



# DYNAMIC SPEECH RATE ADJUSTMENT

Manavpreet Singh Cheema, Mudit Surana, Vaisakh M Menon

“Words have incredible power. They can make people’s hearts soar,  
or they can make people’s hearts sore.”

-Dr. Mardy Grothe



# WHAT ARE WE DOING?

We are creating a **machine learning model** that analyzes **speech patterns** to dynamically adjust speech rates and providing users with **feedback**.






23.47  
MILLION

People face dysarthria.

1 billion

People require speech therapy.



# FEATURES OF THE MODEL

Detailed analysis of **speech rate** and **emotion**, including **visualization**.

**Real-time feedback** to optimize speech rate.

**Speech Rate Classification** with Statistical Analysis.



# LITERARY REVIEW-1

## Feature Extraction :

- Utterance Duration and Word Count
- Syllable Count

## Calculation of Speed Rate

- Using both WPM and SPS metrics

## Inferences

- Developed a method to measure the pace of speech in short utterances, addressing limitations of the WPM metric.
- Utilized the ILMT-s2s and HCRC corpora to establish a comparison between human speech and TTS output rates.

## Accuracy

They provided an adjusted R-squared value of 0.883 for their regression model, indicating a strong fit between the TTS output duration and subject utterance duration.

(School of Computer Science and Statistics, Trinity College Dublin, Ireland;  
Usher Institute of Population Health Sciences & Informatics, University of Edinburgh, UK)

Word Count	<i>Mdn</i>	<i>M</i>	<i>SD</i>	Count
ILMT-s2s All Subjects	4	5.168	6.01	3,628
ILMT-s2s English Subjects	4	4.919	6.76	1,980
ILMT-s2s Portuguese Subjects	4	5.466	4.97	1,648
Duration (sec.)	<i>Mdn</i>	<i>M</i>	<i>SD</i>	Count
ILMT-s2s All Subjects	1.493	2.244	2.75	3,628
ILMT-s2s English Subjects	1.285	1.939	2.96	1,980
ILMT-s2s Portuguese Subjects	1.874	2.610	2.49	1,648

Table 1: Summary of word count and duration in corpus

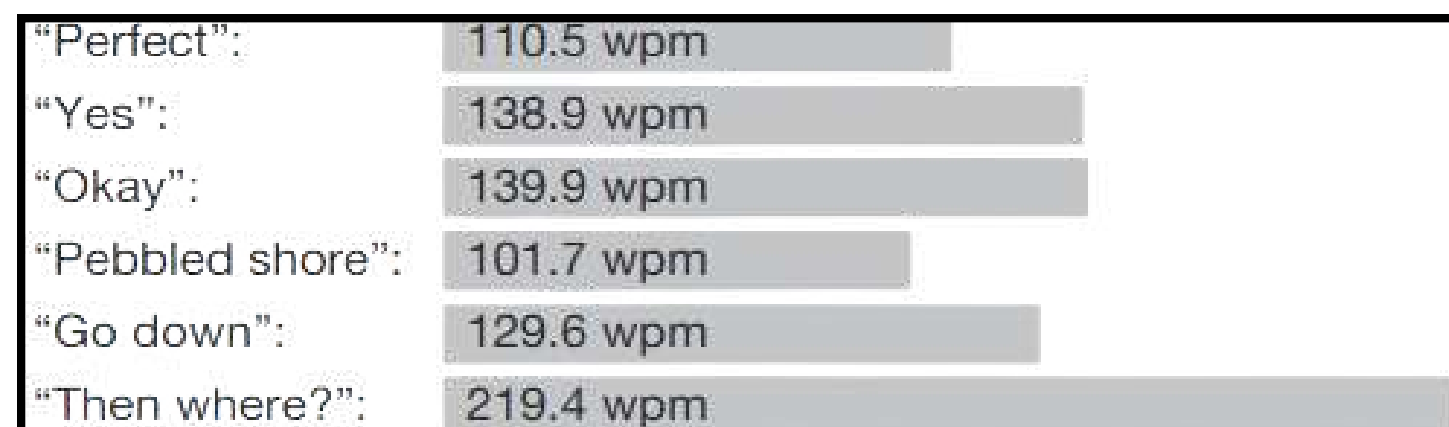


Figure 1: *Mdn* wpm for the six samples — Subject

## Speech Rate Calculations with Short Utterances

Hayakawa, A., Vogel, C., Luz, S., & Campbell, N. (2018). Speech Rate Calculations with Short Utterances: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA).<https://aclanthology.org/L18-1502>

# LITERARY REVIEW-2

## Feature Extraction

- Using Mel-Frequency Cepstral Coefficients **MFCCs**
- Perceptual Linear Prediction (**PLP**)

