A report on Speech Emotion Recognition

Postgraduate Science and Technology Conference

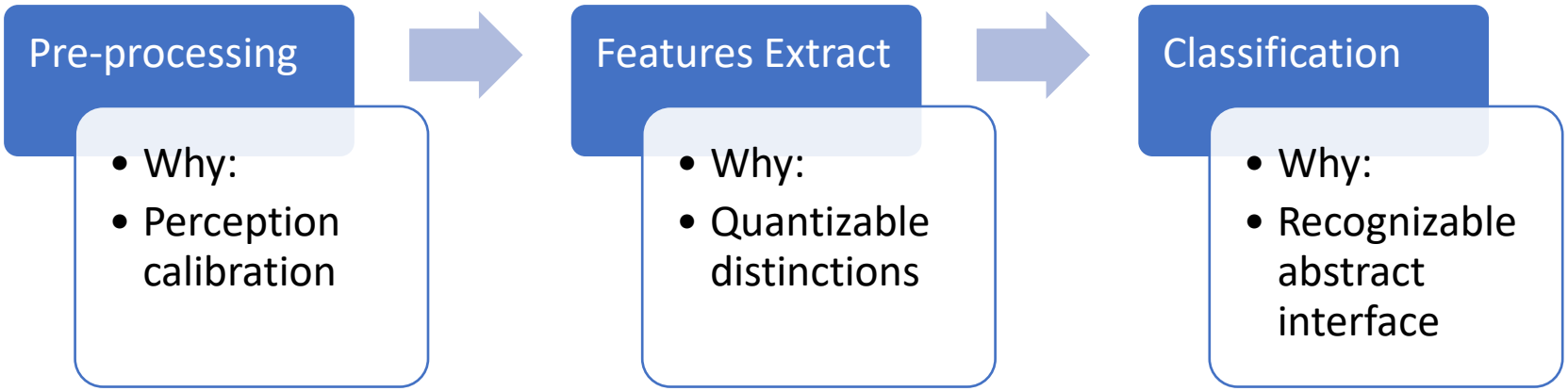
School of Automation, CUG

October 10, 2020

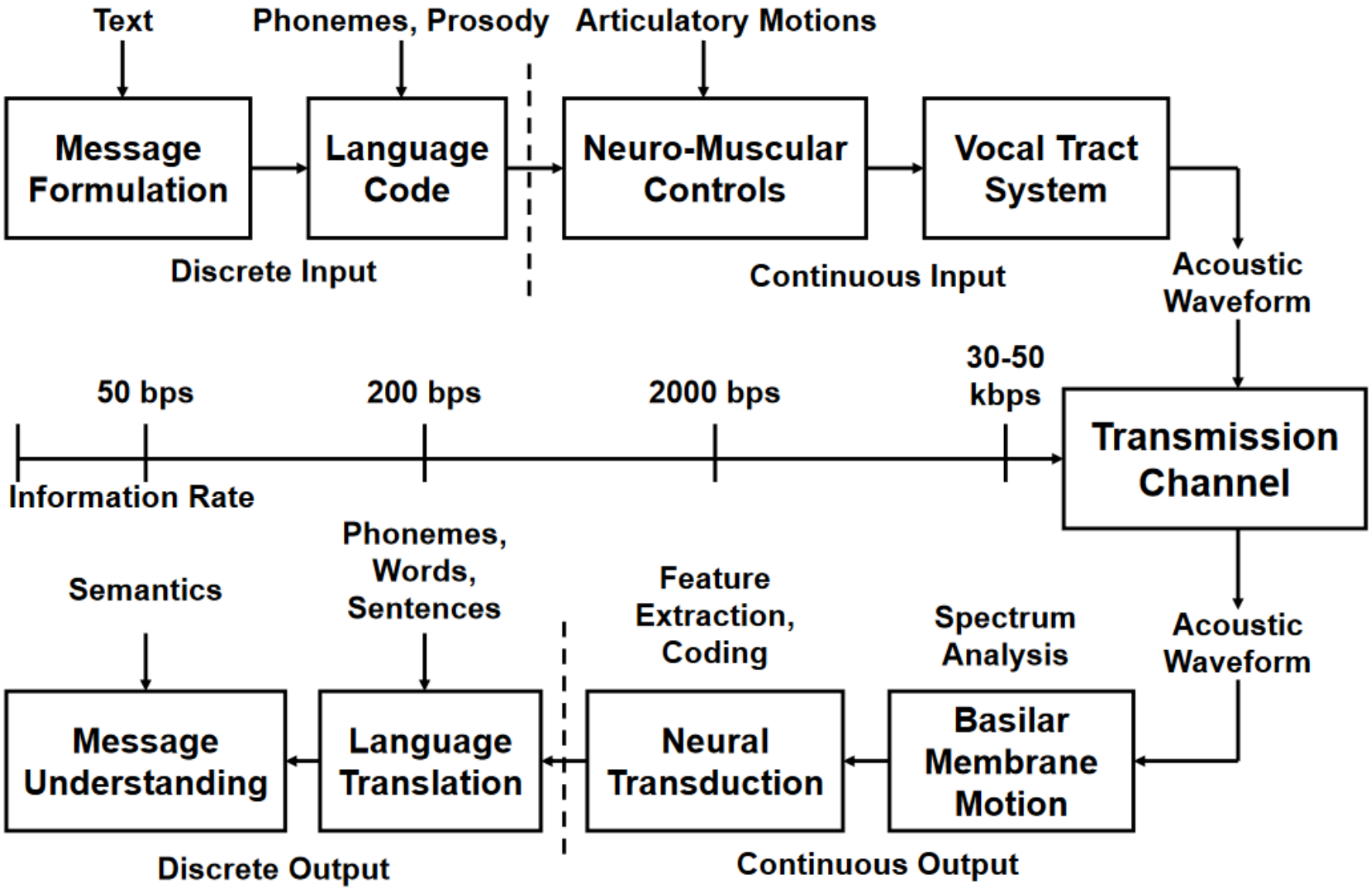
Abdul Rehman

Affective Computing Lab Group

Speech Recognitions Steps



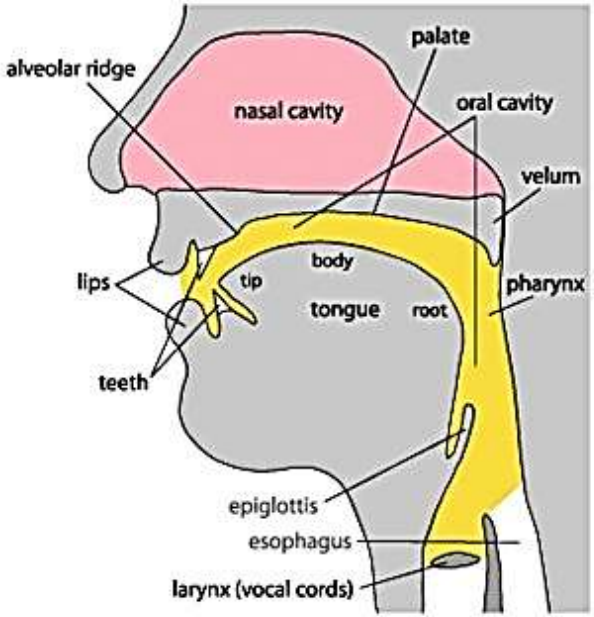
Speech Transmission

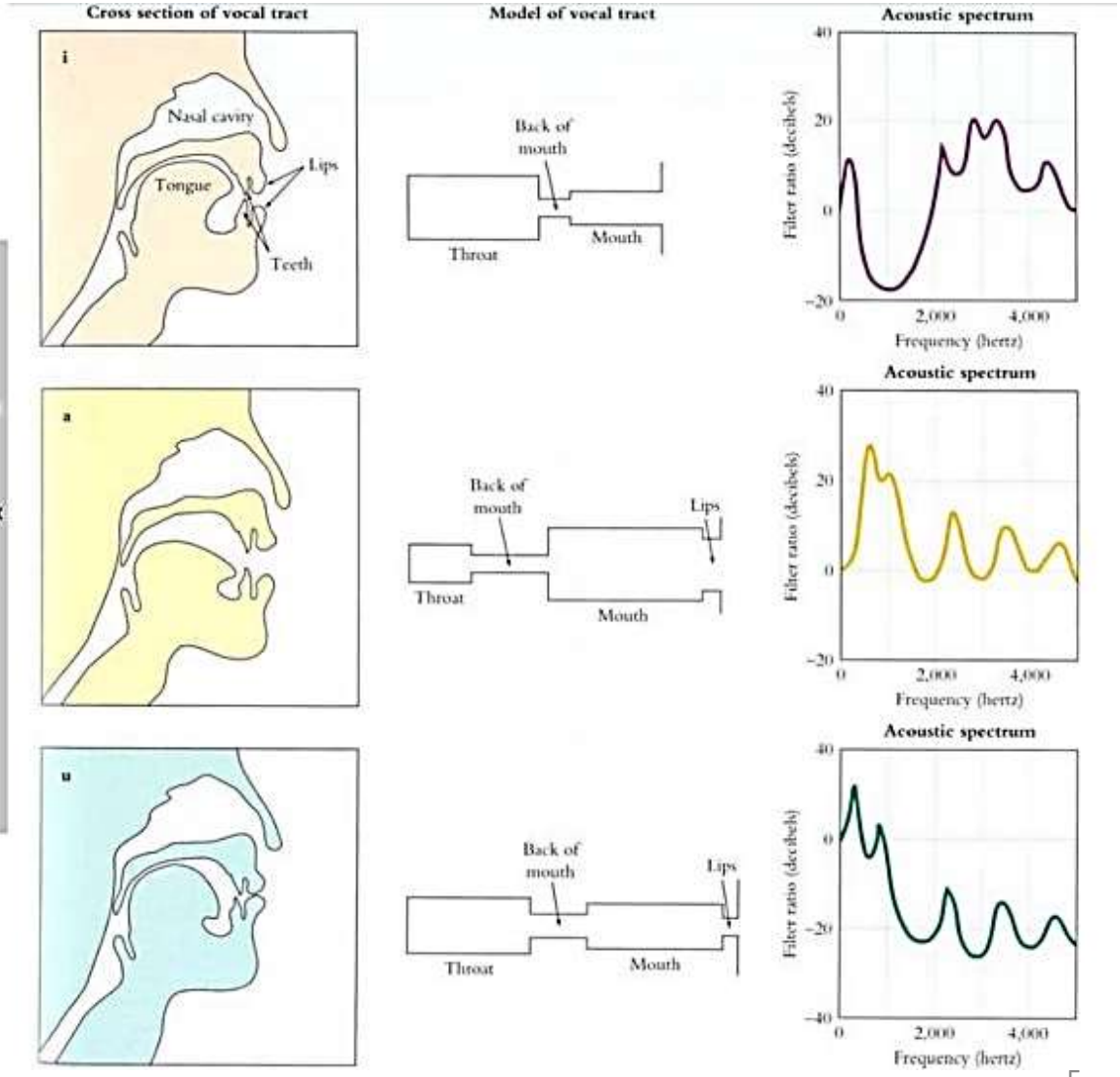


Holistic mediums from speaker to listeners

Layers	Sublayers	Opacity	Uncertainty	Possibilities
Speaker	Subject's subjective Consciou.	100%	Infinite	A kid or an old lady , or a goat… Infinite
	Character	10%	Indefinite	Spectrum of persona
	External Stimuli	20%	Indefinite	Affective environment
	Active Emotion	40%	Small	Happy, Sad, Angry
	Lexicon	40%	Infinite	
	Intensity-Direction	20%	Small	e.g., Distance, volume, angle
Intermediate	Environment	5%	Indefinite	Noise, weather, etc.
	Language Barrier	0%	Small	e.g., accent mismatch
Listener	Listener-Env Barrier	10%	Indefinite	Affective environment
	Sensitivity	10%	Small	Audibility
	Biases	10%	Indefinite	e.g., Prejudices
	Perceiver's subjective Conscio.	100%	Infinite	A kid or an old lady … Infinite

