

The background is a vibrant blue with a subtle, repeating pattern of circuit board traces and circular nodes. A central black rectangular box with rounded corners contains the title text in white, uppercase, sans-serif font. The text is centered within the box and reads: "THE DEVELOPMENT OF AN AUTOMATED RATING SYSTEM FOR A SIMULATED ORAL PROFICIENCY INTERVIEW".

THE DEVELOPMENT OF AN  
AUTOMATED RATING SYSTEM  
FOR A SIMULATED ORAL  
PROFICIENCY INTERVIEW



# OUR TEAM



Payman Vafaee

Teachers College,  
Columbia University



Manraj Grover

MIDAS Labs, IIITD



Pakhi Bamdev

The University of  
Edinburgh



Yaman Kumar

University at Buffalo



Mika Hama

Second Language  
Testing, Inc. (SLTI)



VOLUME 3



# AUTOMATED SPEAKING ASSESSMENT

Using Language Technologies to Score Spontaneous Speech

Edited by Klaus Zechner and Keelan Evanini

Innovations in **LANGUAGE LEARNING** and **ASSESSMENT** at *ETS*



Measuring the Power of Learning®

The background is a vibrant blue with a subtle, repeating pattern of circuit board traces and circular nodes. A central black rectangular box with rounded corners contains the title text in white, uppercase letters. The text is centered within the box and reads: "THE DEVELOPMENT OF AN AUTOMATED RATING SYSTEM FOR A SIMULATED ORAL PROFICIENCY INTERVIEW".

THE DEVELOPMENT OF AN  
AUTOMATED RATING SYSTEM  
FOR A SIMULATED ORAL  
PROFICIENCY INTERVIEW

# APPROACHES



Supervised feature-based classical machine learning



Unsupervised deep learning

# SIMULATED ORAL PROFICIENCY INTERVIEW

Task 1

Next: Task 1: Preparation Time

NEXT ►



You are at a job interview for an amusement park in California. Mr. Brown, the interviewer, asks you to tell him about yourself. He asks about your background, why you want to work at the amusement park, and what your long-term goals are for the future. Tell Mr. Brown about yourself, why you want to work at the amusement park, and what you plan on doing in the future.

# SIMULATED ORAL PROFICIENCY INTERVIEW

Task 1: Preparation Time

Next: Task 1

28

NEXT ▶



**You have 30 seconds to prepare**

You are at a job interview for an amusement park in California. Mr. Brown, the interviewer, asks you to tell him about yourself. He asks about your background, why you want to work at the amusement park, and what your long-term goals are for the future. Tell Mr. Brown about yourself, why you want to work at the amusement

park, and what you plan on doing in the future.

# SIMULATED ORAL PROFICIENCY INTERVIEW

Task 1

Next: Break 2



Recording

58

SAVE & NEXT ▶



**You have 60 seconds to respond**

You are at a job interview for an amusement park in California. Mr. Brown, the interviewer, asks you to tell him about yourself. He asks about your background, why you want to work at the amusement park, and what your long-term goals are for the future. Tell Mr. Brown about yourself, why you want to work at the amusement

park, and what you plan on doing in the future.



# THE SLTI SOPI



EIGHT EQUAL FORMS



SIX TEST TASKS

B1

▶ Task 1 - Task 1

B1

→ Task 2

B2

→ Task 3

C1

→ Task 4

C1

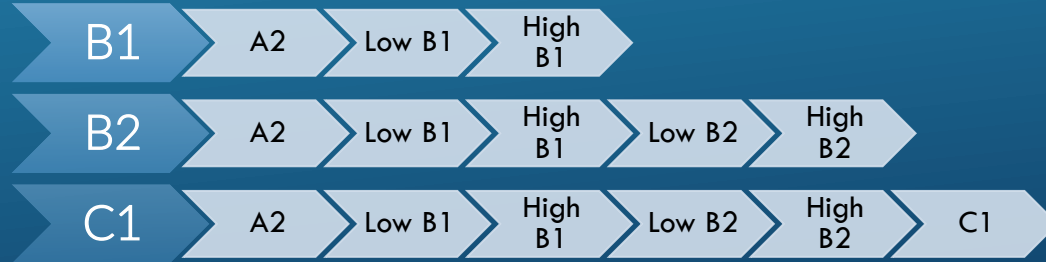
→ Task 5

B1

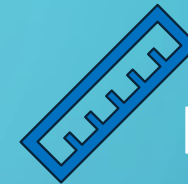
→ Task 6



You are at a job interview for an amusement park in California. Mr. Brown, the interviewer, asks you to tell him about yourself. He asks about your background, why you want to work at the amusement park, and what your long-term goals are for the future. Tell Mr. Brown about yourself, why you want to work at the amusement park, and what you plan on doing in the future.



# RATING



HOLISTIC RUBRIC



CEFR ALIGNED



DOUBLE-BLIND



SCORING ALGORITHM

## Fluency

Has a sufficient range of language to be able to give clear descriptions **without much conspicuous searching for words** about the topic. Can give **clear and detailed** descriptions in response to the provided scenario with **relevant examples**, though his/her language lacks expressive power and idiomaticity. Can **sustain speech without many noticeable hesitations**, but the listener may have to occasionally make an effort to understand.

## Vocabulary

## Content

## Grammar

Has a good command of simple **language structures** and some **complex grammatical forms**. Can use word **stress, intonation, rhythm** to support the message he/she intends to convey, though influence from other languages he/she speaks is present.

## Pronunciation

Low B2



# TEST VALIDATION

# GENERALIZABILITY

Form	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
<i>N</i>	5485	5450	5391	5384	5274	5206	5345	5244


Facet/Percentages	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
<b>p</b>	89.1	89.93	88.27	88.48	91.23	88.5	89	89.42
<b>t</b>	2.5	2.04	4.01	3.98	1.25	3.11	2.79	2.34
<b>r'</b>	0	0	0	0	0	0	0	0
<b>pt</b>	6.54	6.37	6.13	5.93	3.99	6.81	6.6	6.6
<b>pr'</b>	.58	.43	.31	.31	0	.28	.4	.32
<b>tr</b>	0	0	0	0	.78	0	0	0
<b>ptr,e</b>	1.28	1.23	1.28	1.3	2.75	1.3	1.21	1.32
<b>G</b>	<b>.91</b>	<b>.91</b>	<b>.92</b>	<b>.92</b>	<b>.93</b>	<b>.91</b>	<b>.92</b>	<b>.91</b>
<b>Phi</b>	<b>.91</b>	<b>.9</b>	<b>.88</b>	<b>.88</b>	<b>.91</b>	<b>.88</b>	<b>.89</b>	<b>.89</b>

# RELIABILITY

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
<b>Person Reliability</b>	.93	.93	.93	.94	.92	.93	.93	.93
<b>Number of Levels</b>	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4
<b>Task Reliability</b>	1	1	1	1	1	1	1	1
<b>Number of Levels</b>	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4
<b>Ratings Reliability</b>	0	0	0	0	0	0	0	0
<b>Number of Levels</b>	1	1	1	1	1	1	1	1