## Methodology Used

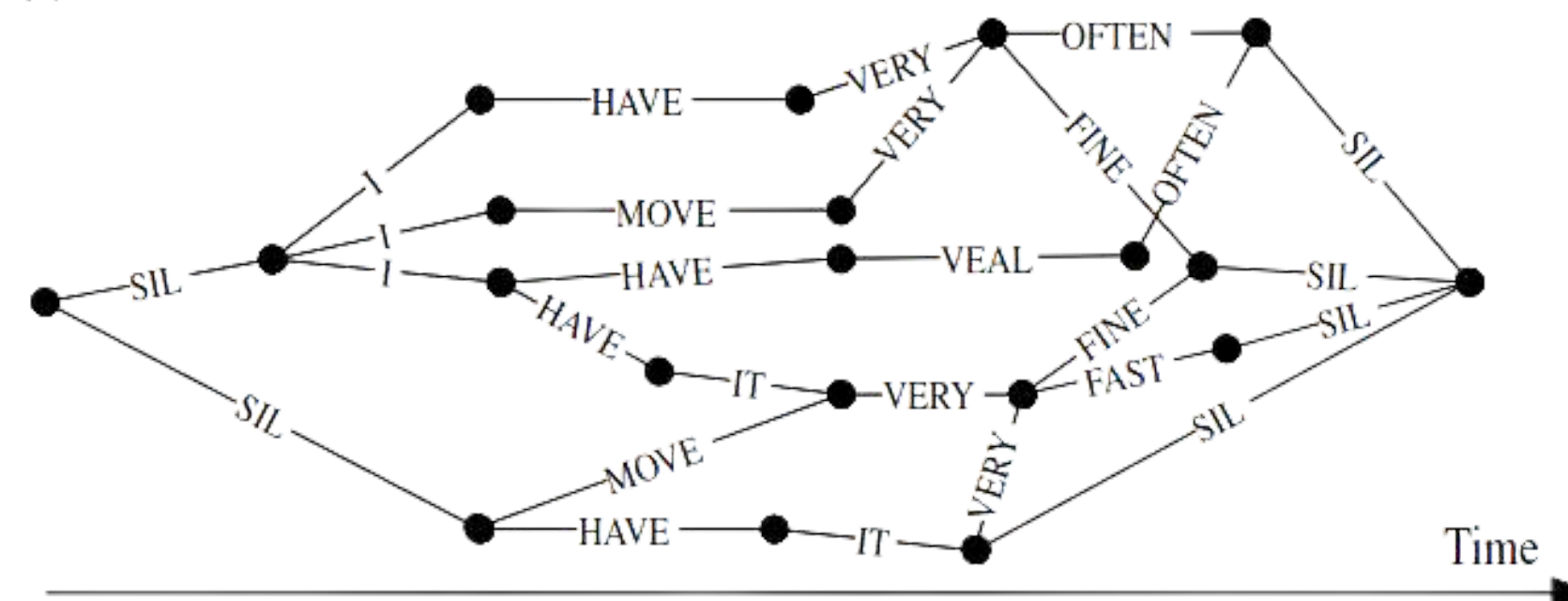
- Machine Learning Model: Uses **HMMs** with **ML** and **MMI** training for **speech recognition**, improved with **VTLN** and **MLLR**.
- Results and Accuracy: No specific metrics – **focused discriminative training**.

## Relevance to Dynamic Speech Rate

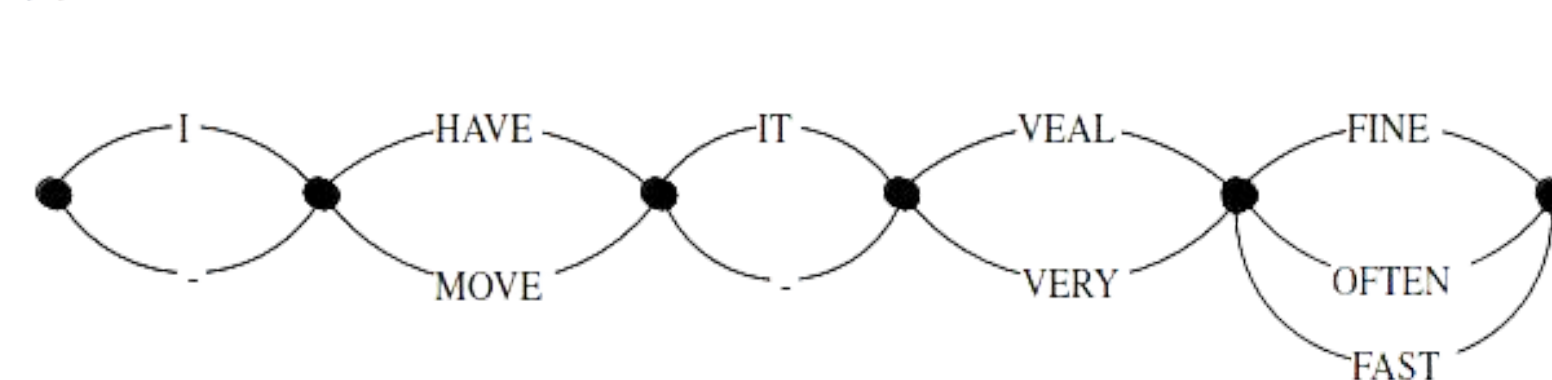
### Adjustment:

- Geared towards **speech recognition**, not rate adjustment;
- Feature extraction methods (**MFCCs**, **PLP**) are **not designed** for dynamic rate modification.

(a) Word Lattice



(b) Confusion Network



## The application of hidden Markov models in speech recognition



# LITERARY REVIEW-3

## Feature Extraction:

- Uses prosodic (pitch, energy, duration) and spectral (MFCCs, LPCC, formants) features for emotion recognition.
- Involves segmenting signals into frames and mapping to feature vectors.

## Machine Learning Application:

- Employs classifiers like linear Bayes, k-NN, SVM, GMM, and neural networks for emotion detection.
- Utilizes HMM and RNN for sequence-based feature classification.

## Relevance to Our Project:

Similar feature extraction methods (especially MFCCs) applicable for speech rate analysis.