Vocal Tract 人声道

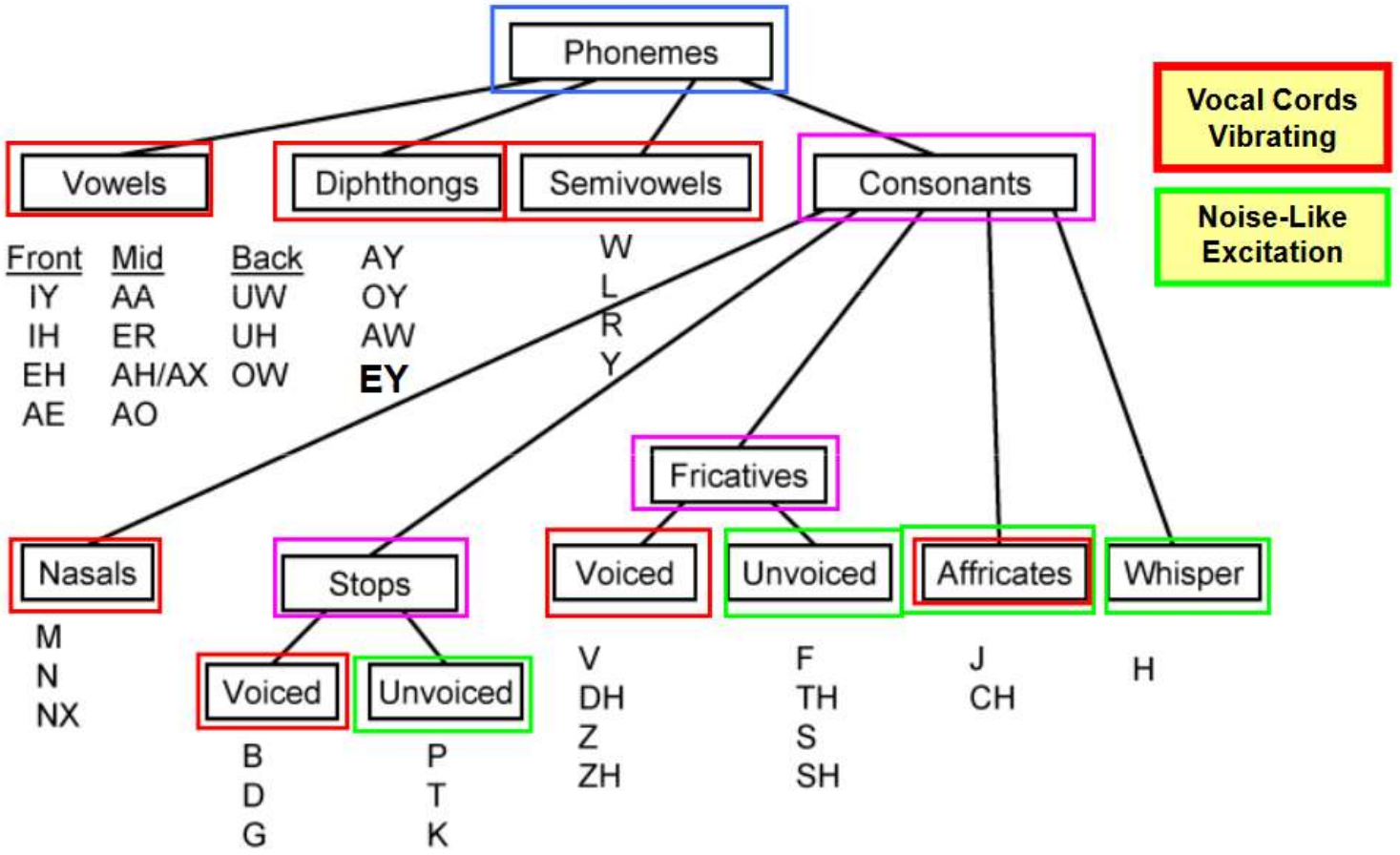




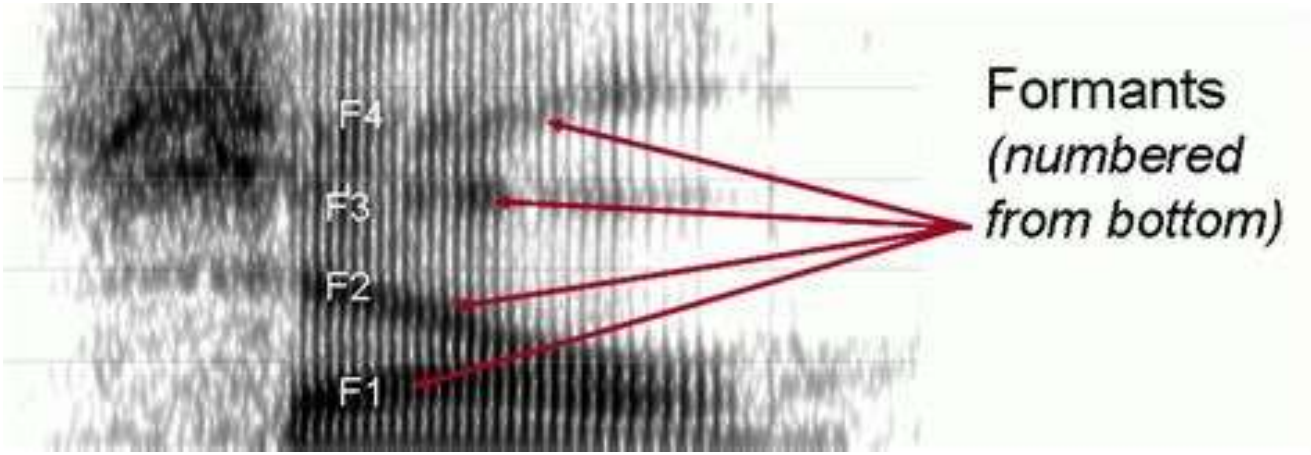
Parts of Speech

phoneme is an indivisible unit of sound in a given language a phoneme is an abstraction of the physical speech sounds (phones) and may encompass several different phones

Phoneme Classification Chart
音素分类

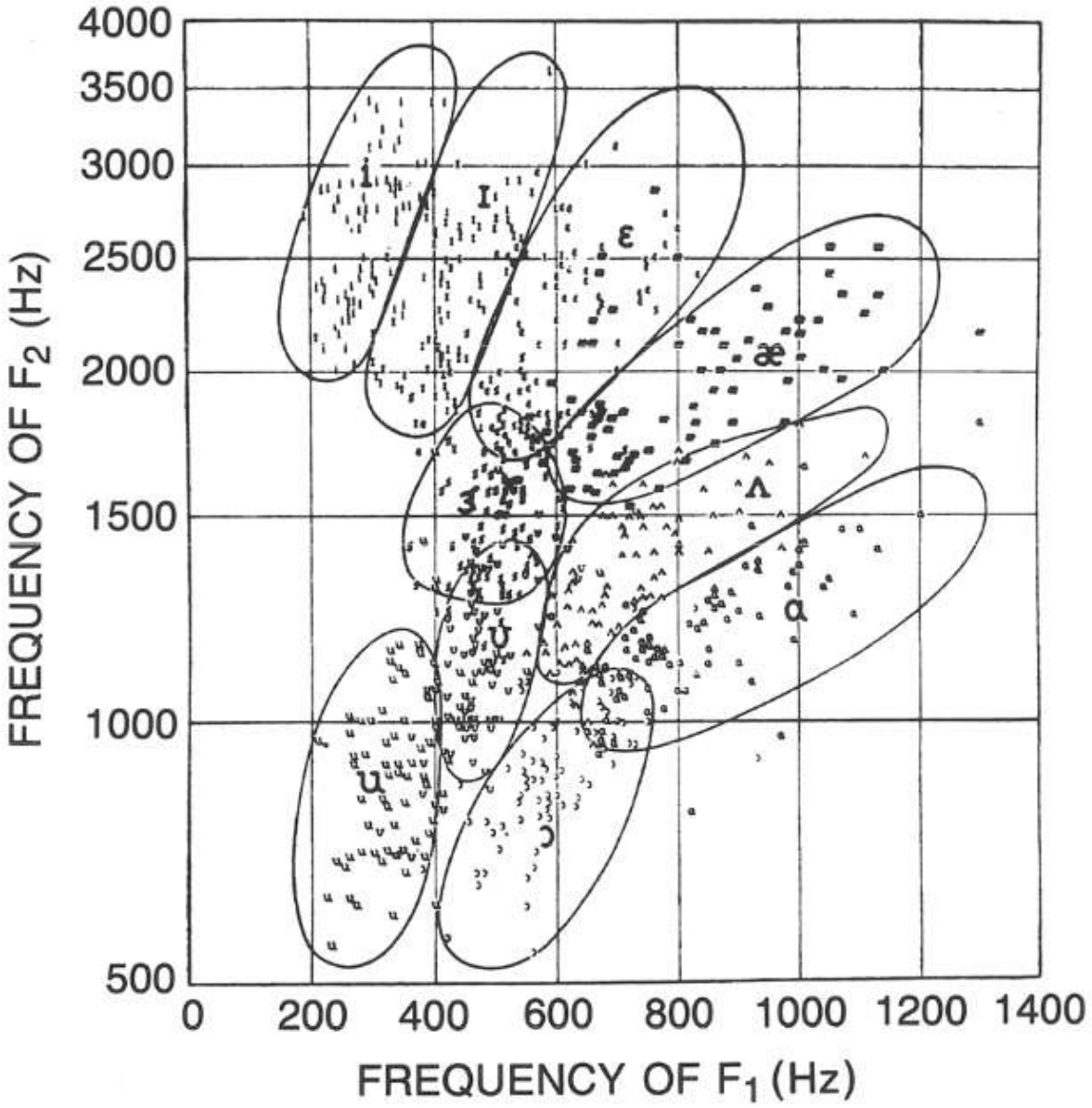


Formants 共振峰



Phonetic Symbol	Example Word	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
/ow/	bought	570	840	2410
/oo/	boot	300	870	2240
/u/	foot	440	1020	2240
/a/	hot	730	1090	2440
/uh/	but	520	1190	2390
/er/	bird	490	1350	1690
/ae/	bat	660	1720	2410
/e/	bet	530	1840	2480
/i/	bit	390	1990	2550
/iy/	beet	270	2290	3010