Task	Form 1		Form 2	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.17	1.13	1.23	1.13
B1-2	1.5	1.39	1.41	1.32
B1-3	1.49	1.35	1.76	1.7
B2	1.24	1.21	.88	.72
C1-1	.95	.8	1.03	.99
C1-2	1.06	1.01	1	.99
	Form 3		Form 4	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.3	1.2	1.28	1.16
B1-2	1.22	1.14	1.33	1.24
B1-3	1.14	1.06	1.13	1.06
B2	.95	.85	.89	.74
C1-1	1.21	1.28	1.32	1.76
C1-2	1.39	1.06	1.17	1.31
	Form 5		Form 6	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.17	1.07	1.23	1.1
B1-2	1.5	1.39	1.28	1.19
B1-3	1.13	1.09	1.87	1.76
B2	1.14	1.08	.77	.61
C1-1	.85	.69	1.17	1.25
C1-2	1.31	1.69	1.1	1.13
	Form 7		Form 8	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.22	1.11	1.31	1.2
B1-2	1.68	1.55	1.3	1.2
B1-3	1.46	1.38	1.68	1.6
B2	1.05	.97	1.01	.9
C1-1	.94	.86	1.07	1.06
C1-2	1.23	1.32	1.03	.92

Rating	Form 1		Form 2	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.24	1.15	1.21	1.14
2	1.24	1.14	1.2	1.13
	Form 3		Form 4	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.21	1.19	1.19	1.21
2	1.19	1.18	1.18	1.22
	Form 5		Form 6	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.19	1.18	1.23	1.17
2	1.18	1.16	1.22	1.17
	Form 7		Form 8	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.26	1.2	1.22	1.13
2	1.25	1.19	1.23	1.16



# SCORING SYSTEM BUILDING



# DATA SET

(FORM 3)

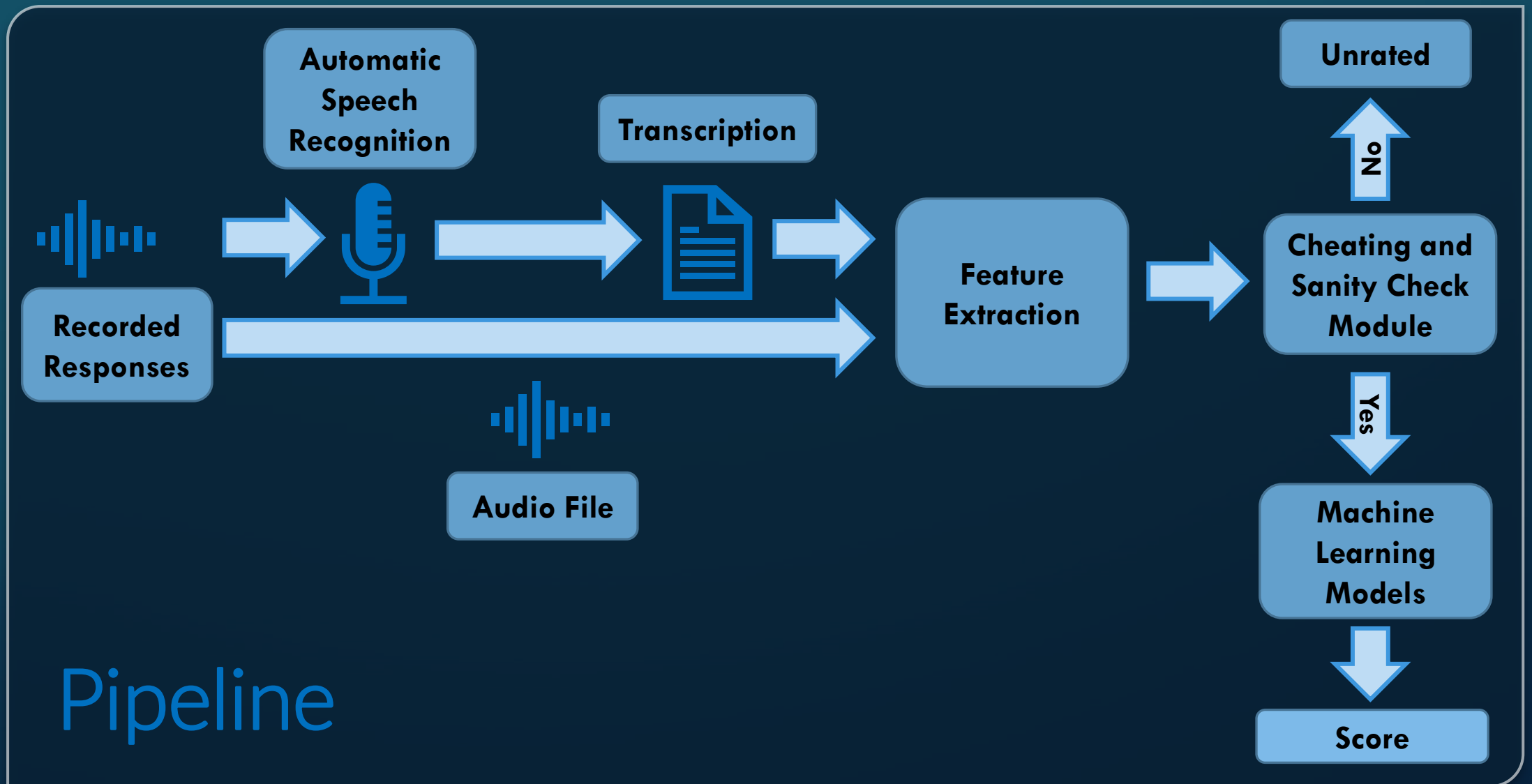
			Sz		DS				
P	N	L	Duration	Length	A2	LB1	HB1	LB2	HB2
1	7877	B1	57.67	100.69	275	1557	6045	–	–
2	7432	B1	58.72	110.03	465	2824	4143	–	–
3	8042	B2	81.43	148.96	117	664	3493	3666	102
4	8020	C1	104.15	180.73	121	720	3536	3534	109
5	7936	C1	105.95	196.55	110	551	3004	4120	151
6	8002	B1	55.87	109.38	119	1028	6855	–	–
<b>Total</b>			<b>47309</b>						

# DATA SET

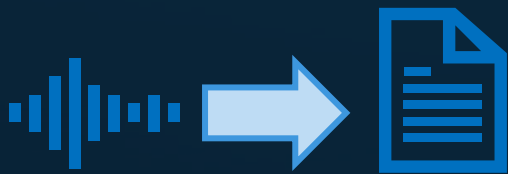
(FORM 3)