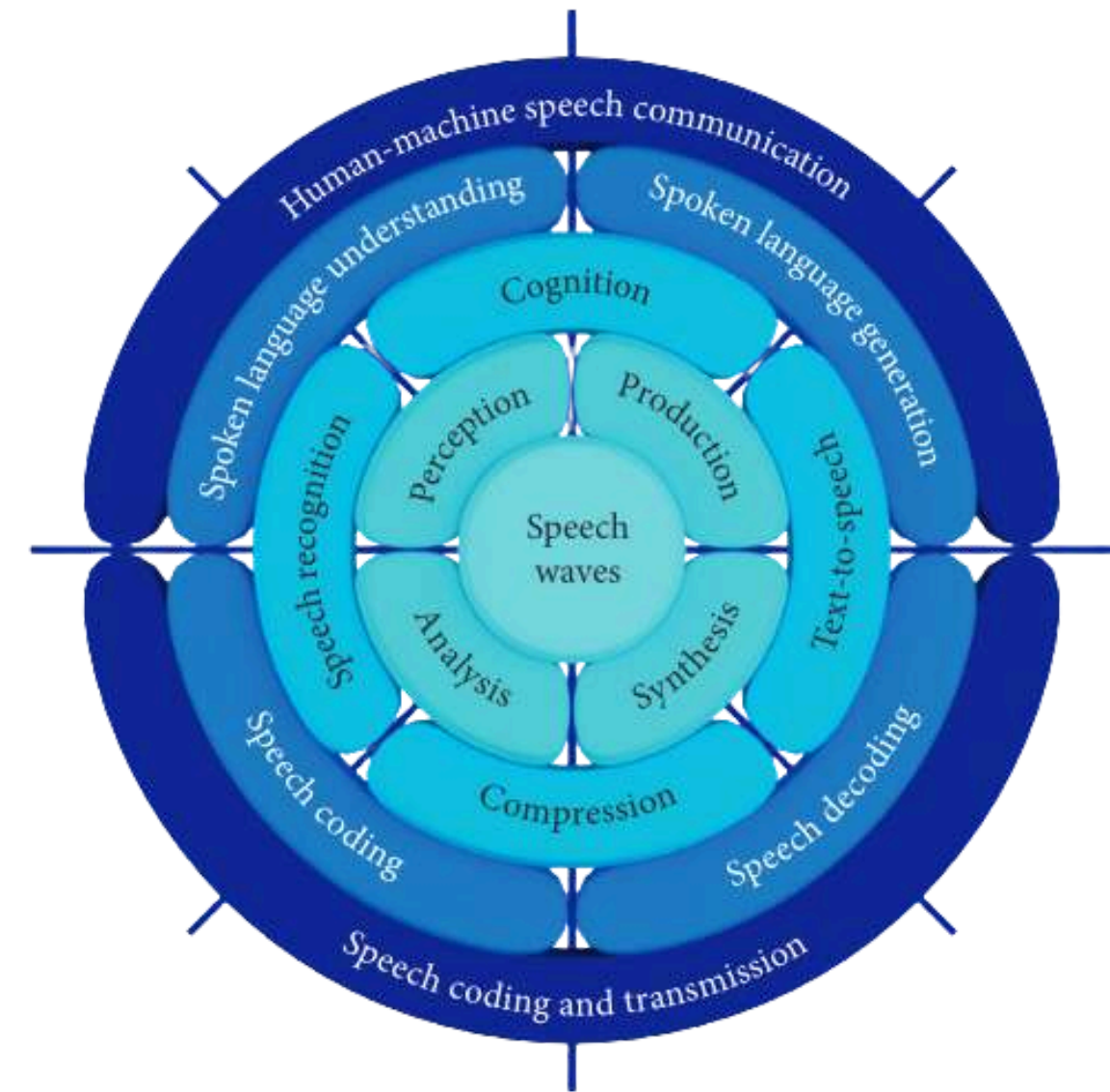


Figure 2: Unified framework that encompasses speech signal processing fields in the scope of the article.

## Speech Technology Progress Based on New Machine Learning Paradigm

# LITERARY REVIEW-4

## Feature Extraction:

- Mel-frequency cepstral coefficients (MFCCs) for capturing speech characteristics.
- Filter-bank energies (Fbanks) that represent the short-term power spectrum of the speech.
- Rate of Speech (ROS)

## Machine Learning Model Used :

- Deep Neural Networks were deployed
- Multilayered architecture for complex pattern learning

## Application in Dynamic Speech

### Rate Adjustment:

- Recognizes and adapts to the rate of speech (ROS)
- Compensates for distortions in fast or slow speech
- Improves accuracy in speech recognition systems
- Can be combined with HMM transition adjustments for enhanced performance

(CSLT, RIT, Tsinghua University;  
TNList, Tsinghua University;  
Beijing University of Posts and Telecommunications;  
Chongqing University of Posts and Telecommunications)

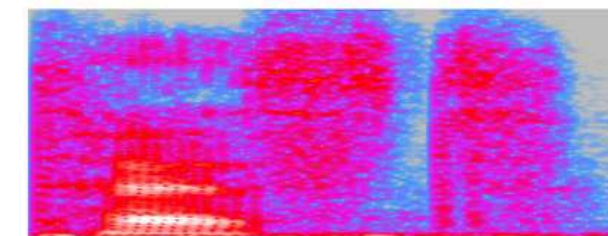


Figure 1: The spectrogram of a fast reading for word 'test'.

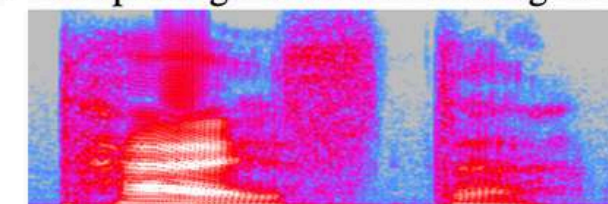


Figure 2: The spectrogram of a slow reading for word 'test'.

Table 7: Results with both the DNN- and HMM-based ROS compensation.

Test set	WER/%	
	Slow	Fast
ROS	< 4	> 10
DNN Baseline	45.71	31.22
+DNN-based compensation	44.92	29.54
+HMM-based compensation	44.76	29.08

## Learning Speech Rate in Speech Recognition.

Zeng, X., Yin, S., & Wang, D. (2015). Learning Speech Rate in Speech Recognition.