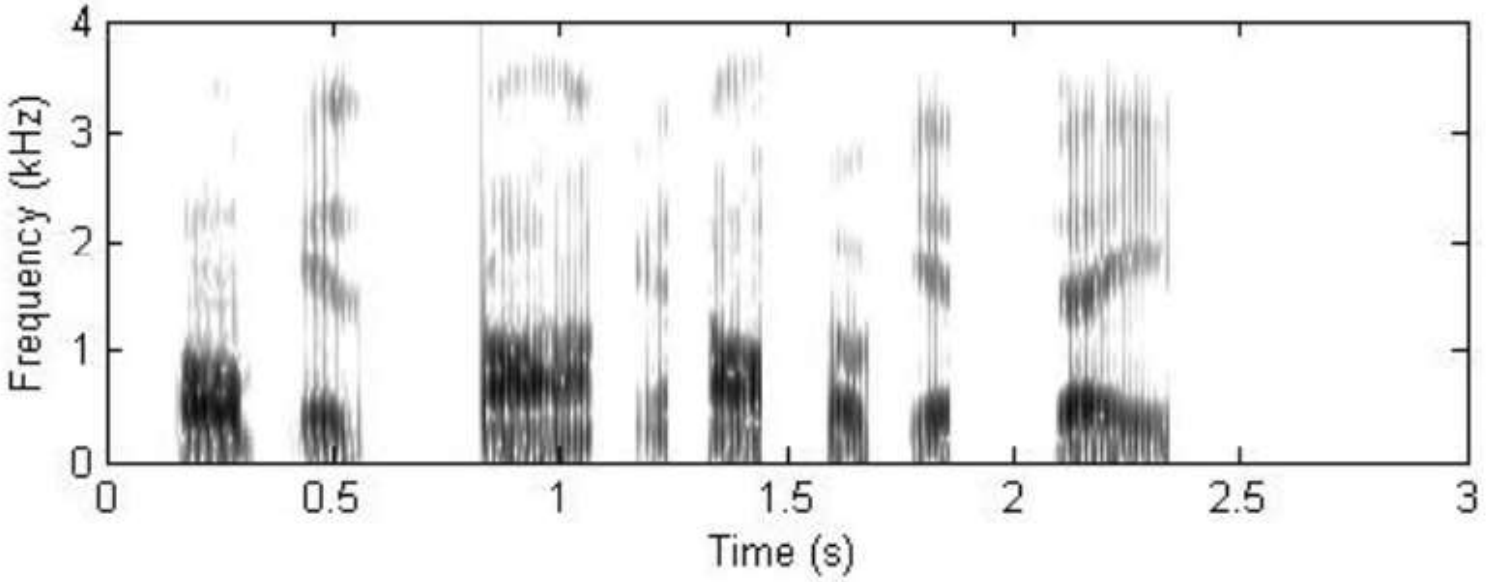
Vowel Formants 元音共振峰

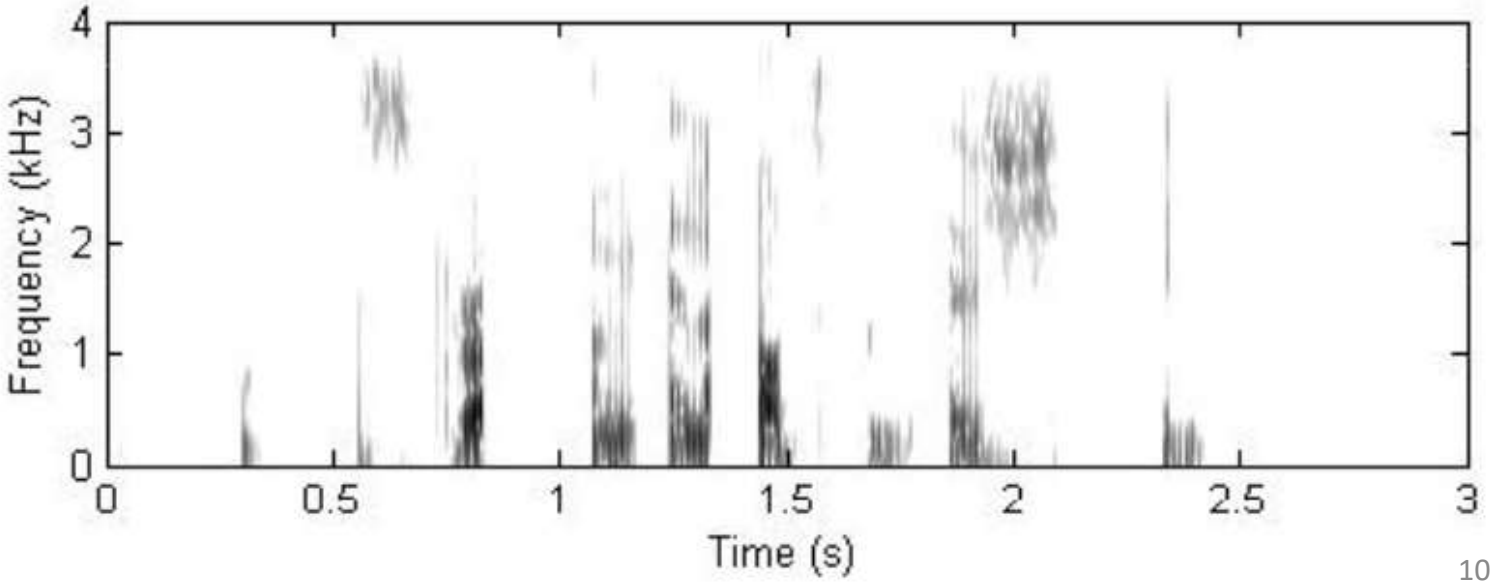


Clear pattern of variability of vowel pronunciation among men, women and children

Strong overlap for different vowel sounds by different talkers => no unique identification of vowel strictly from resonances => need context to define vowel sound

