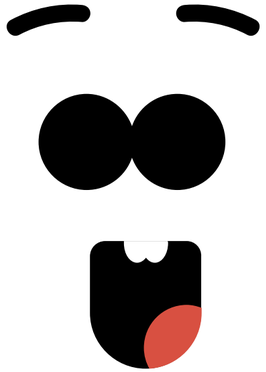
AUTOMATIC ASSESSMENT OF SPEAKING SKILLS USING AURAL AND TEXTUAL INFORMATION

Presentation By
Likhith Asapu

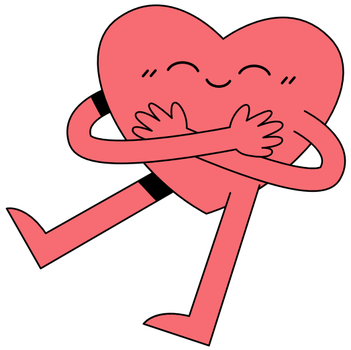
Conference
ACL - International Conference on Natural Language and Speech Processing (ICNLSP 2021)

AIM

- Aim is to a provide a multimodal speech analytics framework for automatically assessing the quality of a public speaker’s capabilities.
- Here, judgement based on two parameters:



Expressiveness



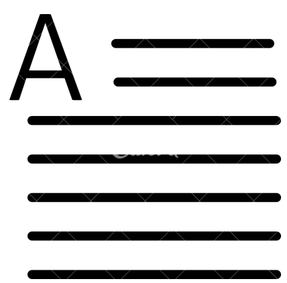
Enjoyment

PARAMETERS

Expressiveness: How active, emotional or passionate the speech is, regardless of its content.

Enjoyment: How exciting, entertaining or motivating the content of the speech was.