# DATA SET

## Source:

The CREMA-D dataset is a collection of **audio-visual emotional** expressions.

## Data Collection:

**Actors:** 91 **actors** (48 male, 43 female) aged between 20 and 74.

**Emotions:** The dataset includes **recordings** of 12 different emotions and the neutral state. The specific emotions are Anger, Disgust, Fear, Happiness, Neutral, Sadness.

**Modalities:** Both audio and video recordings are available. For audio, the dataset features **high-quality** recordings.

**Phrases:** Actors performed a series of 12 scripted phrases, each **articulated** in multiple emotional tones to reflect different emotions.

**Variability:** Each phrase was spoken in multiple emotional tones and at different **intensity** levels to capture a range of expressions.

## Ethical Concerns in Data Collection:

**Consent:** All actors provided **informed consent** for the use of their performances in research and educational efforts. This consent includes the use of their voice and facial expressions.

**Privacy:** Steps were taken to ensure that the identity and performance data were handled according to ethical guidelines, with **personal identifiers** removed where possible.

**Compensation:** Actors were **compensated** for their participation, which is a standard practice to ethically incentivize participation without coercion.



cleaned\_tabulatedVotes

A	D	F	H	N	S	numResponses	agreement	emoVote	meanEmoResp	meanAngerResp	meanDisgustResp	meanFearResp	meanHappyResp	meanNeutralResp	meanSadResp	medianEmoResp	meanEmoRespNorm	meanAngerRespNorm	meanDisgustRespNorm	meanFearRespNorm	meanHappyRespNorm	meanNeutralRespNorm	meanSadRespNorm	medianEmoRespNorm	
0	0	0	1	10	0	11	0.909090909090909	N	71.7272727272727	-1.0	-1.0	-1.0	98.0	69.1	-1.0	87.0	66.1713199530713	-1.0	-1.0	-1.0	97.8947368421053	62.9989782641679	-1.0	83.3333333333333	
0	0	0	3	6	0	9	0.666666666666667	N	62.0	-1.0	-1.0	-1.0	42.6666666666667	71.6666666666667	-1.0	72.0	53.8161602678193	-1.0	-1.0	-1.0	36.292735042735	62.5778728803614	-1.0	53.8461538461539	
0	0	0	4	7	0	11	0.636363636363636	N	63.6363636363636	-1.0	-1.0	-1.0	56.5	67.7142857142857	-1.0	60.0	61.874712936474	-1.0	-1.0	-1.0	55.1652298850575	65.7087032515691	-1.0	61.1111111111111	
2	0	0	6	2	0	10	0.6	H	57.6	82.5	-1.0	-1.0	63.5	15.0	-1.0	69.5	52.9273089246813	84.4201850780798	-1.0	-1.0	56.5427080837364	10.5882352941176	-1.0	66.3081395348837	
1	1	0	0	7	1	10	0.7	N	63.9	21.0	96.0	-1.0	-1.0	73.7142857142857	6.0	80.0	59.4867281993627	12.9411764705882	94.9367088607595	-1.0	-1.0	69.1234852374685	3.125	75.1820728291317	
0	1	1	0	8	0	10	0.8	N	40.4	-1.0	62.0	5.0	-1.0	42.125	-1.0	45.0	35.3467894511312	-1.0	49.0566037735849	2.29885057471264	-1.0	-1.0	37.7640550203768	-1.0	36.3339070567986
1	0	0	1	7	0	9	0.777777777777778	N	51.8888888888889	16.0	-1.0	-1.0	30.0	60.1428571428571	-1.0	61.0	46.297590934203	0.0	-1.0	-1.0	27.8350515463917	55.5490381230622	-1.0	50.6493506493507	
0	1	0	2	7	0	10	0.7	N	52.7	-1.0	29.0	-1.0	40.5	59.5714285714286	-1.0	50.0	50.2121802259455	-1.0	27.0833333333333	-1.0	-1.0	29.6296296296296	59.3970299524089	-1.0	47.803776683087
5	5	0	0	1	1	12	0.416666666666667	A:D	50.5833333333333	37.2	58.6	-1.0	-1.0	57.0	71.0	54.5	41.872206565809	30.1925235258569	50.6189540502666	-1.0	-1.0	51.1363636363636	47.2727272727273	40.3288740245262	
2	6	0	0	1	0	9	0.666666666666667	D	73.4444444444444	79.5	67.0	-1.0	-1.0	100.0	-1.0	81.0	70.0720414899677	78.7064190289997	62.2059225586184	-1.0	-1.0	100.0	-1.0	77.4193548387097	
0	0	2	3	4	1	10	0.4	N	58.7	-1.0	-1.0	55.0	56.0	73.0	17.0	69.0	53.7894321064298	-1.0	-1.0	54.5194508009153	50.3339074306816	68.5913312693498	3.48837209302326	60.3132161955691	
0	3	5	0	3	0	11	0.454545454545455	F	57.0	-1.0	48.0	65.8	-1.0	51.3333333333333	-1.0	57.0	53.8666691664978	-1.0	36.9019327555913	65.9885371073936	-1.0	50.6282923425781	-1.0	52.8571428571429	
2	3	3	1	2	0	11	0.272727272727273	D:F	67.7272727272727	93.0	81.3333333333333	63.6666666666667	26.0	49.0	-1.0	76.0	64.5385339729222	92.6315789473684	81.4244186046512	52.5182457202435	27.7777777777778	47.5274725274725	-1.0	80.7692307692308	