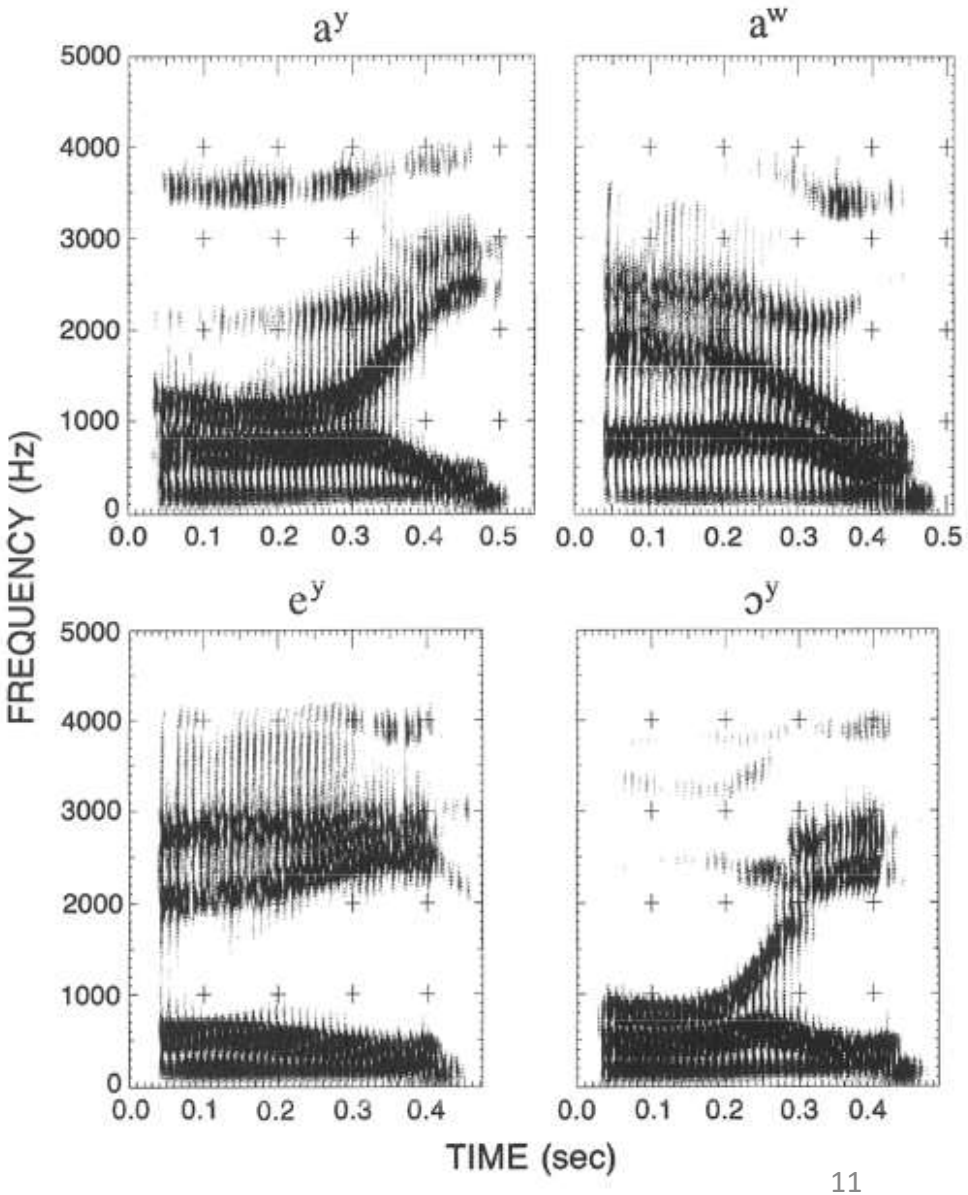
Separating Vowels & Consonants



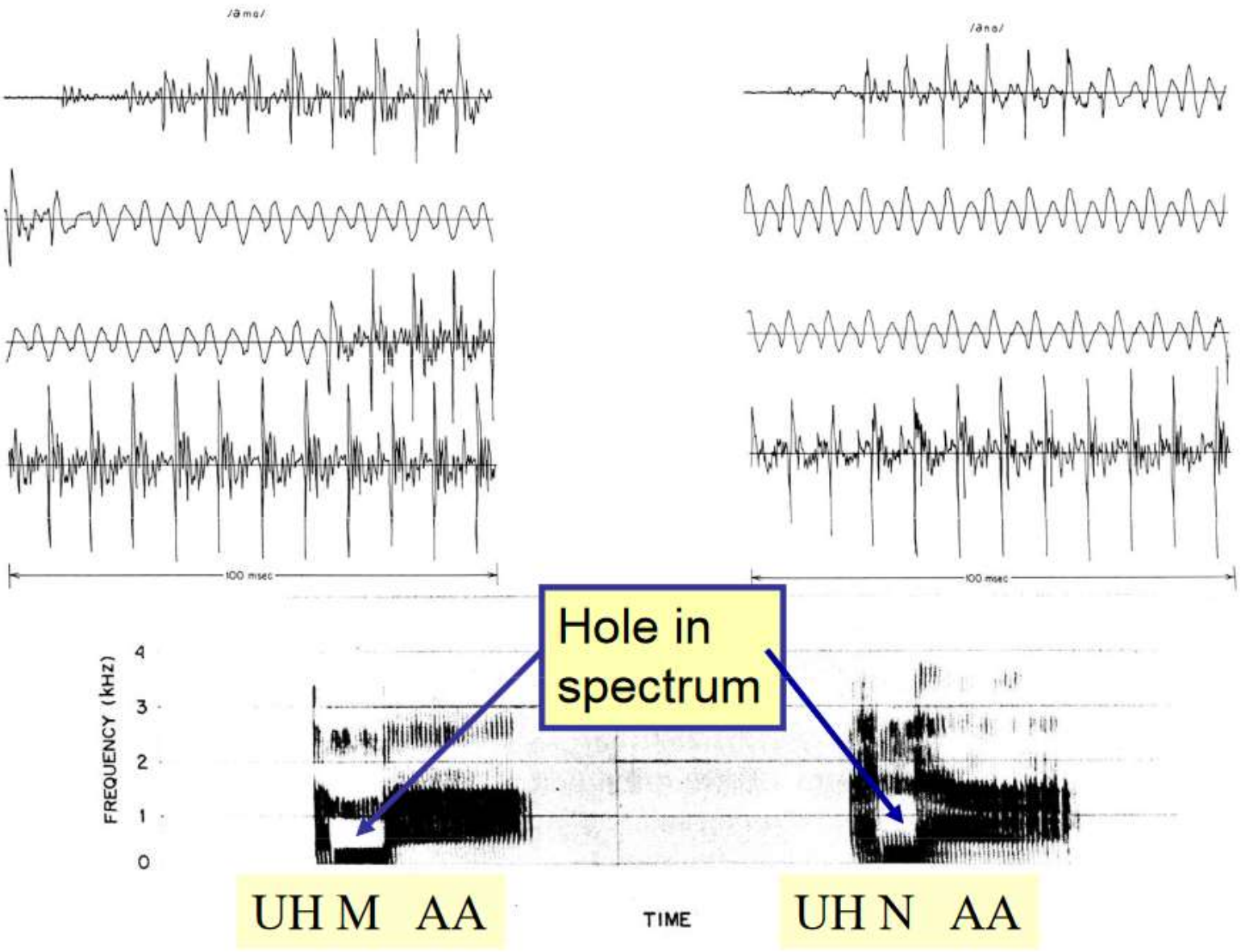


Diphthongs 白喉

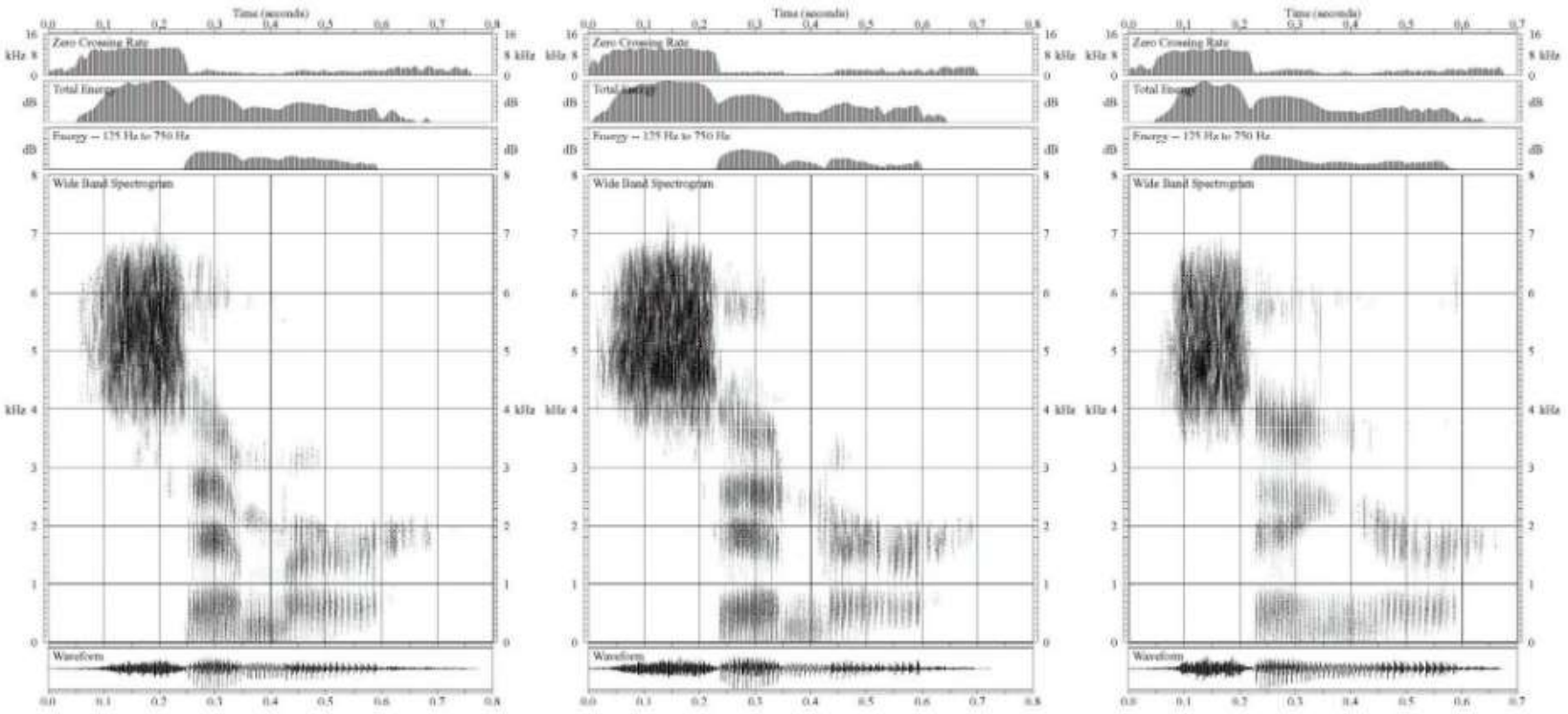
- Gliding speech sound that starts at or near the articulatory position for one vowel and moves to or toward the position for another vowel
  - /AY/ in buy
  - /AW/ in down
  - /EY/ in bait
  - /OY/ in boy
  - /OW/ in boat (usually classified as vowel, not diphthong)
  - /Y/ in you (usually classified as glide)