MODEL



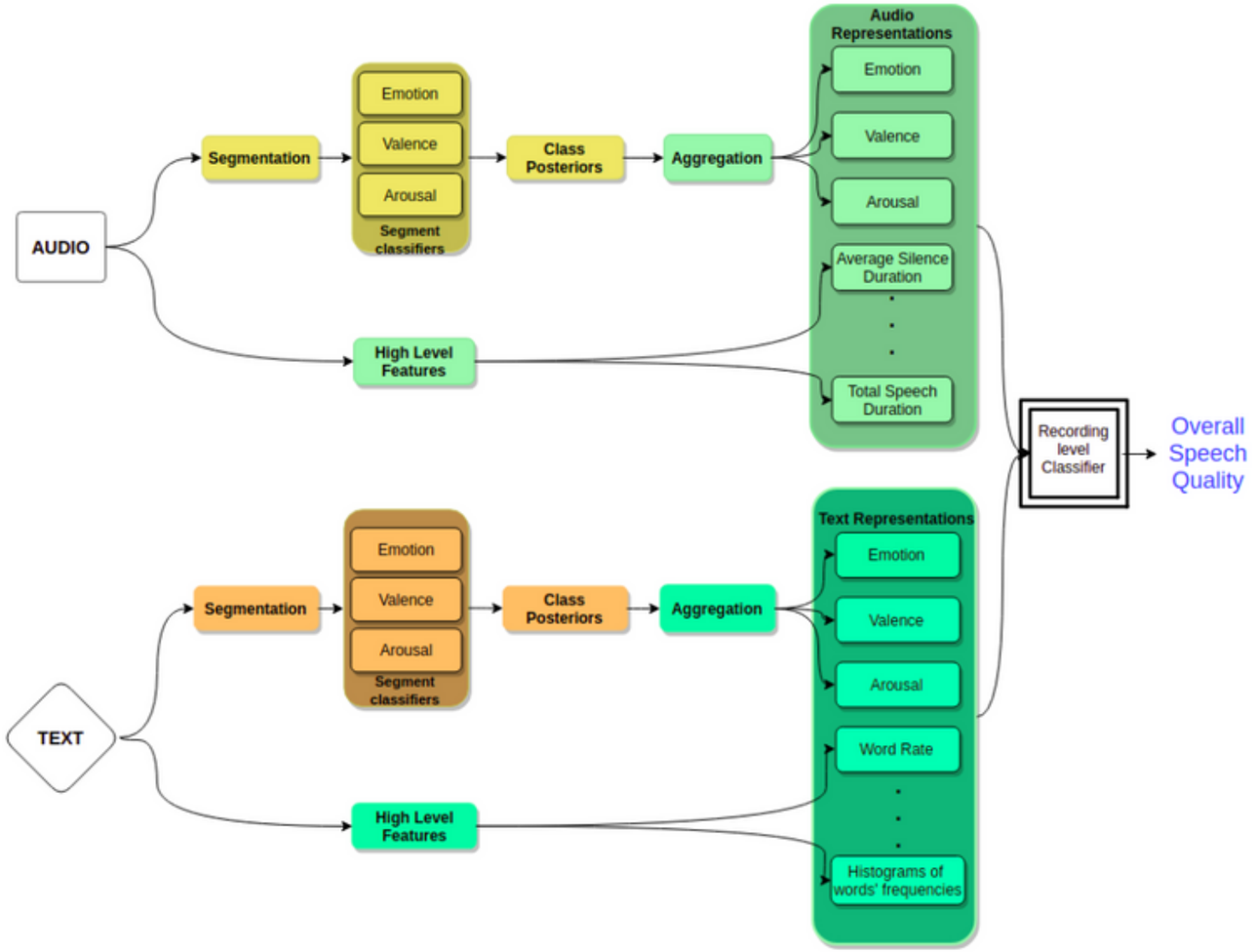
Text Based Component

Audio Based Component

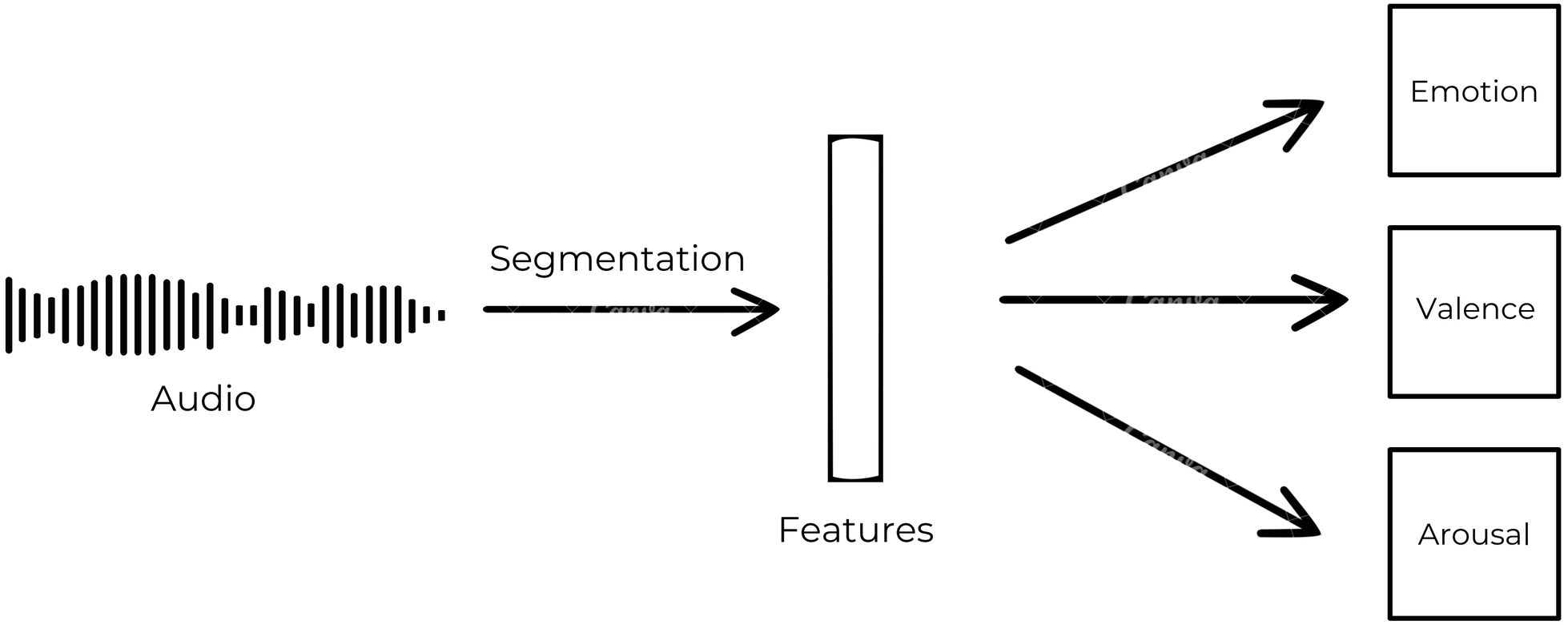
Recording Level Classifier

Overall Speech Quality

SYSTEM ARCHITECTURE



SEGMENT LEVEL AUDIO FEATURE EXTRACTION



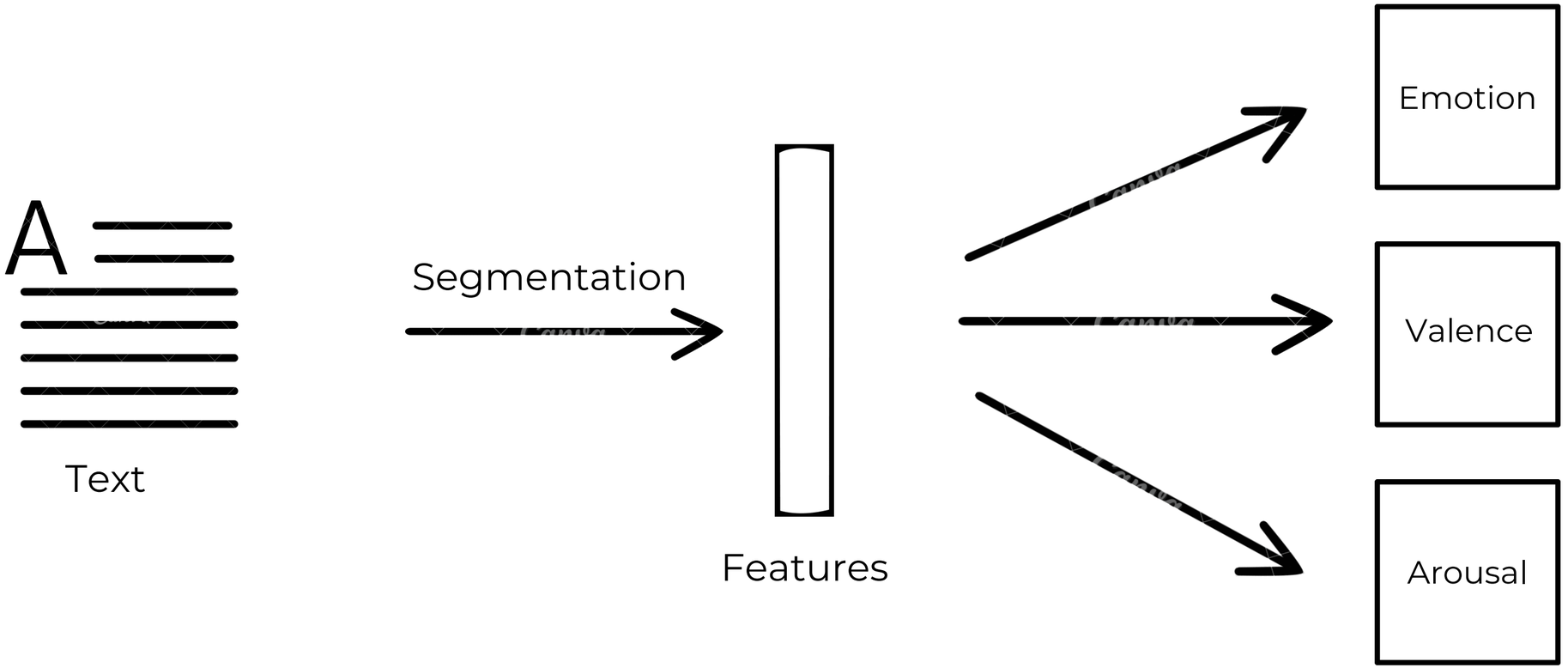
SEGMENT LEVEL AUDIO ANALYSIS - RESULTS

	emoDB	emovo	ravedess	savee	IEMOCAP	ALL
Valence(ANN)	**0.743055555**	0.585964912	**0.555130434**	0.561777505	0.533148756	**0.567069051**
Valence(SVM)	0.594883009	**0.646722204**	0.473057552	**0.657404185**	**0.593961419**	0.540206419