0	2	1	0	7	1	11	0.636363636363636	N	50.0909090909091	-1.0	46.5	22.0	-1.0	59.5714285714286	19.0	49.0	41.1255497986241	-1.0	40.9206234142805	9.09090909090909	-1.0	49.5278276435342	14.7540983606557	38.8888888888889	
1	1	1	0	8	0	11	0.727272727272727	N	72.1818181818182	76.0	96.0	40.0	-1.0	72.75	-1.0	74.0	68.3419805391695	59.0909090909091	94.3661971830986	40.9836065573771	-1.0	69.6651341374349	-1.0	61.7283950617284	
6	4	0	1	0	0	11	0.545454545454545	A	53.4545454545455	59.5	48.25	-1.0	38.0	-1.0	-1.0	42.0	43.8503948357614	51.0448730690402	38.8129428614503	-1.0	20.8333333333333	-1.0	-1.0	28.3582089552239	
0	0	0	0	10	0	10	1.0	N	77.8	-1.0	-1.0	-1.0	-1.0	77.8	-1.0	81.0	77.5921250280795	-1.0	-1.0	-1.0	-1.0	-1.0	77.5921250280795	-1.0	82.4970131421744
2	1	0	1	3	0	7	0.428571428571429	N	62.5714285714286	48.0	64.0	-1.0	85.0	64.3333333333333	-1.0	64.0	48.2732682816543	31.5072463768116	51.3513513513513	-1.0	83.1460674157303	46.8003221502917	-1.0	51.3513513513513	
0	1	1	0	8	0	10	0.8	N	59.3	-1.0	35.0	28.0	-1.0	66.25	-1.0	61.5	54.5933975206774	-1.0	26.9662921348315	14.2857142857143	-1.0	63.0852460982785	-1.0	63.0318743327741	
5	4	0	0	2	0	11	0.454545454545455	A	59.4545454545455	54.8	47.75	-1.0	-1.0	94.5	-1.0	62.0	47.4240307965086	45.9740546184036	29.4961354650132	-1.0	-1.0	86.9047619047619	-1.0	49.4845360824742	
7	1	1	0	2	0	11	0.636363636363636	A	42.3636363636364	36.8571428571429	54.0	43.0	-1.0	55.5	-1.0	42.0	36.640363295745	26.6758285805905	69.4915254237288	52.112676056338	-1.0	47.3544973544973	-1.0	37.1428571428571	
3	4	0	0	2	0	9	0.444444444444444	D	48.1111111111111	44.0	45.75	-1.0	-1.0	59.0	-1.0	49.0	34.7221180260333	24.7333966846162	35.7589285714286	-1.0	-1.0	47.6315789473684	-1.0	40.0	
0	1	0	4	6	0	11	0.545454545454545	N	56.6363636363636	-1.0	60.0	-1.0	38.0	68.5	-1.0	60.0	49.4409414701444	-1.0	56.5217391304348	-1.0	-1.0	31.0245398061854	60.5384096360686	-1.0	46.5909090909091
1	0	0	4	6	0	11	0.545454545454545	N	52.9090909090909	75.0	-1.0	-1.0	47.5	52.8333333333333	-1.0	52.0	46.4979215057814	70.2380952380952	-1.0	-1.0	39.7652736770384	47.0296577695577	-1.0	44.7058823529412	
2	0	1	0	8	0	11	0.727272727272727	N	54.3636363636364	84.0	-1.0	70.0	-1.0	45.0	-1.0	60.0	47.7824762591977	76.8796992481203	-1.0	76.5625	-1.0	36.9106675443668	-1.0	48.6842105263158	
1	0	2	0	7	0	10	0.7	N	58.1	55.0	-1.0	30.0	-1.0	66.5714285714286	-1.0	61.5	49.8124006330946	2.17391304347826	-1.0	17.6923076923077	-1.0	65.795068271836	-1.0	48.37842846553	
5	4	0	0	2	0	11	0.454545454545455	A	67.7272727272727	67.2	66.25	-1.0	-1.0	72.0	-1.0	68.0	63.0088664773876	62.7593130633501	60.4543157991434	-1.0	-1.0	68.74185136897	-1.0	64.4067796610169	
2	1	1	0	6	0	10	0.6	N	55.2	43.5	57.0	12.0	-1.0	66.0	-1.0	55.5	47.4306928501738	31.7984934086629	41.8918918918919	6.38297872340426	-1.0	60.405845178186	-1.0	45.5222171323866	
1	0	1	0	8	0	10	0.8	N	60.6	84.0	-1.0	4.0	-1.0	64.75	-1.0	65.0	57.8523628713626	80.2469135802469	-1.0	3.03030303030303	-1.0	61.9058015128845	-1.0	57.3379099923722	

The background features several glowing orange lines of varying thickness and opacity, creating a sense of movement and depth. A prominent dotted orange line follows a wavy path across the upper half of the image. The overall aesthetic is futuristic and digital.

# METHODOLOGY

# IMPLEMENTING ML FOR SPEECH RATE ADJUSTMENT

- **Objective:** Enhance **real-time speech modulation** using advanced machine learning techniques.
- **Approach:** Utilize algorithms designed for **sequential** data analysis to capture the dynamic nature of speech accurately.
- **Progress:** Conducted thorough analysis using feature extraction methods such as **MFCCs**, essential for understanding intricate speech patterns.
- **Goals for the Future:** Focus on refining model **accuracy**, implementing real-time feedback mechanisms, and extending testing to include a wider array of speech datasets to ensure **robustness** and **applicability** across different scenarios.