Nasal Sounds 鼻音



Nasal Spectrograms

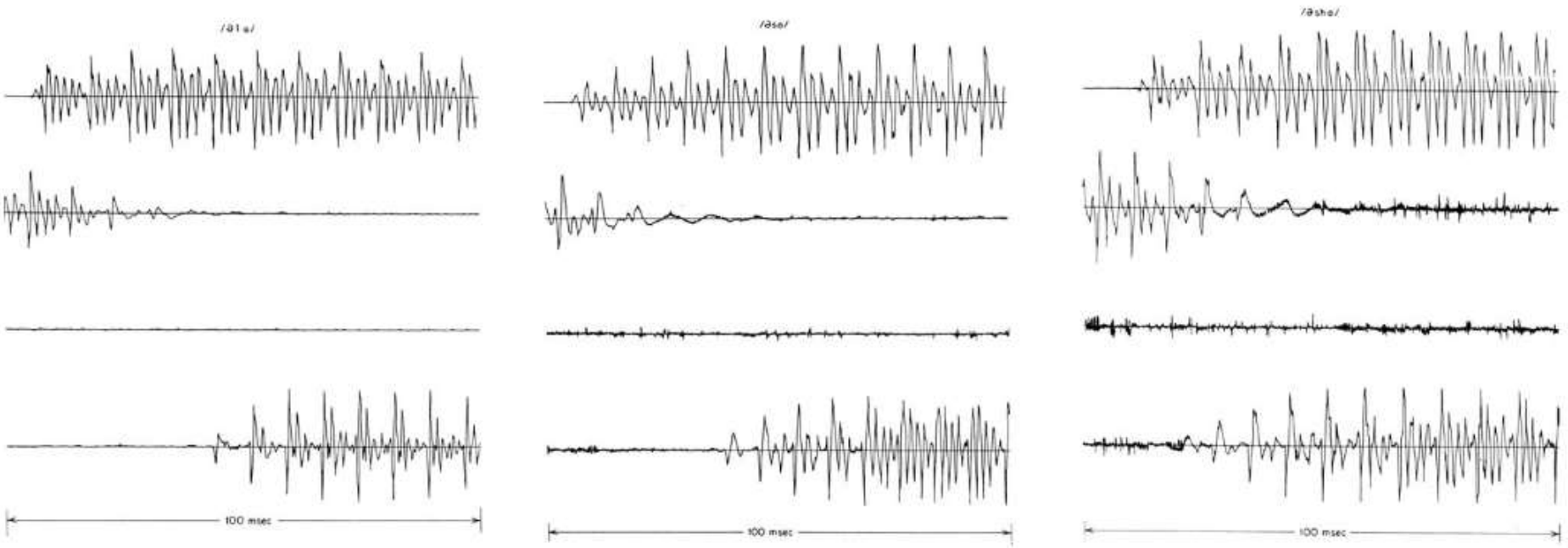


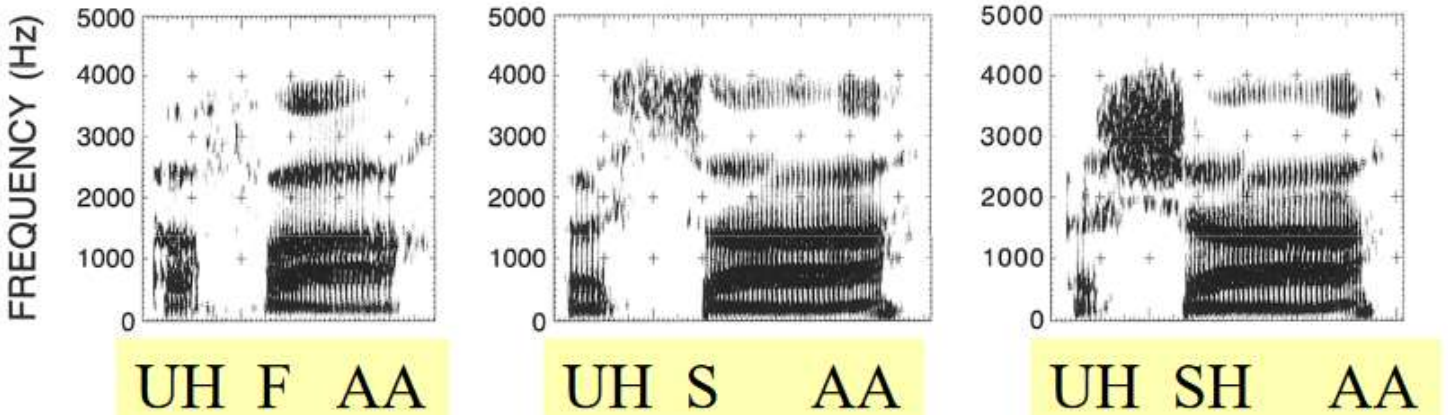
simmer
/sɪmɝ/

sinner
/sɪnɝ/

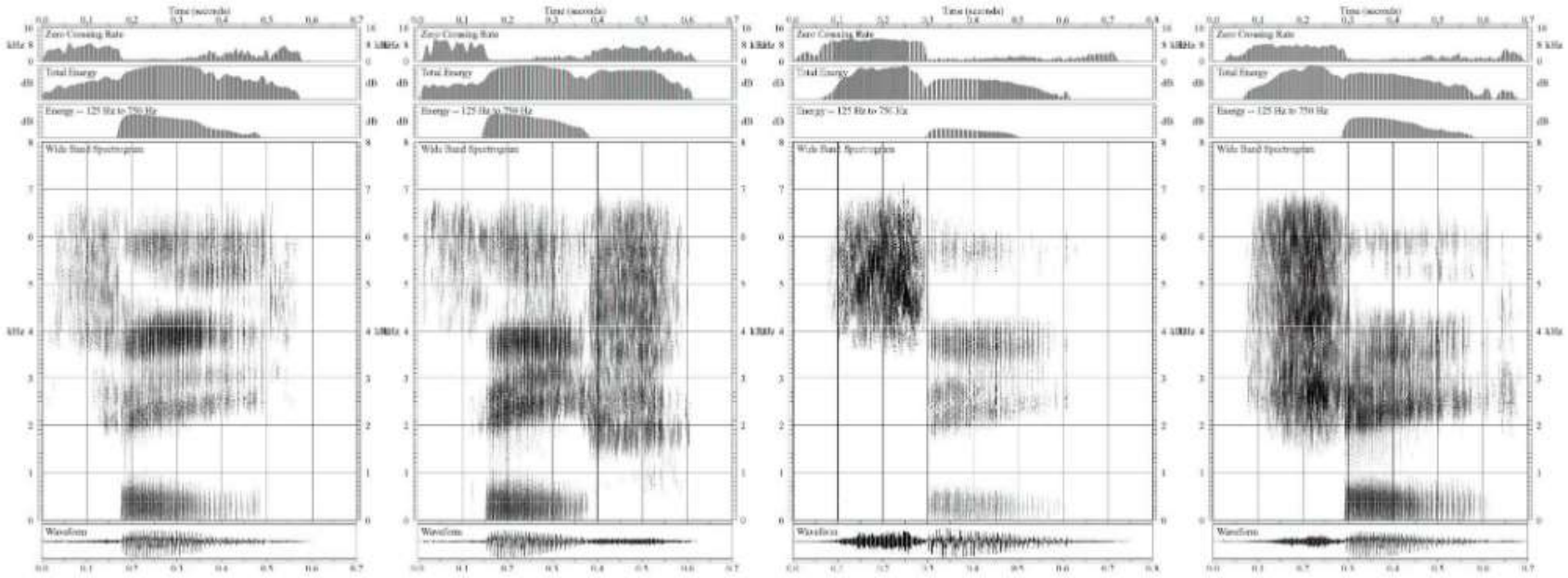
singer
/sɪŋɝ/

Unvoiced Fricatives 清音摩擦



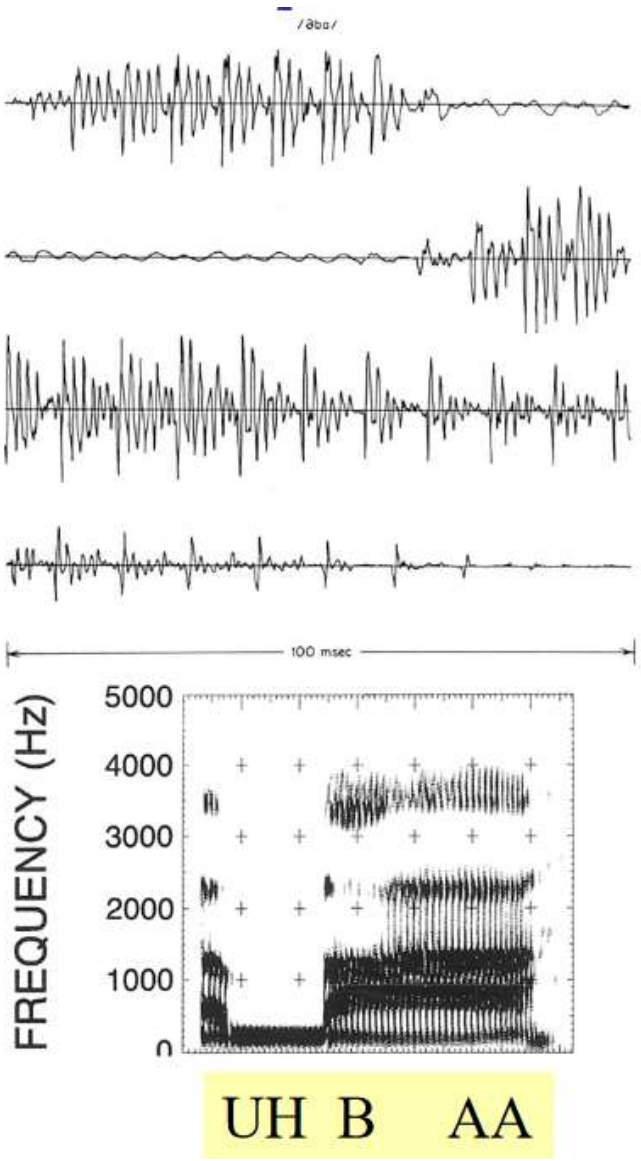


Unvoiced Fricatives Spectrograms

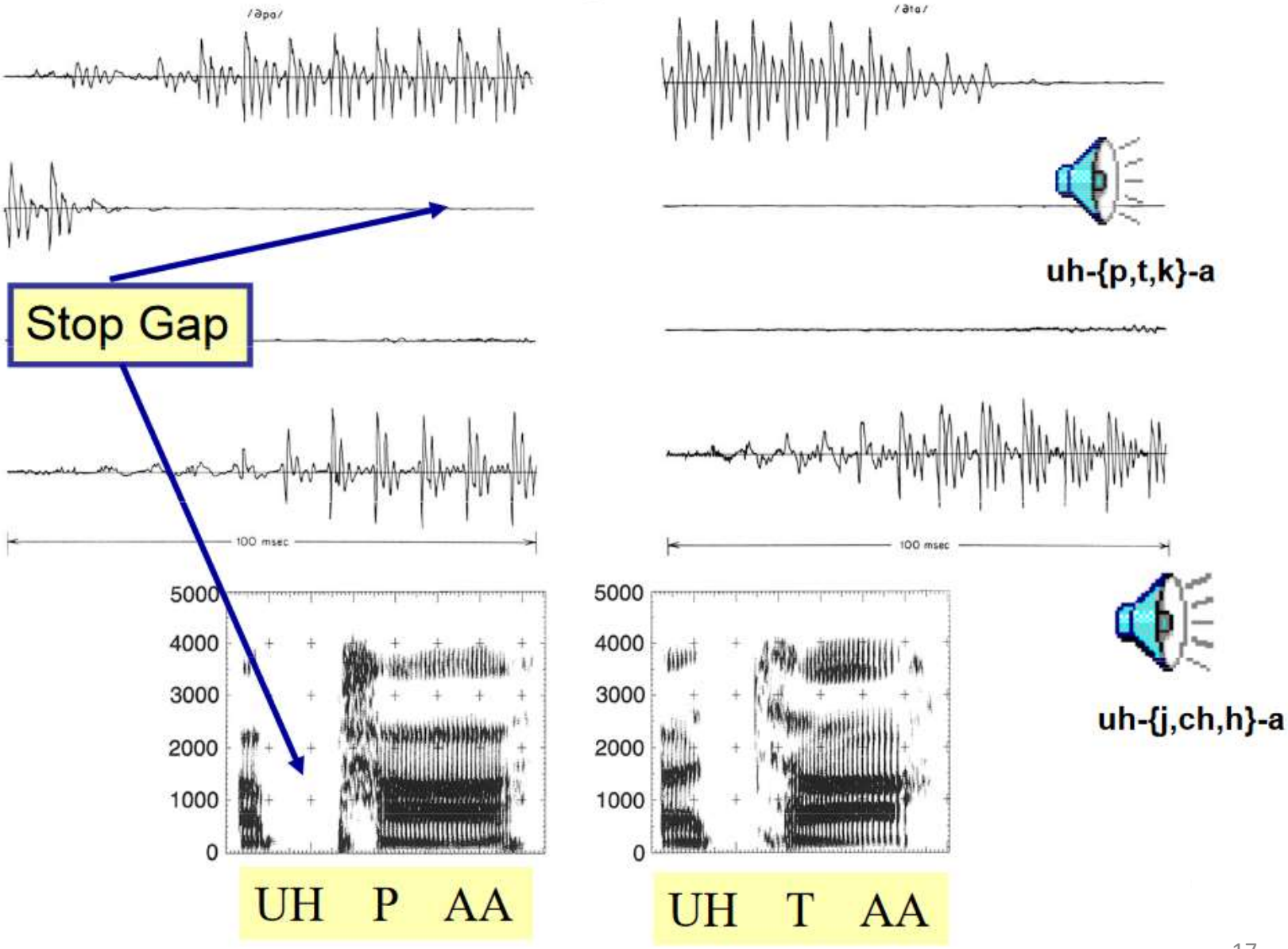


fee /fiʸ/ thief /θiʸf/ see /siʸ/ she /šiʸ/

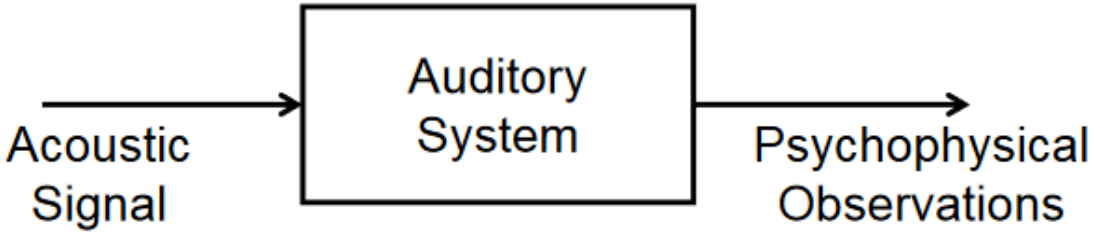
Voiced Stop Consonant



Unvoiced Stop Consonants



Human Auditory System

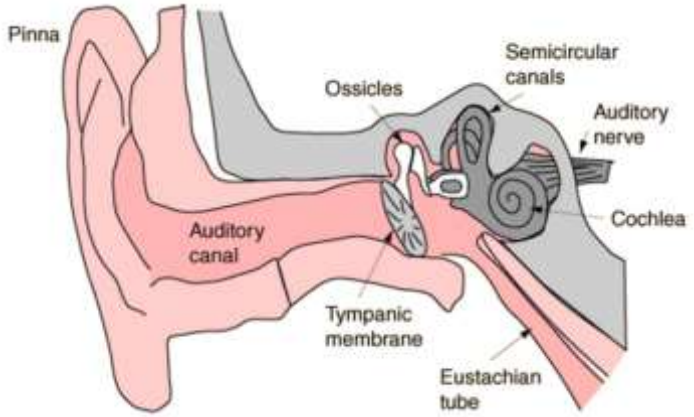


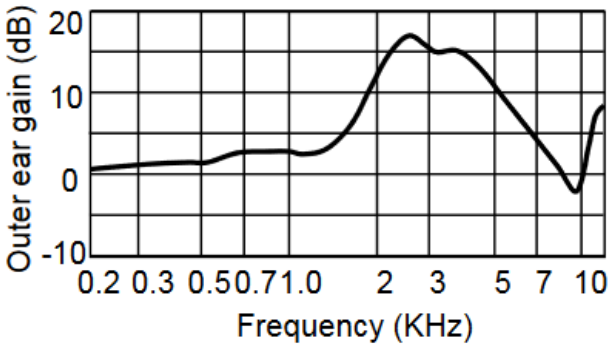
Hearing and perception
听力与知觉

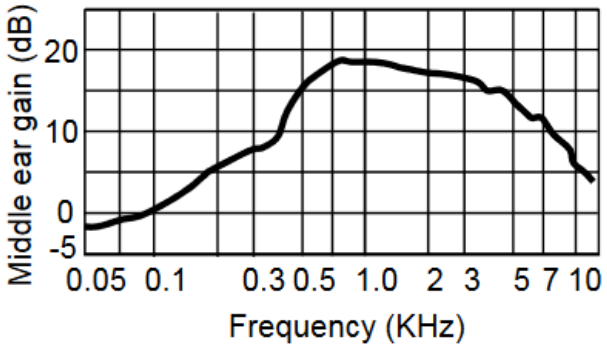
- Correspondences between sound intensity and loudness, and between frequency and pitch are complicated and far from linear.

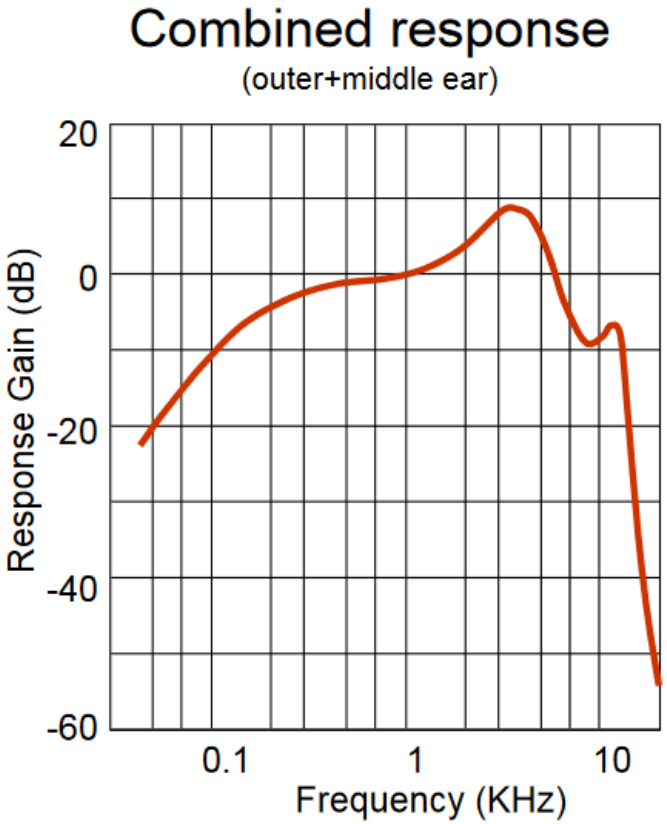
- Attempts to extrapolate from psychophysical measurements to the processes of speech perception and language understanding are, at best, highly susceptible to misunderstanding of exactly what is going on in the brain