A stratified data split with a ratio of 70: 10: 20

P	ST	#R	MS	DS				
				A2 (0)	LB1 (1)	HB1 (2)	LB2 (3)	HB2 (4)
1	Train	5670	1.732	185	1152	4333	-	-
	Validate	631	1.737	27	112	492	-	-
	Test	1576	1.734	63	293	1220	-	-
2	Train	5176	1.496	319	1973	2884	-	-
	Validate	808	1.499	63	279	466	-	-
	Test	1448	1.49	83	572	793	-	-
3	Train	5641	2.375	75	472	2423	2602	69
	Validate	821	2.378	13	66	353	376	13
	Test	1580	2.344	29	126	717	688	20
4	Train	5774	2.341	93	529	2549	2521	82
	Validate	642	2.364	7	58	284	280	13
	Test	1604	2.365	21	133	703	733	14
5	Train	5713	2.451	89	406	2159	2957	102
	Validate	635	2.465	6	37	265	310	17
	Test	1588	2.491	15	108	580	853	32
6	Train	5760	1.839	92	744	4924	-	-
	Validate	641	1.861	5	79	557	-	-
	Test	1601	1.844	22	205	1374	-	-



# Automatic Speech Recognition



An end-to-end [DeepSpeech 2](#) (Amodei et al. [2016](#)) architecture followed by a [4-gram](#) language model decoder.

Trained on approximately [1,000 hours](#) of audio sampled from [CommonVoice](#) (Ardila et al. [2020](#)) and [LibriSpeech](#) dataset (Panayotov et al. [2015](#)).

Further fine-tuned on approximately [45 hours](#) (or 3558 samples) of transcribed non-native spoken responses sampled from our 8-form dataset.

Word Error Rate (WER) of [20.2%](#) on approximately [10 hours](#) of unseen spoken responses.

## Forced Aligner



To produce a [time-aligned](#) representation of [words](#) and [phonemes](#).

We produced [stress markings on the vowels](#) (no stress, primary stress, secondary stress, tertiary stress).

A pre-trained [Montreal Forced Aligner \(MFA\)](#) (McAuliffe et al. [2017](#)) trained on [LibriSpeech](#) dataset (Panayotov et al. [2015](#)).

## Cheating and Sanity Check



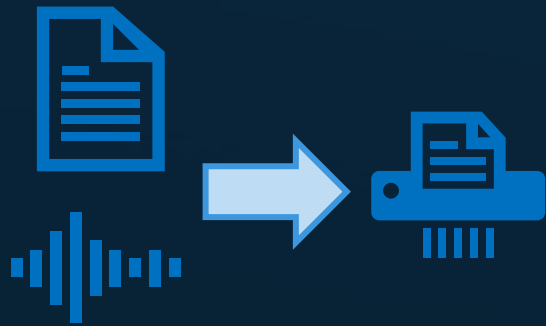
Language detection

Reading off the internet

Repeating question

Reciting prayers and poems

## Feature Extraction



From both **audio** and **text**

Five groups of hand-crafted **linguistic** and **content** features

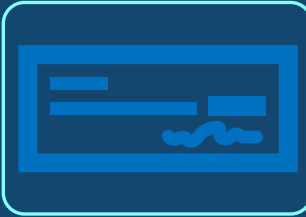
Acoustic features such as **pitch**

# FEATURES



## Delivery

- Fluency
- Pronunciation



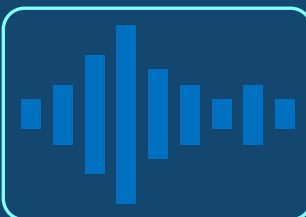
## Language Use

- Grammar
- Vocabulary



## Content

- Response-based features



## Acoustic Features

- Time-sequenced prosodic features



# FEATURES



## Delivery: Breakdown Fluency

Feature Name	Definition	Correlation (SOPI)	Correlation (ETS)
filled_pause_rate	Number of filled pauses (uh, um) per second.	-0.14	-0.23
general_silence	Number of silences. (silent duration between two words greater than 0.145 seconds)	-0.15	-0.26
mean_silence	Mean duration of silences in seconds.	-0.28	-0.32
silence_absolute_deviation	Mean absolute difference of silence duration.	-0.27	-0.32
SilenceRate1	Number of silences divided by total number of words.	-0.38	-0.50
SilenceRate2	Number of silences divided by total response duration in seconds.	-0.24	-0.45
long_silence_deviation	Mean deviation of long silences in seconds. (silent duration between two words greater than 0.495 seconds.	-0.16	-0.26

# FEATURES



## Delivery: Speed Fluency

Feature Name	Definition	Correlation (SOPI)	Correlation (ETS)
speaking_rate	Number of words per second in total response duration.	0.44	0.54
articulation_rate	Number of words per second in total articulation time.	0.21	0.38
longpfreq	Frequency of long pauses normalized by response length in words	-0.42	-

# FEATURES



## Delivery: Supersegmental Pronunciation

### Stress-based features

Feature Name	Definition	Correlation (SOPI)	Correlation (ETS)
StressedSyllPercent	Relative frequency of stressed syllables in percent.	0.04	0.38
StressDistanceSyllMean	Mean distance between stressed syllables in syllables.	-0.12	-0.37
StressDistanceSyllSD	Mean deviation of distances between stressed syllables in syllables.	-0.08	-0.33
StressDistanceMean	Mean distance between stressed syllables in seconds.	-0.20	-0.47
StressDistanceSD	Mean deviation of distances between stressed syllables in seconds.	-0.16	-0.41

# FEATURES



## Delivery: Supersegmental Pronunciation

### Interval-based features

Feature Name	Definition	Correlation (SOPI)	Correlation (ETS)
vowelPercentage	Percentage of speech that consists of vowels.	-0.24	-0.30
vowelDurationSD	Standard Deviation of vowel segments.	-0.15	-0.26
consonantDurationSD	Standard Deviation of consonantal segments.	-0.09	-0.20
syllableSDNorm	Standard Deviation of syllable segments divided by mean length of syllable segments.	-0.15	-0.24
vowelPVI	Raw Pairwise Variability Index for vocalic segments.	-0.19	-0.39
consonantPVI	Raw Pairwise Variability Index for consonantic segments.	-0.13	-0.36
syllablePVI	Raw Pairwise Variability Index for syllable segments.	-0.19	-0.4
vowelPVINorm	Normalized Pairwise Variability Index for vocalic segments.	-0.19	-0.25
consonantPVINorm	Normalized Pairwise Variability Index for consonantic segments.	-0.24	-0.32
syllablePVINorm	Normalized Pairwise Variability Index for syllable segments.	-0.25	-0.29

# FEATURES



## Language Use: Grammatical Complexity

POS-based, Clause-based, and Phrase-based Grammatical Complexity Features

Syntactic Analyzer Tool (Lu [2010](#))