# FEATURE PREPROCESSING

# DATA PREPARATION AND FEATURE ENGINEERING

- **Dataset Utilized:** The Crema D dataset was used, providing a diverse linguistic input.
- **Preprocessing Techniques:** Applied methods include segmentation of audio clips, normalization of audio levels, and silence trimming to enhance data quality.
- **Feature Extraction:**
  - MFCCs (Mel Frequency Cepstral Coefficients):** Crucial for capturing the timbre of speech, which is vital for distinguishing between different emotional tones.
  - Pitch and Energy Features:** Included to capture the dynamic range and intensity of speech, facilitating better classification of emotional states.
  - Speed Classification (WPM):** Calculated the WPM of each audio file and then classified them as slow, fast or at a good pace based on the percentile.



# FEATURE CLASSIFICATION

## EMOTION CLASSIFICATION

- **Random Forest Classifier:** Chosen for its **robustness** in handling **overfitting** and its ability to maintain accuracy across diverse datasets.

### **Advantages:**

- Combines multiple decision trees to ensure **generalizability** and prevent overfitting.
- Effective in **classifying** complex datasets with multiple feature types.

### **Application:**

Used to classify the 'dominant\_emotion' of speech samples, focusing solely on identifying the prevalent **emotional state** without employing metrics like F1 score for model evaluation in this context.

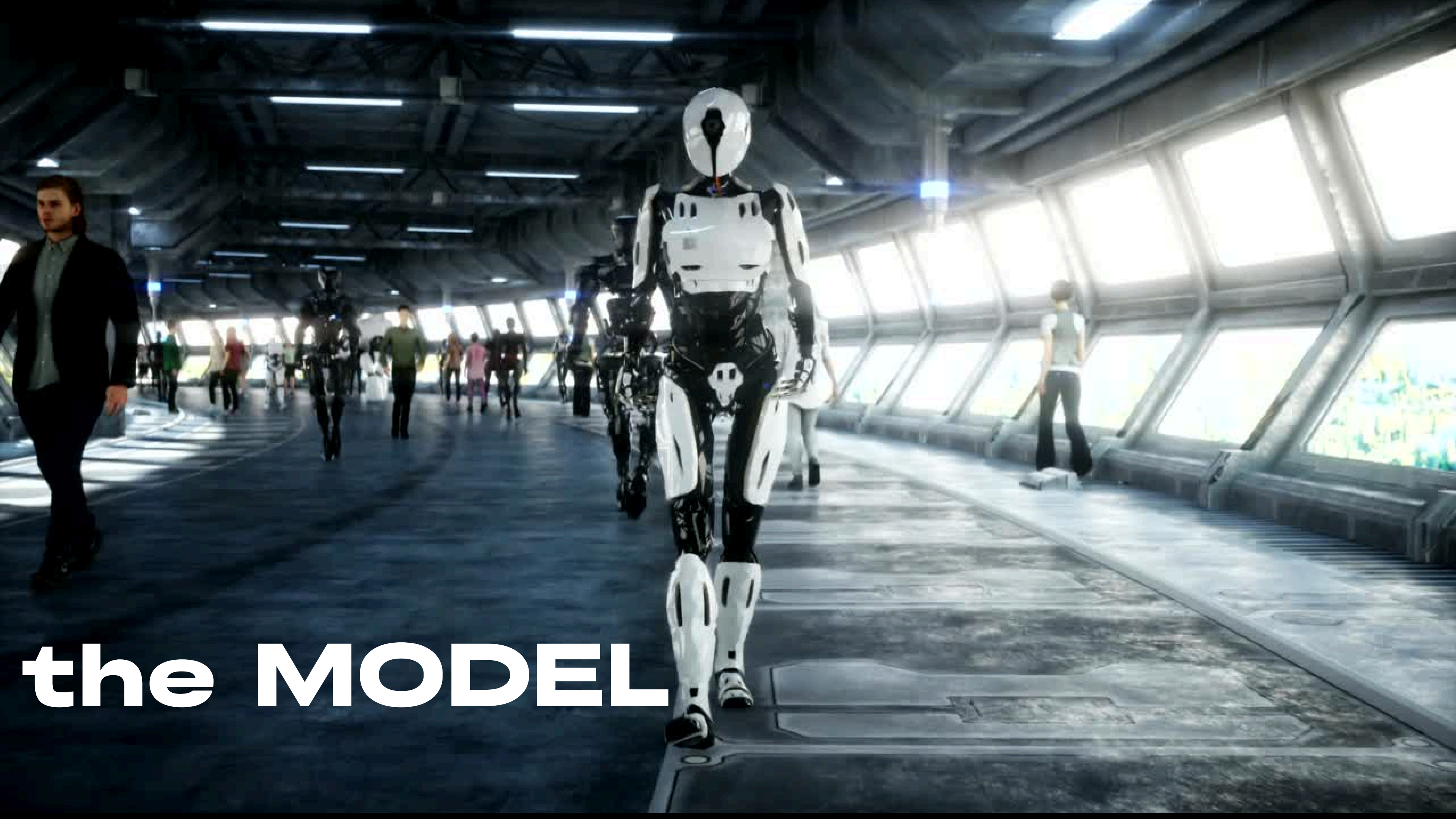
### **Training Process:**

- Split the data into **training** and **testing** sets to validate the effectiveness of the model.
- Trained to recognize various emotions based on the engineered features.