Transfer function for Ear

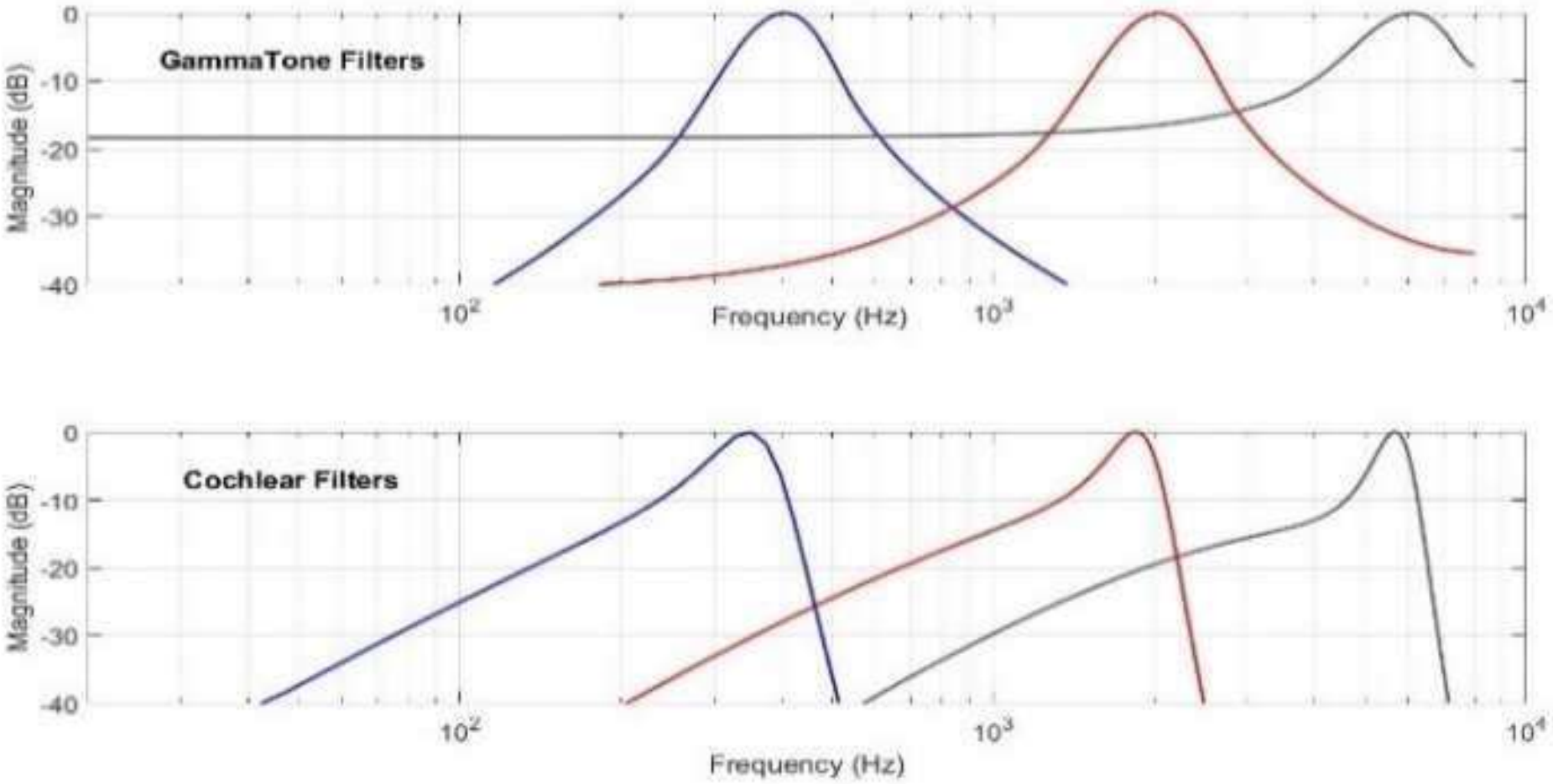




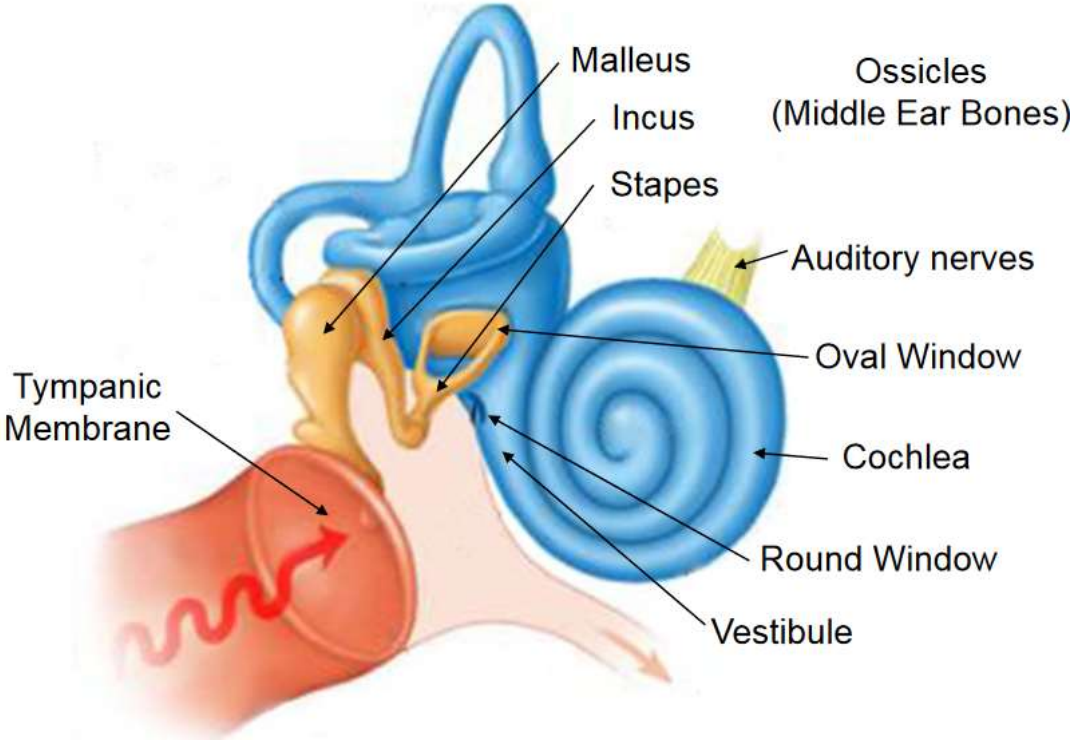




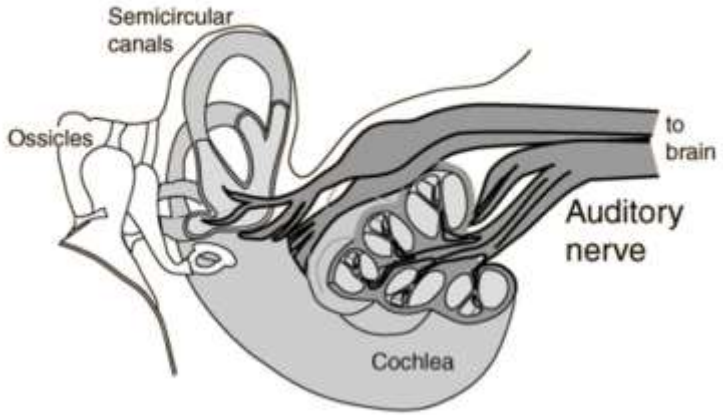
Filters that match Human Ear



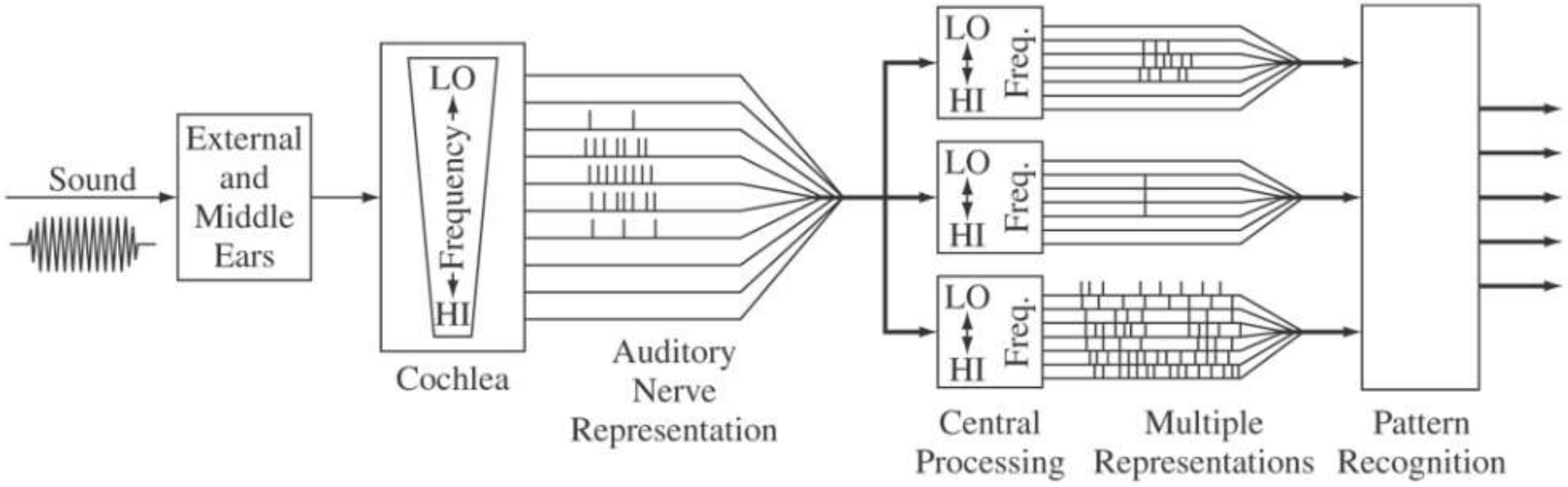
The Cochlea



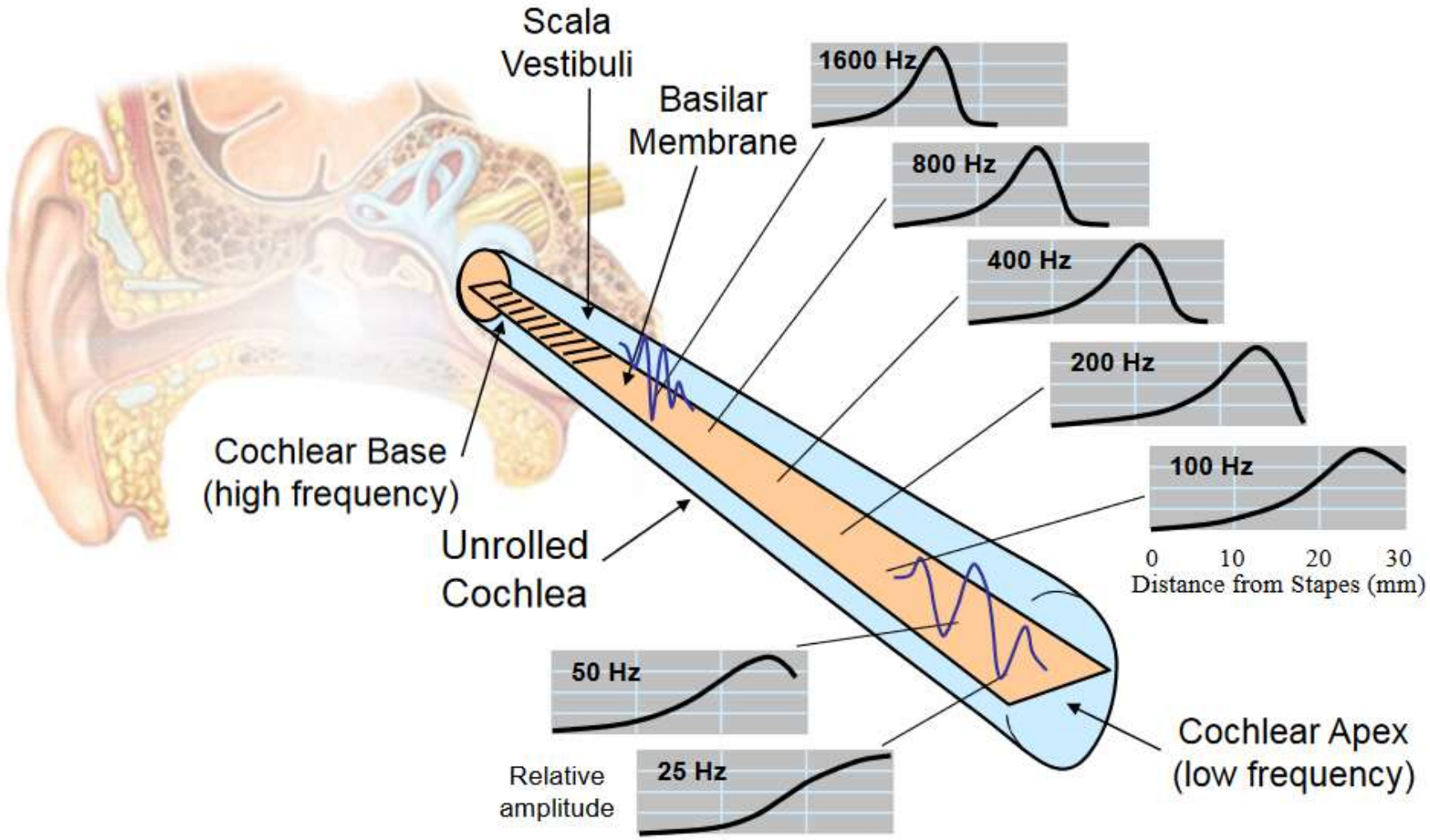
The Auditory Nerve



Abstract Ear 抽象的耳朵

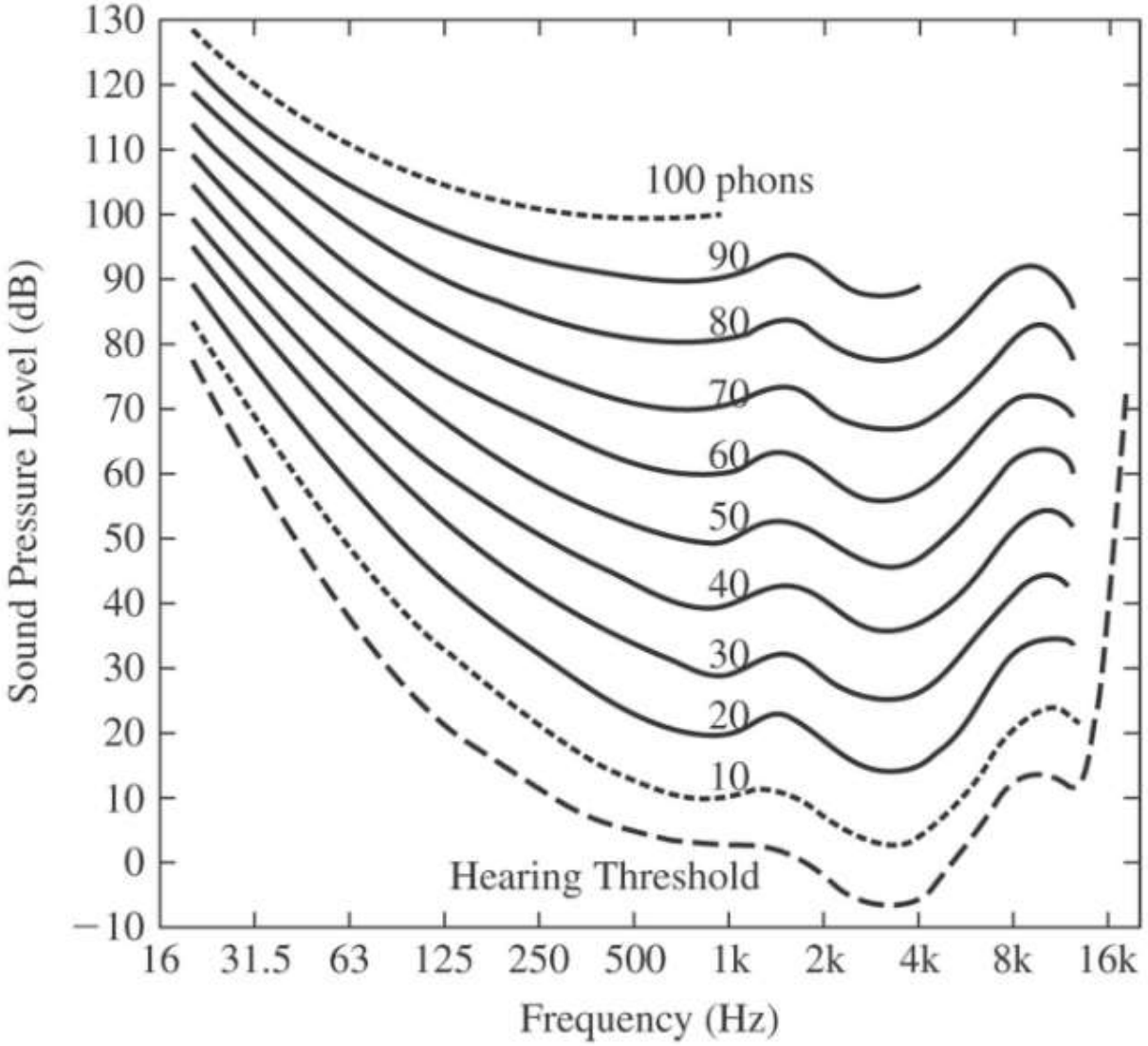


Cochlea and Basilar Membrane



Loudness Level

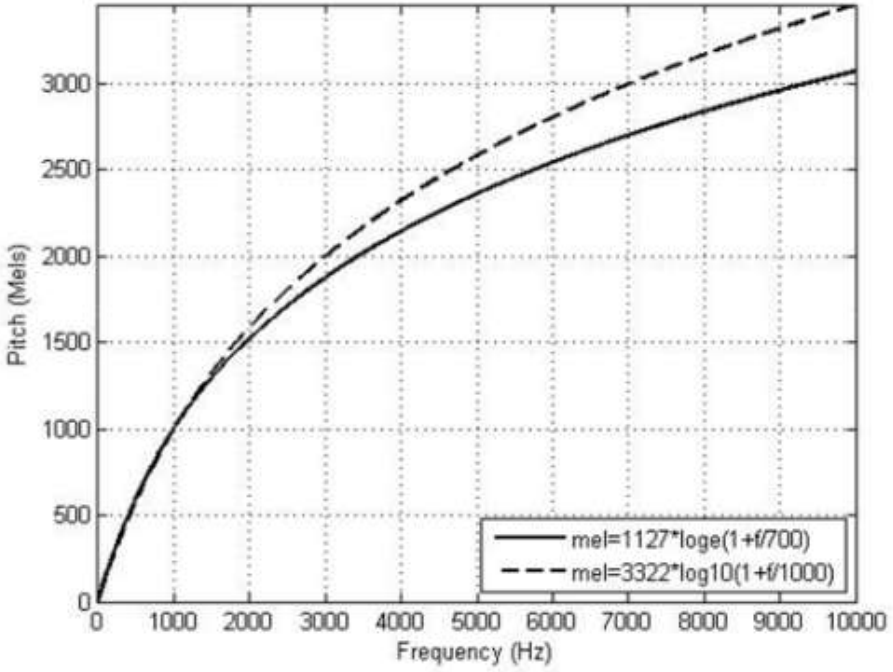