Feature Name	Definition	Correlation (SOPI)
MLS	Mean length of sentence.	0.07
MLT	Mean length of T-unit.	0.02
MLC	Mean length of clause.	-0.04
C/T	Number of clauses per sentence.	0.06
VP/T	Number of verb phrases per T-unit.	0.01
DC/C	Number of dependent clauses per clauses.	0.1
DC/T	Number of dependent clauses per T-unit.	-0.03
T/S	T-units per sentence.	0.05
CT/T	Complex T-unit per T-unit.	0.06
CP/T	Coordinate Phrase per T-unit.	-0.03
CP/C	Coordinate Phrase per clause.	-0.05
CN/T	Complex Nominal per T-unit.	0.01
CN/C	Complex Nominal per clause.	-0.03

# FEATURES



## Language Use: Grammatical Complexity

### Count-based Grammatical Complexity Features

Feature Name	Definition	Correlation (SOPI)
total_adjectives*	Total number of adjectives.	0.3
total_adverbs*	Total number of adverbs.	0.4
total_nouns*	Total number of nouns.	0.31
total_verbs*	Total number of verbs.	0.36
total_pronoun*	Total number of pronouns.	0.28
total_conjunctions*	Total number of conjunctions.	0.37
total_determiners*	Total number of duration.	0.29
total_text_complexity_no_sw_mAvg*	Total text complexity (average values) when stop words are removed.	0.39
average_word_complexity_no_sw_mAvg*	Average text complexity (average values) when stop words are removed.	0.03
total_text_complexity_mAvg*	Total text complexity (average values) when no stop words are removed.	0.39

# FEATURES



## Language Use: Vocabulary Diversity & Sophistication

Lexical Analyzer  
Tool (Ai &  
Lu 2010)

Feature Name	Definition	Correlation (SOPI)
total_adjectives*	Total number of adjectives.	0.3
total_adverbs*	Total number of adverbs.	0.4
total_nouns*	Total number of nouns.	0.31
total_verbs*	Total number of verbs.	0.36
total_pronoun*	Total number of pronouns.	0.28
total_conjunctions*	Total number of conjunctions.	0.37
total_determiners*	Total number of duration.	0.29
total_text_complexity_no_sw_mAvg*	Total text complexity (average values) when stop words are removed.	0.39
average_word_complexity_no_sw_mAvg*	Average text complexity (average values) when stop words are removed.	0.03
total_text_complexity_mAvg*	Total text complexity (average values) when no stop words are removed.	0.39

# FEATURES

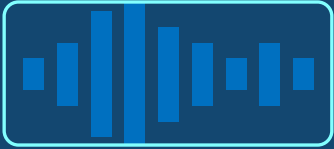


## Content

- Response-based content analysis
- Content Vector Analysis (Attali & Burstein, 2006)
- Extracted word frequencies weighted by inverse document frequency values (TF-IDF) word vectors for each test response




# FEATURES



## Acoustic Features

Feature Name	Definition	Correlation (SOPI)
stdev_energy	Standard deviation of energy of the response.	0.04
mean_pitch	Mean pitch of the response.	-0.05
stdev_pitch	Standard deviation of pitch of the response.	-0.04
range_pitch*	Range of pitch of the response.	0.14
zero_crossing_rate	Rate of sign change across the audio signal.	-0.05
energy_entropy	Entropy of the normalised energy of sub-frames of the audio signal.	0.24
spectral_centroid	Weighted average of all the frequencies in the given response signal. It is closely related to the brightness of a sound.	-0.01
Jitter and its variants	Measure of frequency instability. Variants that were calculated are rapJitter, ppq5Jitter, and ddpJitter.	-0.03
Shimmer and its variants	Measure of amplitude instability. Variants that were calculated are localShimmer, apq3Shimmer, aqq5Shimmer and ddaShimmer.	-0.04



# MACHINE LEARNING MODELS

- Both **classification** and **regression** models
  - For regression, **Linear Multiple Regression**
  - For classification, **Logistic Regression**
- Alternative machine learning models:
  - **XGBoost**
  - **Gradient Boosted Trees**
  - **Random Forest**
  - **Support Vector Machine**



## EVALUATION INDICES

- Quadratic Weighted Kappa (QWK)
- Pearson correlation coefficient ( $r$ )
- Mean Squared Error (MSE)

Model	P 1	P 2	P 3	P 4	P 5	P 6
	QWK	QWK	QWK	QWK	QWK	QWK
LR (baseline)	.3	.13	.34	.36	.4	.28
GBT	.46	.3	.3	.5	.53	.44
RF	.52	.3	.48	.55	.53	.4
SVM	.42	.15	.24	.3	.35	.42
XGBoost	.52	.3	.5	.56	.54	.44
HH	.69	.56	.78	.8	.84	.68

Ave: .25

Ave: .48

Ave: .72

# SELECTED RESULTS

# APPROACHES



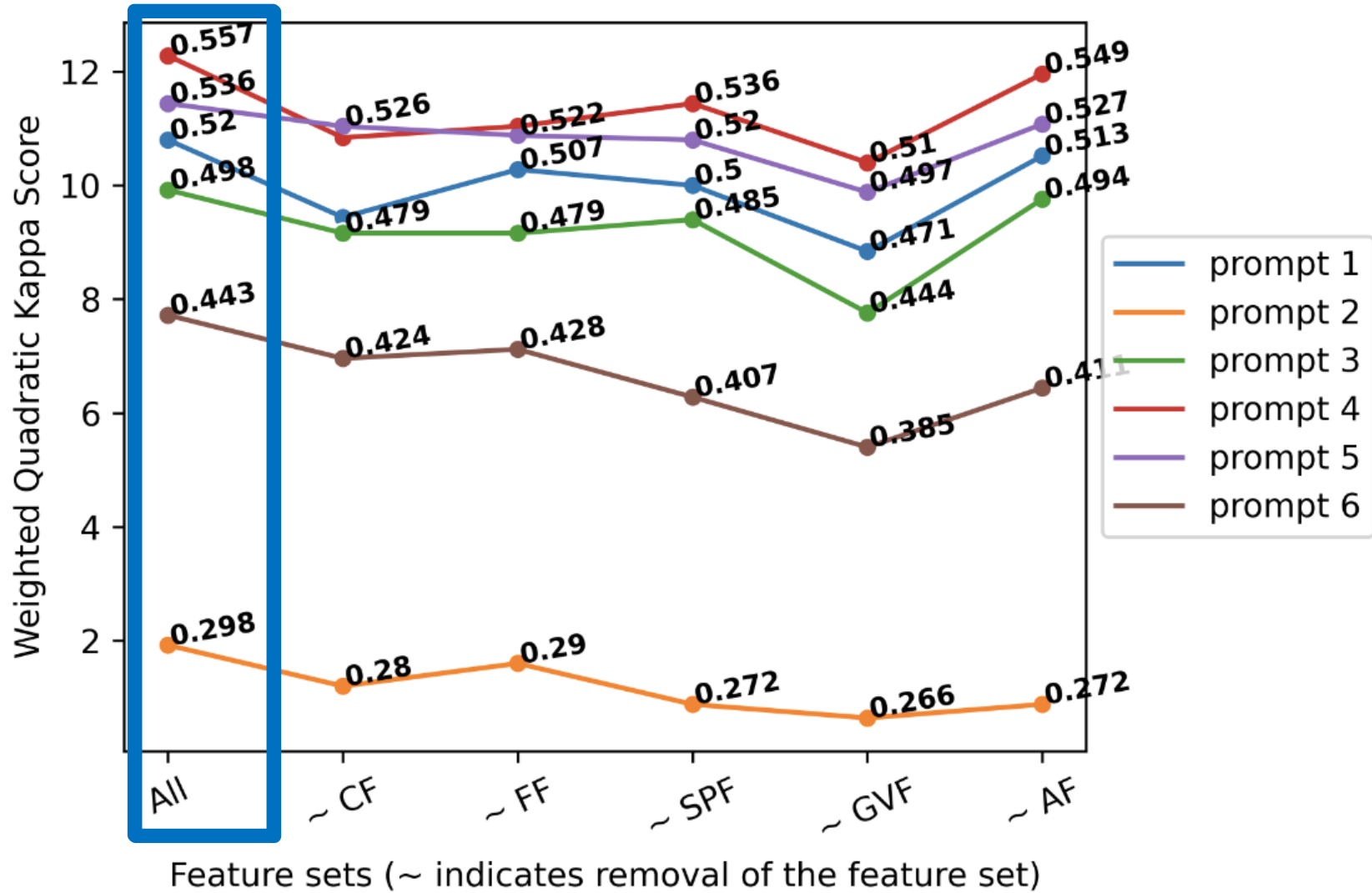
Supervised feature-based classical machine learning