fileName	dominant_emotion	wpm	mfcc_1	mfcc_2	mfcc_3	mfcc_4	mfcc_5	mfcc_6	mfcc_7	mfcc_8	mfcc_9	mfcc_10	mfcc_11	mfcc_12	mfcc_13	mfcc_14	mfcc_15	mfcc_16	mfcc_17	mfcc_18	mfcc_19	mfcc_20	speed_category
1001_IEO_HAP_HI	Anger	66.57608696	-388.03778	131.7907	-14.469674	15.20501	-4.615429	4.121954	-7.558937	-1.4746752	-6.45826	-11.721743	-2.3501334	-9.309282	-9.791393	3.5260203	-2.82946	-9.73492	-5.8396273	0.6956932	-8.079308	-9.566392	slow
1001_IEO_HAP_HI	Happy	66.57608696	-388.03778	131.7907	-14.469674	15.20501	-4.615429	4.121954	-7.558937	-1.4746752	-6.45826	-11.721743	-2.3501334	-9.309282	-9.791393	3.5260203	-2.82946	-9.73492	-5.8396273	0.6956932	-8.079308	-9.566392	slow
1001_IEO_HAP_HI	Happy	66.57608696	-388.03778	131.7907	-14.469674	15.20501	-4.615429	4.121954	-7.558937	-1.4746752	-6.45826	-11.721743	-2.3501334	-9.309282	-9.791393	3.5260203	-2.82946	-9.73492	-5.8396273	0.6956932	-8.079308	-9.566392	slow
1001_IEO_SAD_MD	Disgust	67.55514706	-518.3452	160.27013	2.008292	37.162014	-2.8761728	27.385954	-11.937206	7.3184323	-11.25938	-1.8923516	-7.9093537	-3.4414966	-10.494672	2.3391242	0.9061716	-6.1450157	-0.7448092	-7.7682376	-4.241329	-8.927371	slow
1001_IEO_SAD_MD	Neutral	67.55514706	-518.3452	160.27013	2.008292	37.162014	-2.8761728	27.385954	-11.937206	7.3184323	-11.25938	-1.8923516	-7.9093537	-3.4414966	-10.494672	2.3391242	0.9061716	-6.1450157	-0.7448092	-7.7682376	-4.241329	-8.927371	slow
1001_IEO_SAD_MD	Neutral	67.55514706	-518.3452	160.27013	2.008292	37.162014	-2.8761728	27.385954	-11.937206	7.3184323	-11.25938	-1.8923516	-7.9093537	-3.4414966	-10.494672	2.3391242	0.9061716	-6.1450157	-0.7448092	-7.7682376	-4.241329	-8.927371	slow
1001_IEO_SAD_HI	Neutral	88.34134615	-523.32855	150.40701	7.7846985	34.56123	2.3921921	25.240118	-6.3619666	5.3296776	-8.034051	-0.9371134	-6.482719	-2.7363355	-7.662001	1.8578376	-1.8092355	-8.129779	-1.5842165	-5.070552	-3.012066	-10.162157	good pace
1001_IEO_SAD_HI	Sad	88.34134615	-523.32855	150.40701	7.7846985	34.56123	2.3921921	25.240118	-6.3619666	5.3296776	-8.034051	-0.9371134	-6.482719	-2.7363355	-7.662001	1.8578376	-1.8092355	-8.129779	-1.5842165	-5.070552	-3.012066	-10.162157	good pace
1001_IEO_SAD_HI	Fear	88.34134615	-523.32855	150.40701	7.7846985	34.56123	2.3921921	25.240118	-6.3619666	5.3296776	-8.034051	-0.9371134	-6.482719	-2.7363355	-7.662001	1.8578376	-1.8092355	-8.129779	-1.5842165	-5.070552	-3.012066	-10.162157	good pace
1001_IEO_ANG_LO	Neutral	56.71296296	-509.58966	149.28673	6.21179	25.928122	9.122308	16.558119	-1.4808415	4.692389	-4.407613	-3.39041	-1.447334	-4.2012563	-9.667849	0.5657872	0.16165301	-9.0126095	-3.6566296	-2.6409717	-7.940986	-8.460317	slow
1001_IEO_ANG_LO	Neutral	56.71296296	-509.58966	149.28673	6.21179	25.928122	9.122308	16.558119	-1.4808415	4.692389	-4.407613	-3.39041	-1.447334	-4.2012563	-9.667849	0.5657872	0.16165301	-9.0126095	-3.6566296	-2.6409717	-7.940986	-8.460317	slow
1001_IEO_ANG_LO	Neutral	56.71296296	-509.58966	149.28673	6.21179	25.928122	9.122308	16.558119	-1.4808415	4.692389	-4.407613	-3.39041	-1.447334	-4.2012563	-9.667849	0.5657872	0.16165301	-9.0126095	-3.6566296	-2.6409717	-7.940986	-8.460317	slow
1001_IEO_ANG_MD	Sad	45.03676471	-490.34747	156.41154	0.035413384	31.213022	0.7548194	15.093776	-7.9669557	8.810475	-6.6436276	-9.367609	-4.028554	-2.3922653	-10.209055	4.6429234	-1.2676045	-13.408666	-2.3444402	-1.310818	-7.7136354	-9.908755	slow
1001_IEO_ANG_MD	Neutral	45.03676471	-490.34747	156.41154	0.035413384	31.213022	0.7548194	15.093776	-7.9669557	8.810475	-6.6436276	-9.367609	-4.028554	-2.3922653	-10.209055	4.6429234	-1.2676045	-13.408666	-2.3444402	-1.310818	-7.7136354	-9.908755	slow
1001_IEO_ANG_MD	Neutral	45.03676471	-490.34747	156.41154	0.035413384	31.213022	0.7548194	15.093776	-7.9669557	8.810475	-6.6436276	-9.367609	-4.028554	-2.3922653	-10.209055	4.6429234	-1.2676045	-13.408666	-2.3444402	-1.310818	-7.7136354	-9.908755	slow
1001_IEO_ANG_HI	Neutral	90.66611842	-416.7182	131.73732	0.7154218	22.31798	1.6687083	1.6722232	-4.688296	2.7581217	-2.9017353	-12.251036	-0.50289035	-6.377226	-12.19315	4.707449	-2.4693842	-11.639656	-4.545666	1.3567295	-9.157837	-11.573683	good pace
1001_IEO_ANG_HI	Anger	90.66611842	-416.7182	131.73732	0.7154218	22.31798	1.6687083	1.6722232	-4.688296	2.7581217	-2.9017353	-12.251036	-0.50289035	-6.377226	-12.19315	4.707449	-2.4693842	-11.639656	-4.545666	1.3567295	-9.157837	-11.573683	good pace
1001_IEO_ANG_HI	Anger	90.66611842	-416.7182	131.73732	0.7154218	22.31798	1.6687083	1.6722232	-4.688296	2.7581217	-2.9017353	-12.251036	-0.50289035	-6.377226	-12.19315	4.707449	-2.4693842	-11.639656	-4.545666	1.3567295	-9.157837	-11.573683	good pace
1001_IEO_DIS_HI	Anger	58.89423077	-461.78415	137.60162	-2.592366	20.170418	11.439757	8.987415	-3.6565192	1.3593817	-5.0284343	-8.609394	0.77692	-3.2939587	-9.313088	2.2107668	-2.29295	-8.388313	-3.0056298	-1.6189089	-9.51205	-7.6704283	slow