- ***Loudness Level (LL)*** is equal to the *IL* of a 1000 Hz tone that is judged by the average observer to be equally loud as the tone

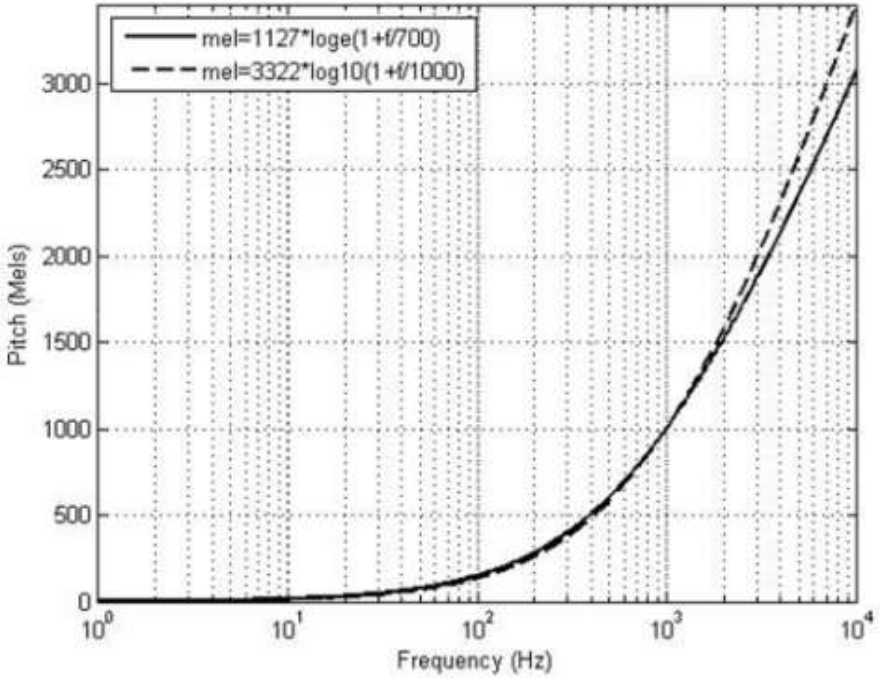


Pitch perception

- Pitch and fundamental frequency are not the same thing.
- Pitch is a perceived quantity while frequency is a physical one.
- Relationship between pitch and fundamental frequency is not simple, even for pure tones.

Mel-Scale for Pitch



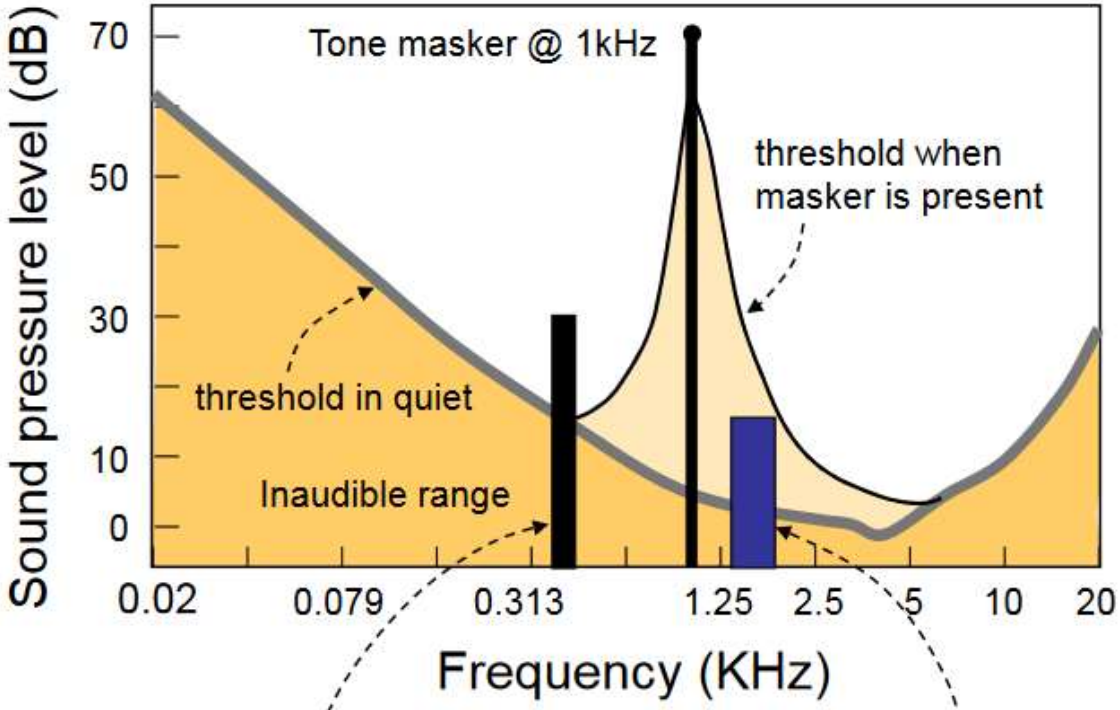


$$\text{Pitch } (mels) = 3322 \log_{10}(1 + f/1000)$$

Alternatively, we can approximate curve as:

