extraction
Feature Extraction
Dimension Reduction
Deep Neural Network

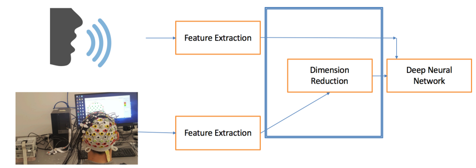
Feature Extraction



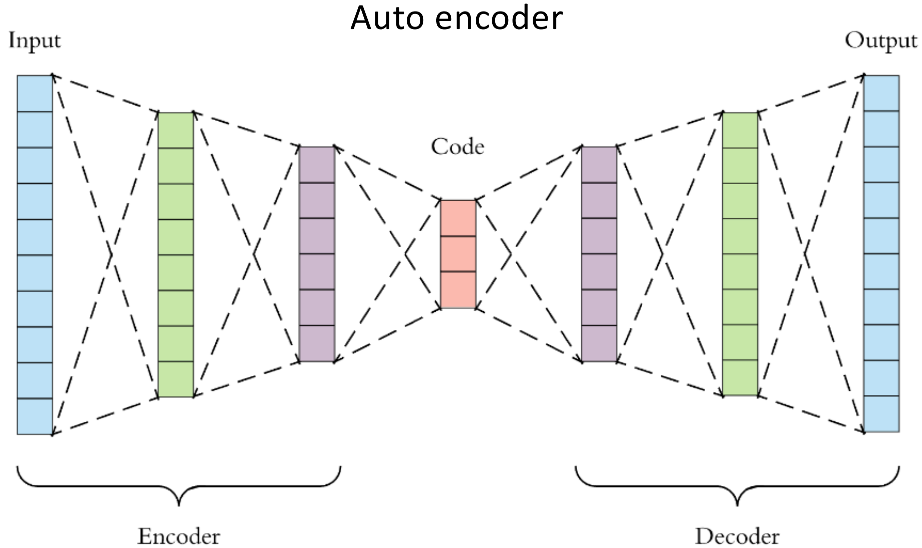




Dimension Reduction



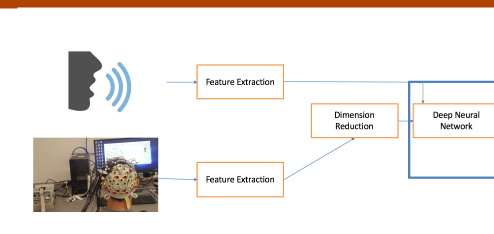
Auto encoder



Layer	Input	1	2	3	4	5	6	7	8	9	Output
Units	155	200	100	40	13	6	13	40	100	200	155

Kernel PCA

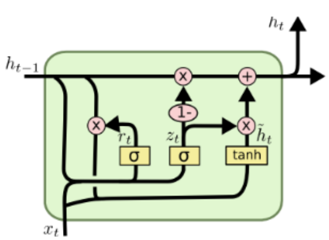
$$k(x_m, x_n) = (x_m . x_n)^d$$



ASR Model

ReLu(x)= max(0,x)