Ave: .48

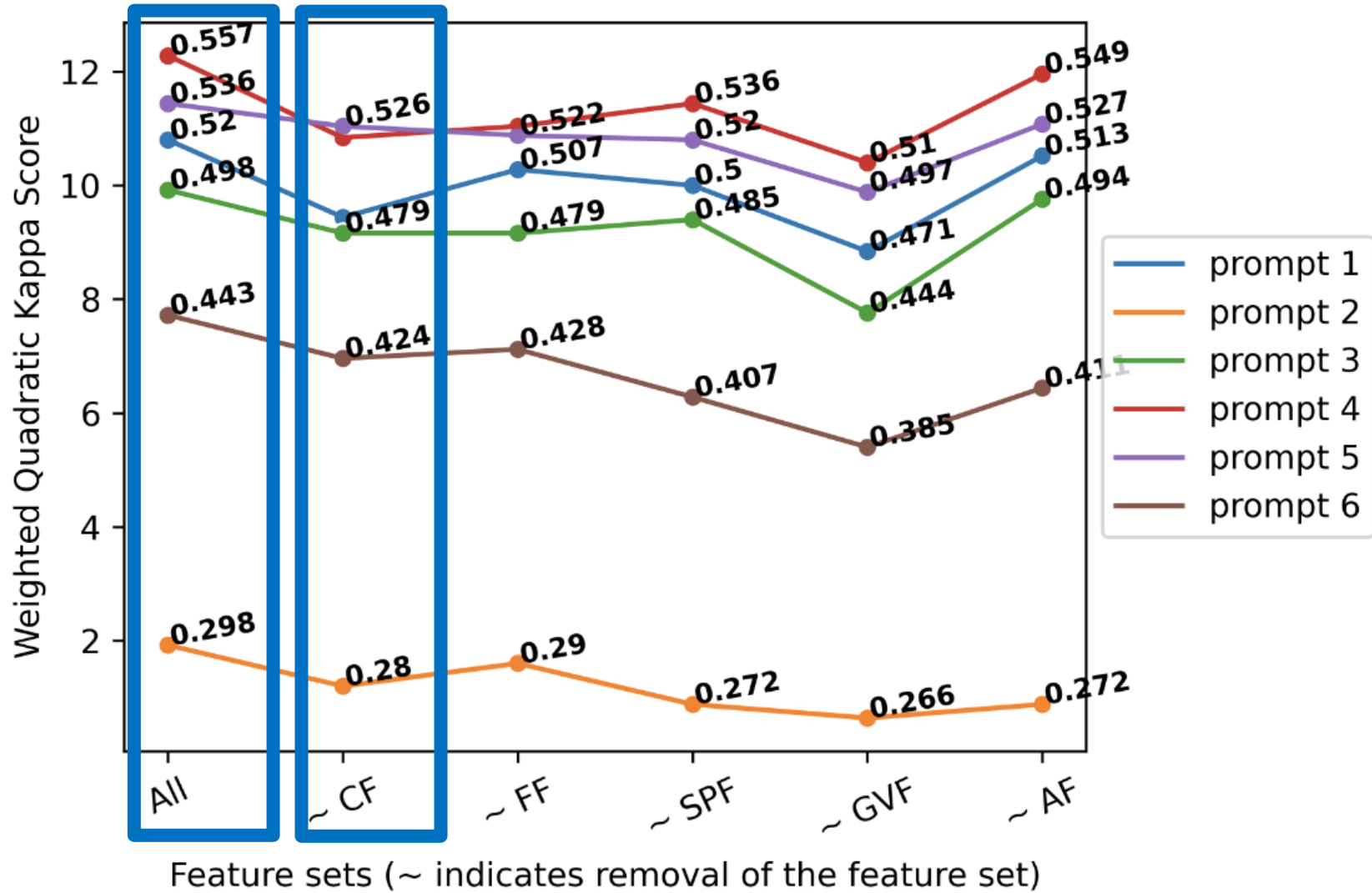


Unsupervised deep learning

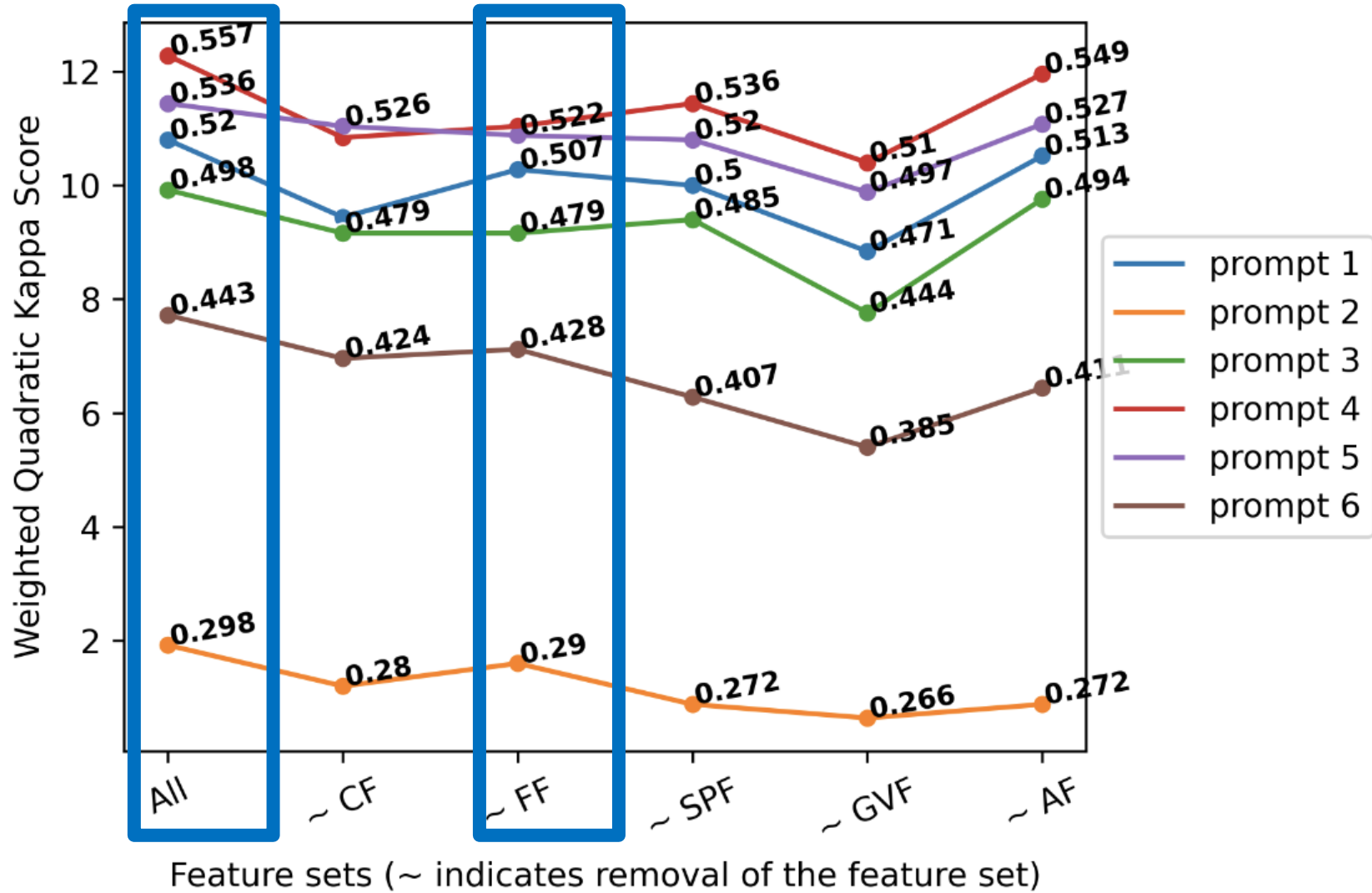
Ave: .62



# ABLATION STUDY

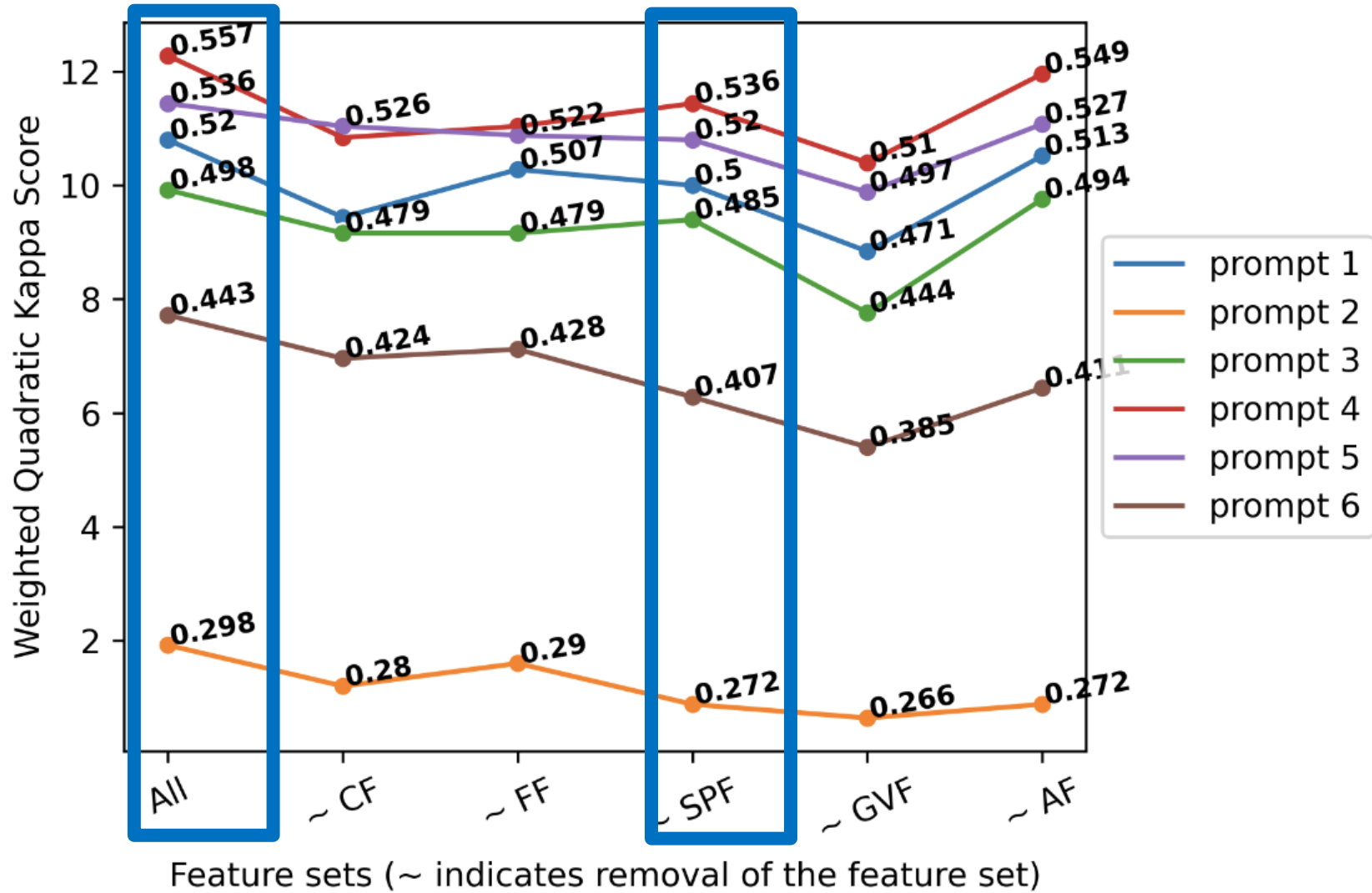


# ABLATION STUDY

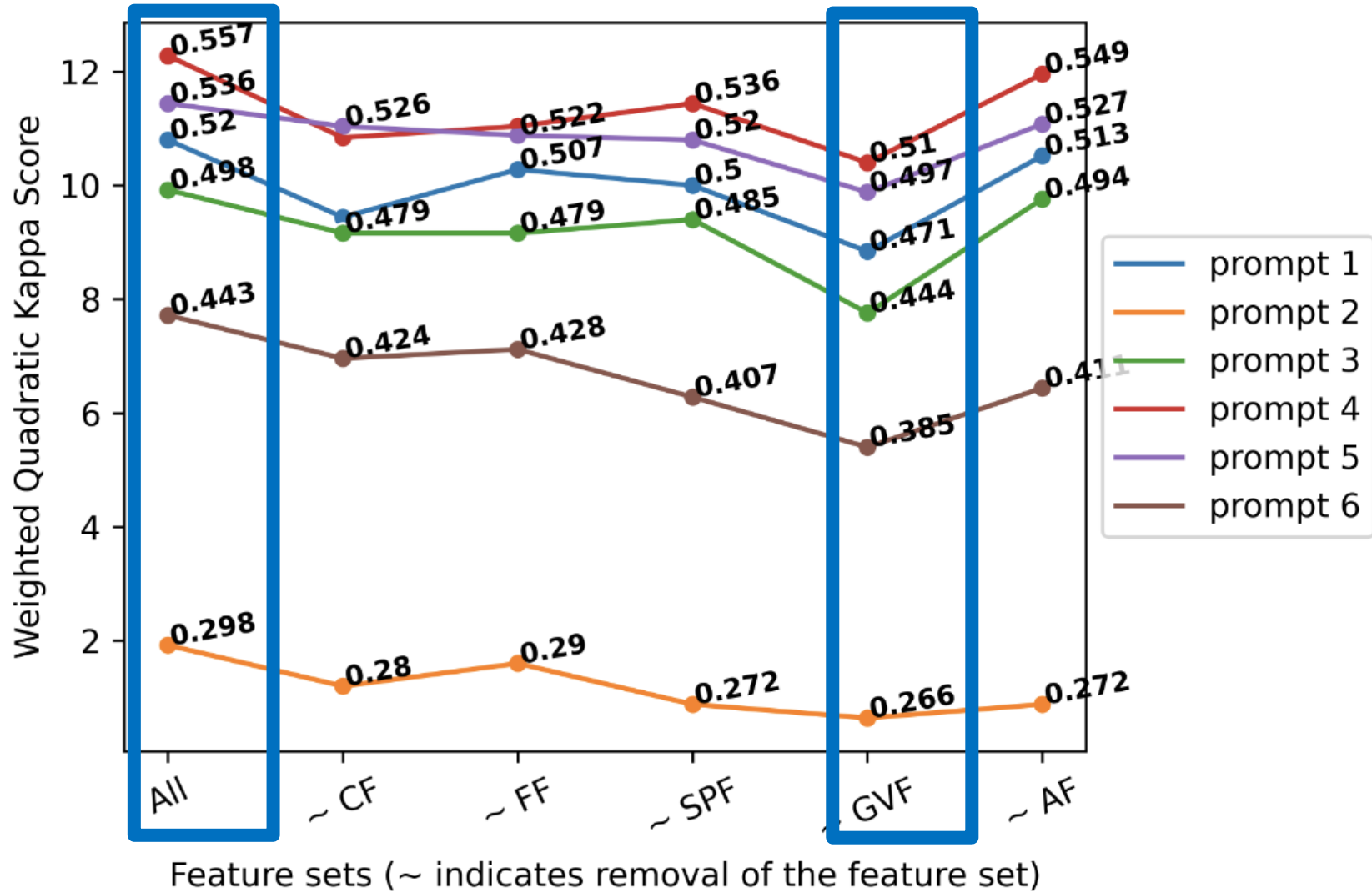


# ABLATION STUDY

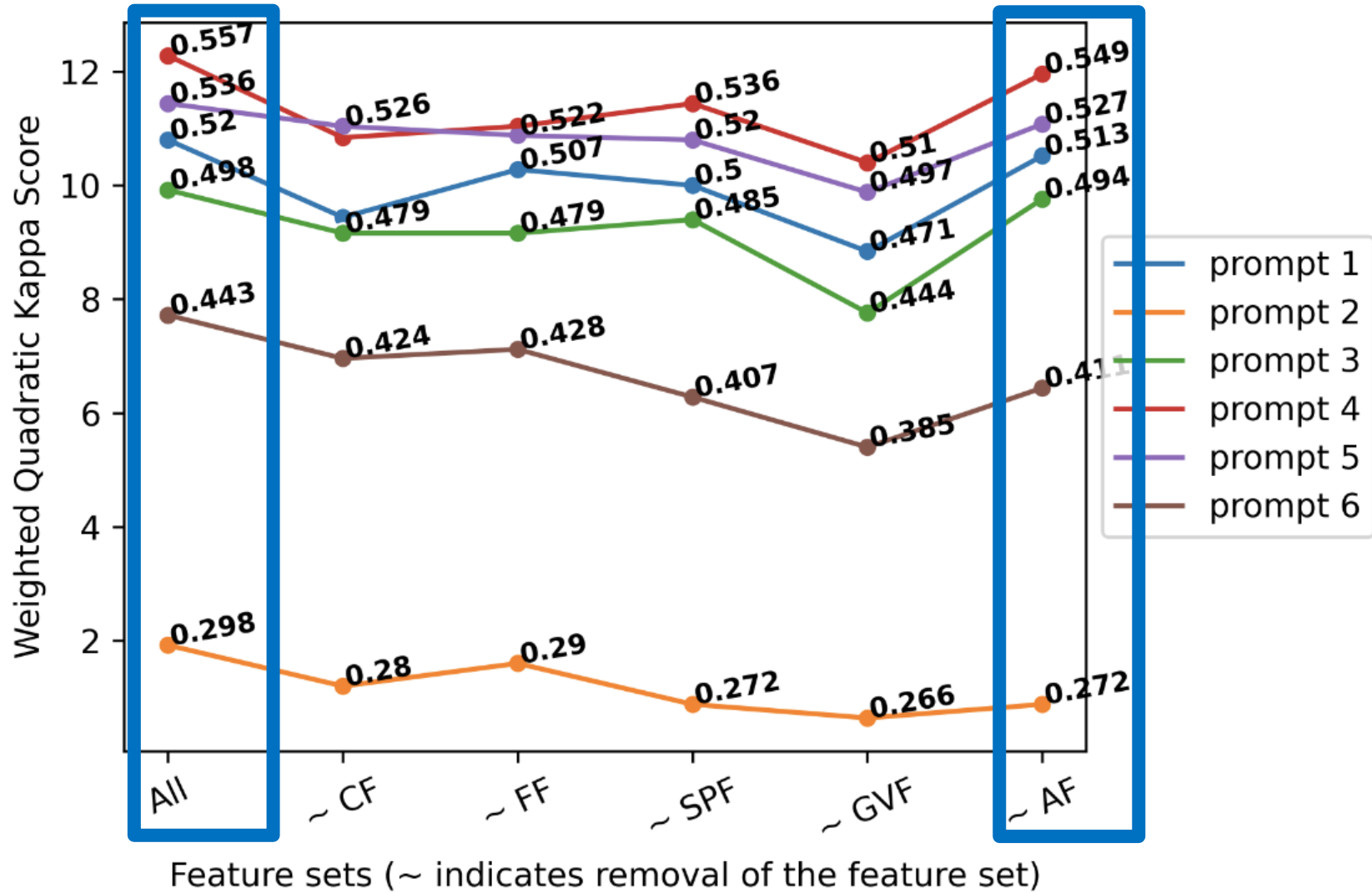




# ABLATION STUDY

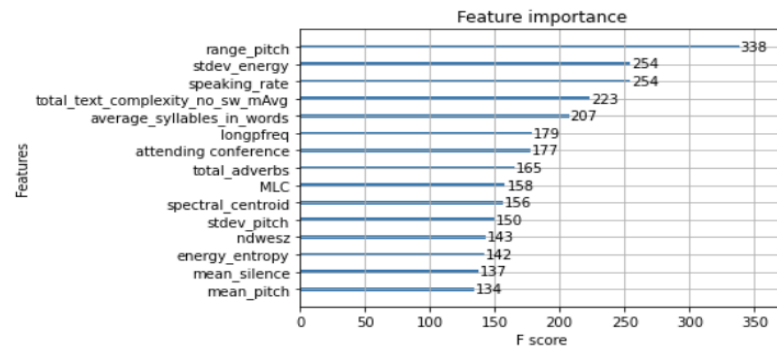