# FEATURES

- **MFCC Mean and Standard deviation:**  
Extracted from mel-scale frequency - captures **speech timbre**.  
Mean shows average spectral shape - std shows **spectral variability**.
- **Dominating\_emotion:**  
Shows the most **prevalent** emotion that the audio file depicts
- **Words per minute:**  
Counts **words spoken per minute** - key metric for speech rate.
- **Speech rate category:** Classifies the **tempo of speech** as 'slow', 'good pace', 'fast'.



**the MODEL**

# Feature Extraction with 1D CNN:

## Role of the CNN with MFCCs

- CNN layers apply **filters** to Mel-frequency cepstral coefficients (MFCC) input.
- They detect and enhance **subtle patterns** in speech, including variations in tone, pace, and emotion.
- Variations in MFCCs can indicate changes in **emotional intonation** or differences between phonetic components such as vowels and consonants.

## Impact on Speech Analysis

- CNN identifies subtle patterns to abstract higher-level features from raw audio data.
- These features are critical for tasks like distinguishing between speech sounds, understanding speech rhythm, and detecting emotional nuances.

## CNN as a sophisticated filter

- The CNN transforms raw audio input into a **refined** set of features.
- These features reveal deeper insights into the speech's structural and emotional composition.

# Temporal Sequence Modeling with LSTM:

## Temporal Dependency Handling

- **Capability:** LSTMs can **remember** and **use past information** over long intervals, essential for capturing the flow and progression in speech patterns.
- **Mechanism:** Through gates—**forget, input, and output**—LSTMs manage the flow of information, deciding what to retain or discard over time.

## Sequence Modeling

- **Process:** As **MFCCs** move through the **LSTM**, each unit processes current inputs along with previously remembered information, gradually building a contextual understanding of speech.
- **Outcome:** This ability allows LSTMs to **recognize and predict speech** characteristics like **emotion or tempo changes** that develop over time.

## Application in Speech Analysis

- **Emotion Detection:** By tracking the emotional tone over sentences, LSTMs can **identify underlying emotions** that might evolve or be emphasized as speech progresses.
- **Speech Rate Classification:** LSTMs assess variations in speed and articulation over time, critical for determining the speech rate or identifying changes in speech dynamics.

# Technical Explanation of Dual Output Architecture in Speech Analysis Model:

## Architecture Design:

- **Integration:** The model integrates convolutional neural networks (CNNs) and long short-term memory (LSTM) units in a unified framework. This setup allows for the extraction and temporal analysis of features like MFCCs, crucial for understanding speech properties.
- **Dual Outputs:** The final layer of the model splits into two branches, each dedicated to a specific task—emotion recognition and speech speed classification.

## Emotion Recognition:

- **Feature Utilization:** Emotion classification exploits subtle variations in MFCCs that correspond to emotional expressions. These features capture essential vocal nuances like pitch and tone, which are indicators of emotional states.
- **Processing Flow:** After feature extraction by CNNs, LSTMs analyze these features over time, enhancing the model's ability to detect emotional nuances that develop throughout the speech.

## Speech Speed Classification:

- **Temporal Dynamics:** The speech speed is determined by analyzing the rate at which words are spoken, which involves calculating the temporal intervals between spoken words or phonetic elements.
- **Categorization Method:** The model categorizes speech into 'slow', 'moderate', or 'fast' based on predefined thresholds of words per minute, utilizing the temporal data processed by LSTMs to assess the overall speech tempo.





# Model Results

## Emotion Classification:

- **F1 Score (Weighted Average): 0.34.** Indicates poor performance, showing that the model struggles to correctly classify various emotions.
- **Accuracy: 44%.** This highlights the model's limited ability to distinguish between different emotions accurately.
- **Confusion Matrix Analysis:** Most emotions are incorrectly predicted as 'Neutral', suggesting a bias or imbalance towards this class.
- Very low recall for most classes except 'Neutral', indicating a lack of sensitivity to other emotional states.

## Speed Classification:

- **F1 Score (Weighted Average): 0.53.** This is better than emotion classification but still indicates room for improvement.
- **Accuracy: 54%.** Higher than emotion accuracy, suggesting that the model is somewhat better at recognizing speech speed.
- **Confusion Matrix Analysis:** The model shows some confusion between 'slow' and 'fast' speeds, often misclassifying them as 'moderate'.
- 'Moderate' speed has the highest recall, indicating that the model is most reliable in detecting this speed category.

These statistics reflect specific challenges in emotion recognition while showing relatively better performance in speed classification.

# Confusion Matrix