GRU Cell

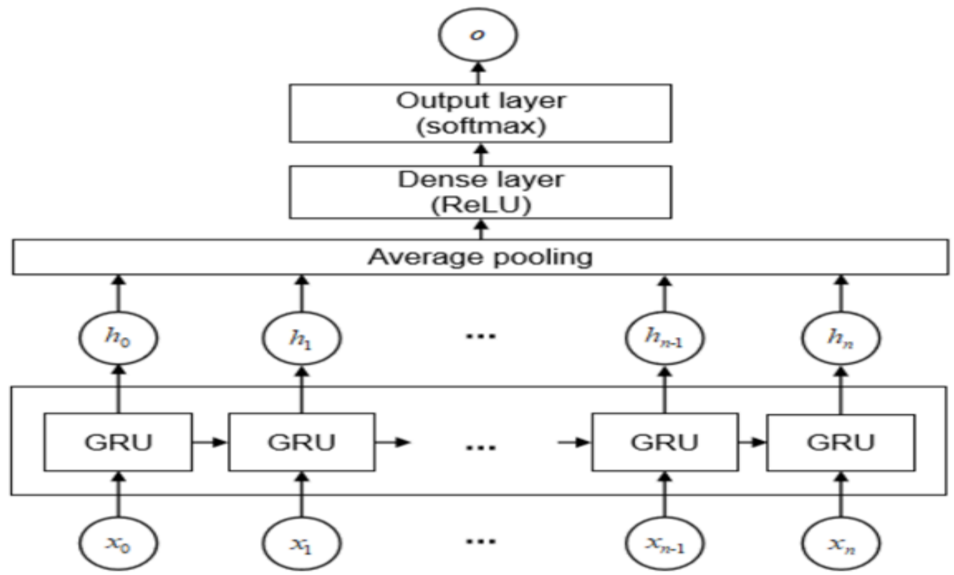


$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

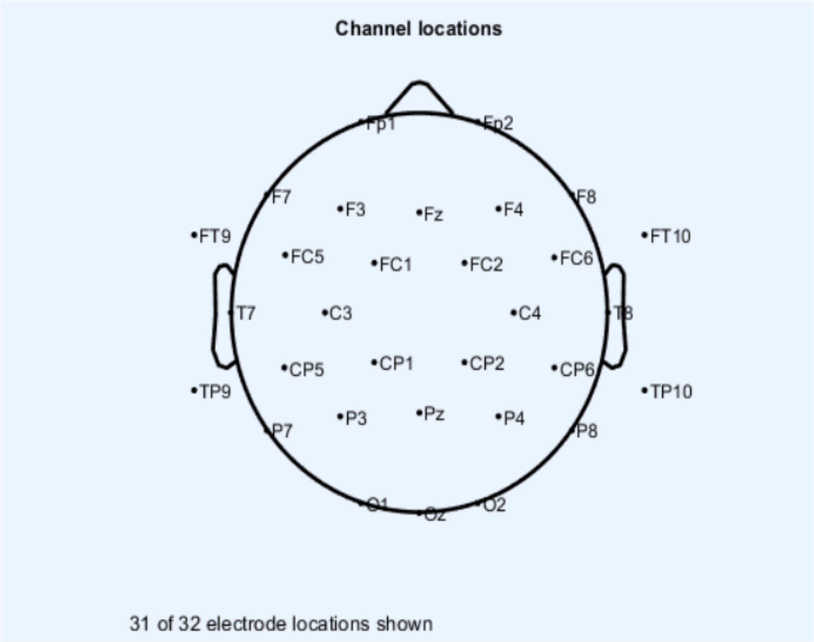


ASR Model

Predicting isolated five English vowels and four English words using only EEG features, EEG + Acoustic features

- Four male subjects
- Three were native English speakers and one non native speaker
- English vowels and four English words
- Background noise of 60 dB
- Data collected from the same subject on different days
- Brain vision EEG hardware
- Simultaneous Speech and EEG signals were recorded

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering



EEG Sensor locations

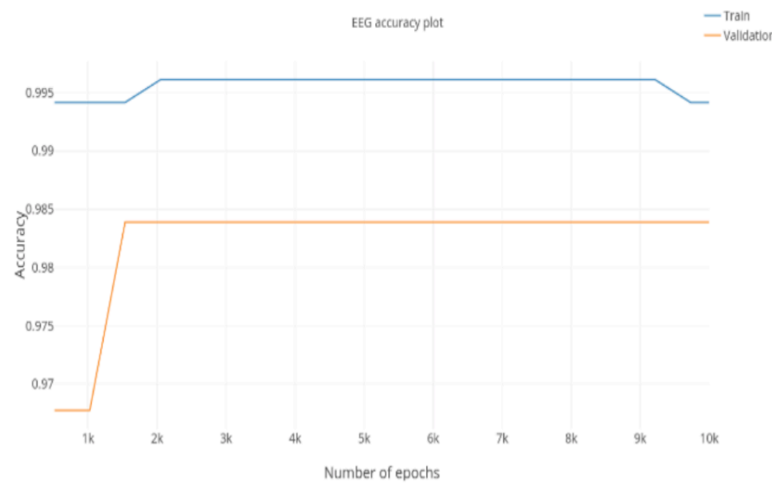
Words/Vowels	Class	Training set	Validation set	Test set	Total
	Ratio	64	16	20	100
Words	yes	195	49	61	305
Words	no	259	66	81	406
Words	right	219	56	68	343
Words	left	214	54	67	335
Vowel	a	170	44	53	267
Vowel	e	170	44	53	267
Vowel	i	170	44	53	267
Vowel	o	170	44	53	267
Vowel	u	170	44	53	267

Data Set used

Results

Words/Vowels	Background noise	MFCC acc	MFCC-EEG acc	EEG acc
Vowels	No	89.09	**96.36**	90.91
Vowels	Yes	74.74	**94.74**	92.63
Words	No	95.63	**97.91**	96.87
Words	Yes	93.00	97.50	**99.38**

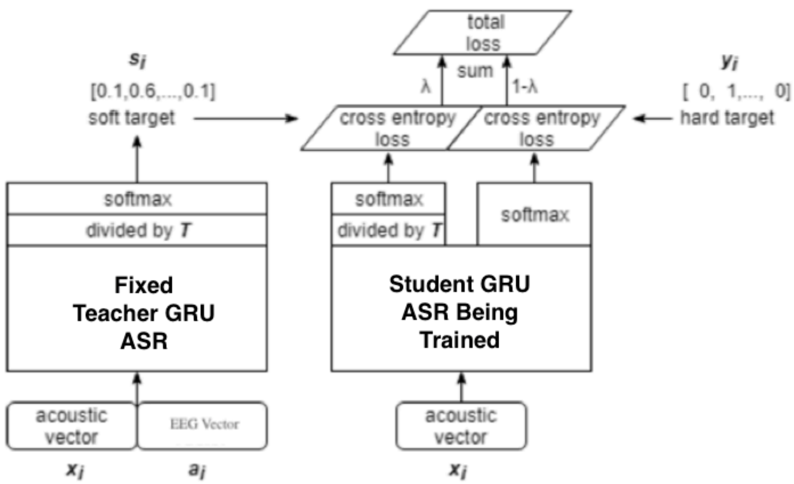
ASR EEG fusion Test time results for words, vowels data set