# ABLATION STUDY

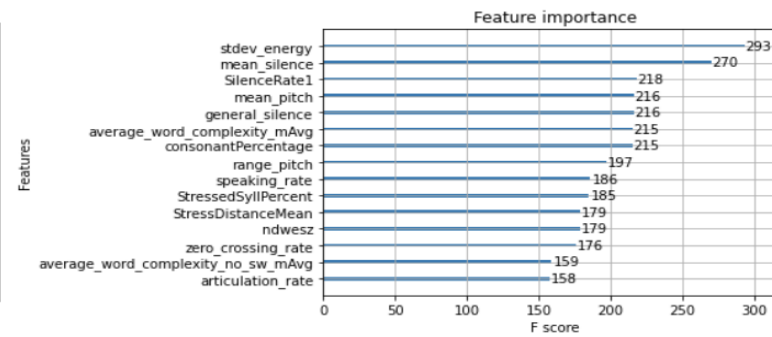


# ABLATION STUDY

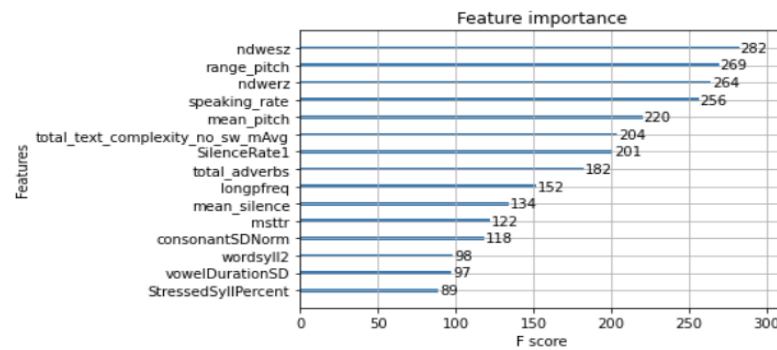
# INDIVIDUAL FEATURE IMPORTANCE



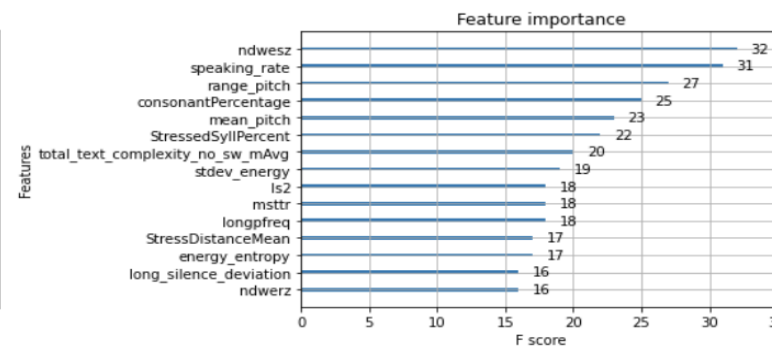
(a) Prompt 1



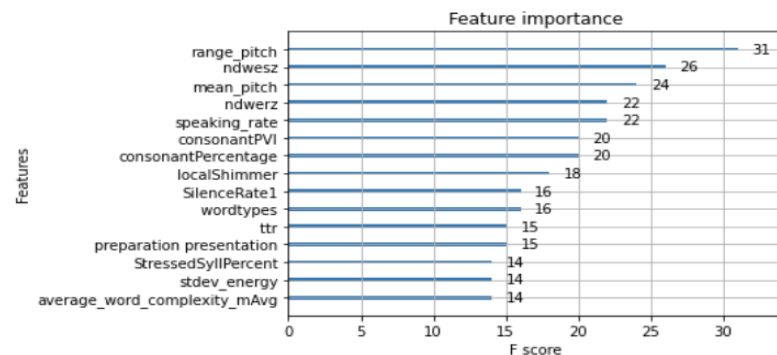
(b) Prompt 2



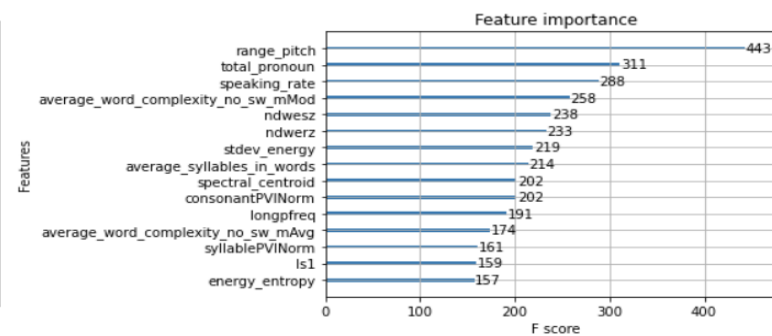
(c) Prompt 3



(d) Prompt 4



(e) Prompt 5



(f) Prompt 6

Pitch  
Speaking rate  
Silence mean  
Audio energy  
Stress patterns  
Number of different words

# IMPROVEMENTS

- Improving the automatic speech recognition system
- More (balanced) data
- Add sub-features for the current features
  - Tone-based suprasegmental features
- Add new features:
  - Repair fluency
  - Segmental pronunciation
  - Grammatical accuracy
  - Discourse coherence

The background is a solid dark blue color. In the four corners, there are decorative white line-art patterns that resemble circuit traces or a network diagram. These patterns consist of thin lines that branch out and terminate in small white circles. The patterns are symmetrical and frame the central text.

THANK YOU