	emoDB	emovo	ravedess	savee	IEMOCAP	ALL
Emotion(ANN)	0.740731008	0.587628865	0.446410142	0.522395292	0.530428230	0.500564971
Emotion(SVM)	**0.812064952**	**0.720613669**	**0.603552805**	**0.684097864**	**0.548204612**	**0.573436570**

	emoDB	emovo	ravedess	savee	IEMOCAP	ALL
Arousal(ANN)	0.783333333	0.618713450	**0.676869565**	0.657269141	0.559306152	0.635989889
Arousal(SVM)	**0.812732930**	**0.712002746**	0.552770676	**0.694198960**	**0.576053771**	**0.645453187**

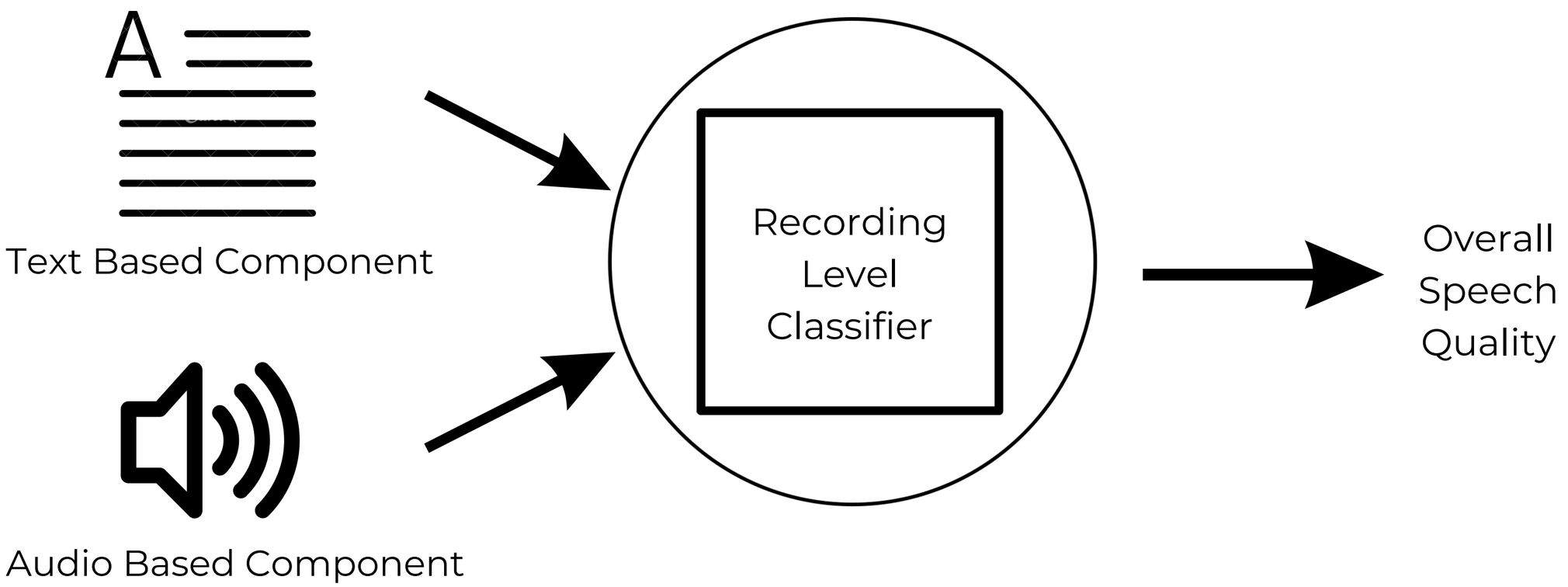
SEGMENT LEVEL TEXT FEATURE EXTRACTION



TEXT ANALYSIS - RESULTS

	SVM + fastText	ANN + fastText	BiLSTM + fastText
Emotion	**0.596961433774467**	0.537647756350246	0.558251263808275
Arousal	0.457903210363394	0.524056042893696	**0.536418273731511**
Valence	**0.599081468424333**	0.556230682650563	0.588661507416422