ASR performance for recognition of words in presence of background noise using only EEG

Results

Source: Adapted from Generalized Distillation paper by Jianguo Yu

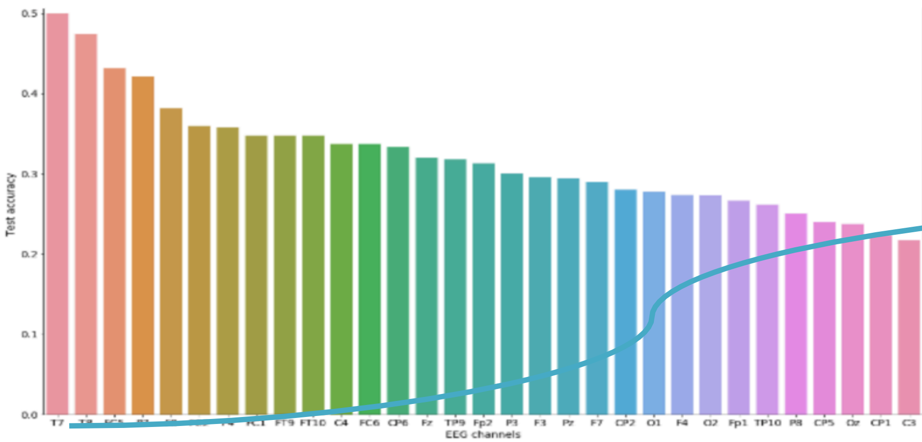


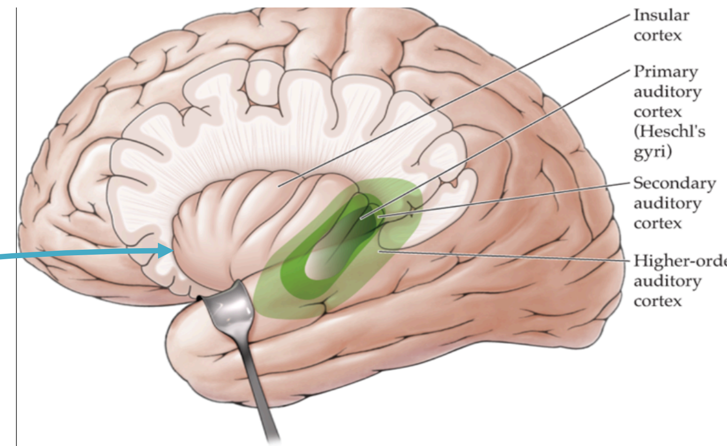
Teacher : mfcc + EEG
Student : soft targets + mfcc
Explains interpretability of the model and shows another way of integrating EEG features with acoustic features for ASR

Words/Vowels	Background noise	Student acc	MFCC acc
Vowels	No	**92.73**	89.09
Vowels	Yes	**76.84**	74.74
Words	No	**98.61**	95.83
Words	Yes	**97.62**	93.00

Test time results after distillation training

Results





Source: John H. Martin:
Neuroanatomy Text and Atlas, Fourth Edition,
http://neurology.mhmedical.com

ASR Test accuracy contribution per each EEG sensor. Sensors T7 and T8 showed highest contribution.
T7 and T8 are located near temporal lobe (auditory cortex)

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Extending the results for continuous speech for English Corpus

Number of Sentences	Number of unique words contained	EEG (CER %)	EEG+ MFCC (CER %)
3	19	2.2	**0**
5	29	1	**0**
7	42	1.8	**0**
10	59	11.6	**9.6**

Character error rate on Test set using CTC model for 65 dB noise data

Number of Sentences	Number of unique words contained	EEG (CER %)
3	19	0.8
5	29	11.6
7	42	18
10	59	22.01

Character error rate on Test set using CTC model for 65 dB noise data by using EEG features from only T7 and T8 electrodes

Submitted to EUSIPCO 2019

CONCLUSION

- EEG can help ASR systems to overcome performance loss due to background noise
- Demonstrated the feasibility of using only EEG signals for Speech Recognition

FUTURE WORK

- Collect clinical data from people with speaking disabilities, disorders
- Develop physics models to give better interpretability
- Build a larger Speech EEG data base and demonstrate results for a much larger English corpus

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Thanks!