RECORDING LEVEL CLASSIFIER



RECORDING LEVEL FEATURES

- Speech Features - average silence duration, silence segment per minute, standard deviation of silence duration, speech ratio and word rate in speech.

- Text Features - word rate, unique word rate and 10-bin histogram of word frequencies

INDIVIDUAL MODALITIES - RESULTS

	Meta Audio(SVM)	Meta Audio(ANN)	Text(SVM)	Text(ANN)	Low Level Audio(SVM)	Low Level Audio(ANN)
Female Expressiveness	0.563423050	**0.682073844**	**0.716765480**	0.657938076	**0.851639061**	0.737974465
Male Expressiveness	0.726851851	**0.781944444**	**0.701388888**	0.659722222	**0.932605820**	0.844973544
Female Enjoyment	0.827638888	**0.843611111**	0.631940427	**0.678000289**	**0.959583333**	0.768263888
Male Enjoyment	**0.739947089**	0.446759259	**0.671990740**	0.635879629	**0.821494708**	0.734722222

FUSION METHODS - RESULTS

	Meta Audio + Text(SVM)	Meta Audio + Text(ANN)	MA and LLA LateFusion(SVM)	MA and LLA LateFusion(ANN)	MA and LLA EarlyFusion(SVM)	MA and LLA EarlyFusion(ANN)
Female Expressiveness	0.654382332	**0.820565907**	0.900103519	**0.975069013**	0.604606625	**0.721152518**
Male Expressiveness	0.855555555	**0.8375**	0.885185185	**0.993055555**	0.739814814	**0.796296296**
Female Enjoyment	0.738333333	**0.776111111**	0.916805555	**0.97**	0.807916666	**0.84375**
Male Enjoyment	**0.735714285**	0.456018518	**0.869444444**	0.444444444	**0.738293650**	0.456018518

FUSION METHODS - RESULTS

	MA + T and LLA Late Fusion(SVM)	MA + T and LLA Late Fusion(ANN)	MA + T and LLA Early Fusion(SVM)	MA + T and LLA Early Fusion(ANN)
Female Expressiveness	0.950793650	**0.984126984**	0.627553485	**0.786973775**
Male Expressiveness	0.928240740	**0.976851851**	**0.841666666**	0.769907407
Female Enjoyment	0.909210526	**0.938345864**	0.708611111	**0.772777777**
Male Enjoyment	**0.908134920**	0.444444444	**0.744973544**	0.449074074

LINGUISTIC ANALYSIS

- The low level features such as Zero-crossing rate, Energy, Energy entropy, MFCC, etc. capture acoustic variances such as **Tone, Frequency and Energy of Speech - ACOUSTIC ANALYSIS**
- Fast Text Embeddings used for classification capture **syntax and semantics** of the text. High level features such as word rate, unique word rate and 10-bin histogram of word frequencies also help in capturing **semantics- TEXTUAL ANALYSIS**
- High level audio features such as average silence duration, silence segment per minute, standard deviation of silence duration, speech ratio and word rate in speech included capture some aspects of **Rhythm of speech and prosody.**

CHALLENGES FACES

- Had to work with large amounts of data 6 datasets in total. Data collection was a huge challenge.
- Large amount of testing needed to be done.
- Had to be acquainted with working of libraries such as PyAudioAnalysis

TIMELINE

DONE - Dataset Collection and Segment Level Audio Analysis

TO BE COMPLETED - Segment Level Text Analysis, Recording Level Analysis - By 29th November

Experimentation - Neural Network Architectures such as ANN, LSTMs instead of SVM - End Submission - 2nd December

THANK YOU