$$\text{Pitch } (mels) = 1127 \log_{e}(1 + f/700)$$

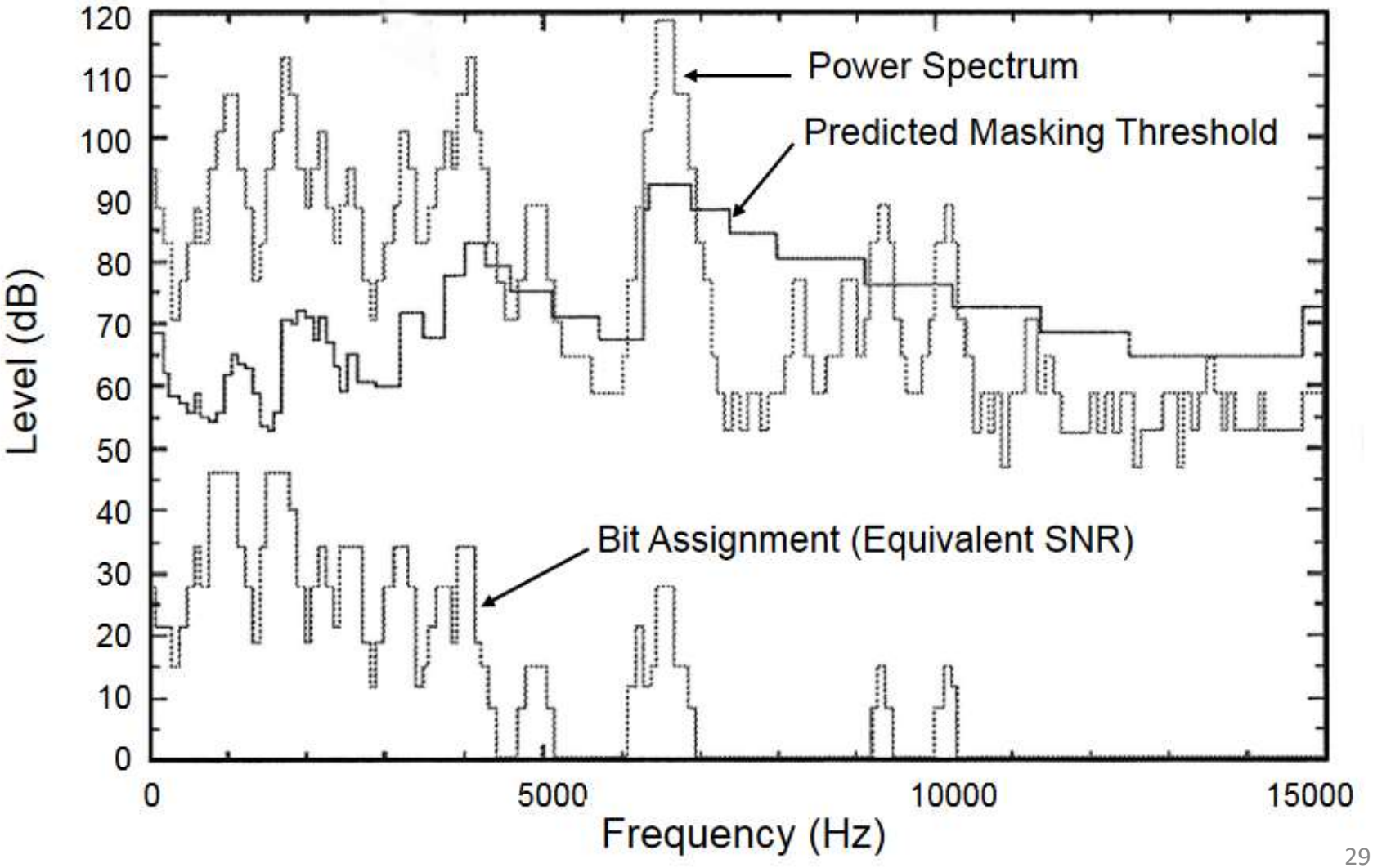
Auditory Masking 听觉掩蔽



Signal perceptible even in the presence of the tone masker

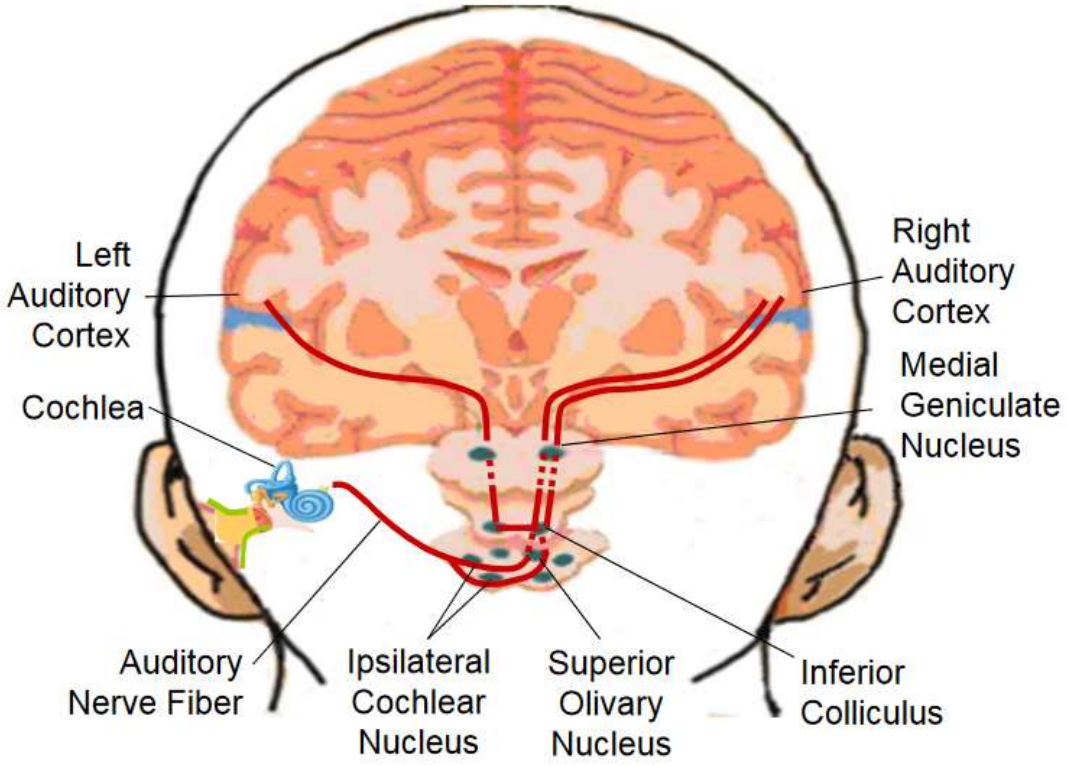
Signal not perceptible due to the presence of the tone masker

Using Masking to Separate Noise



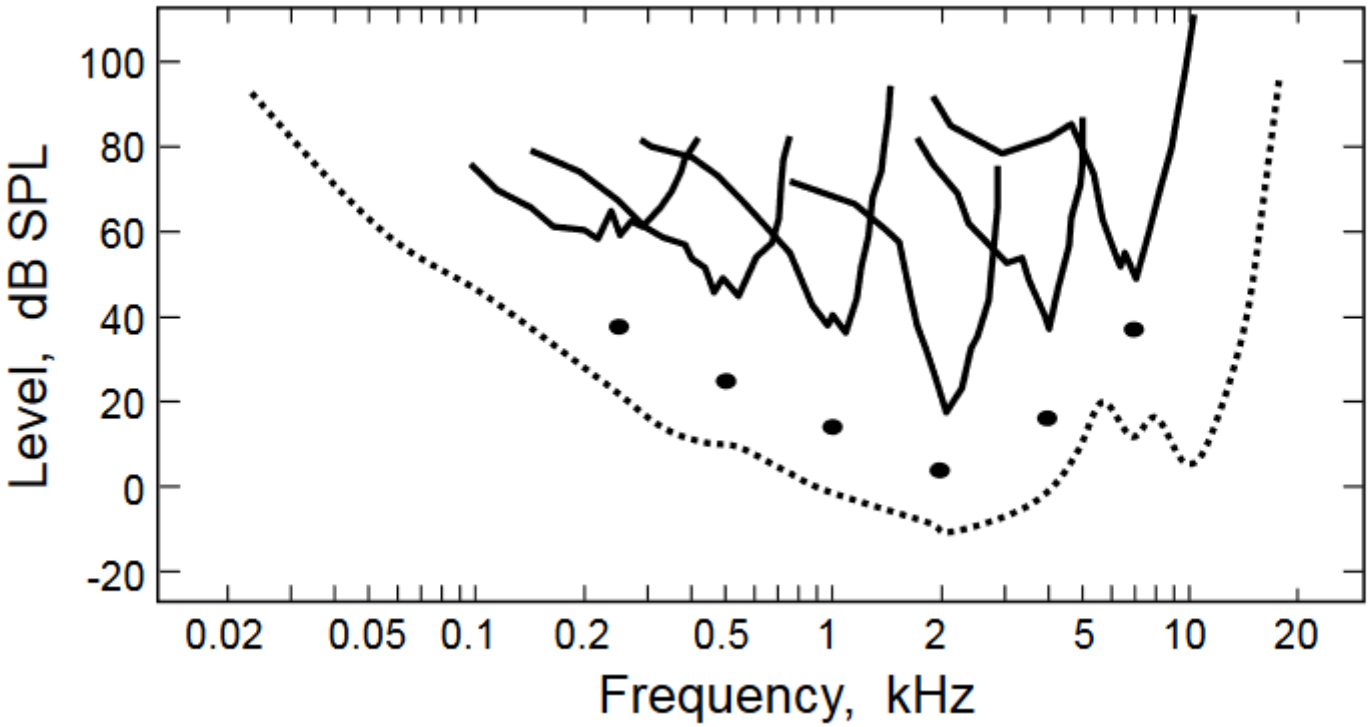
Psychophysical Tuning
心理物理调优

- Tuning curves of the auditory nerve fibers

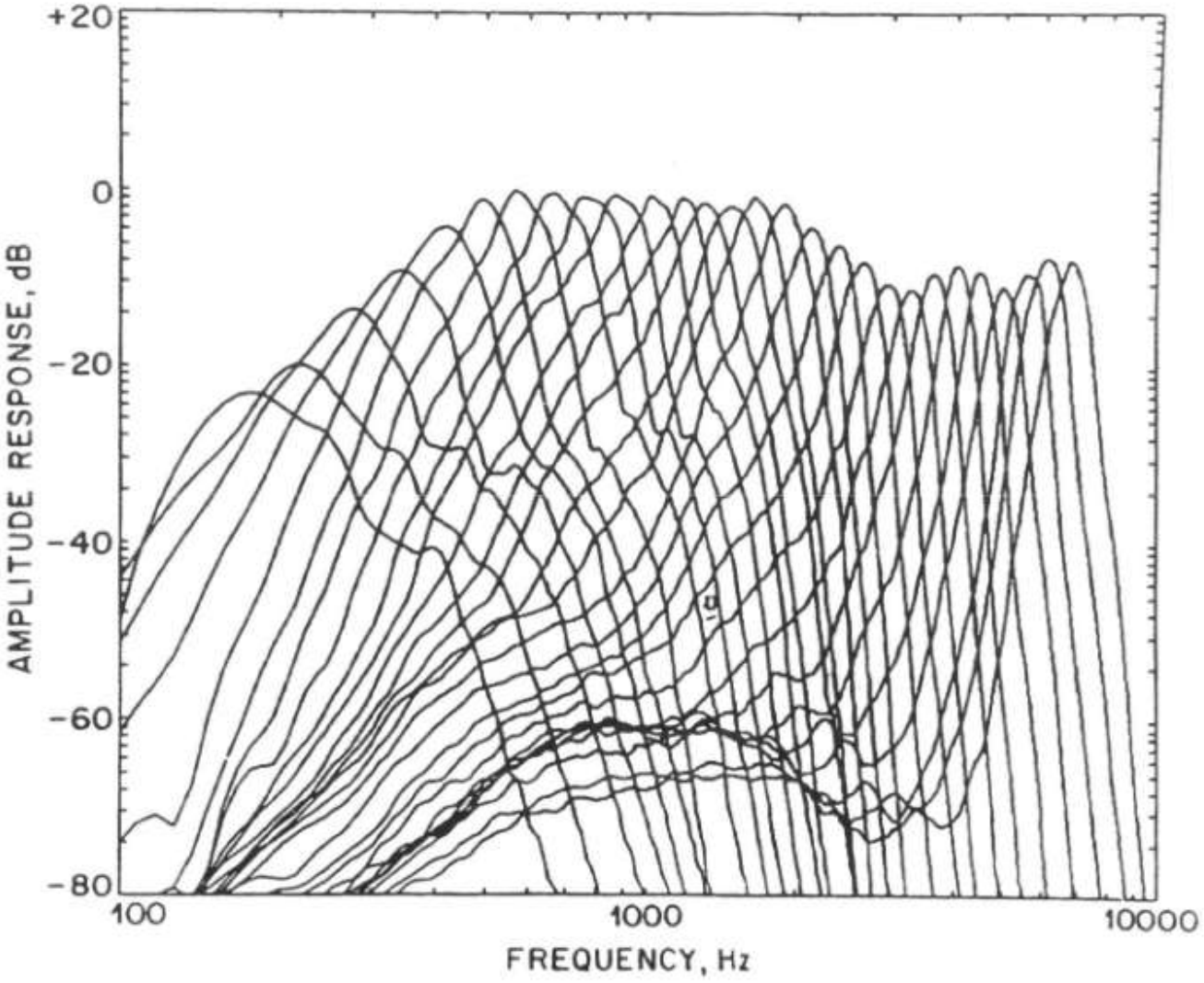


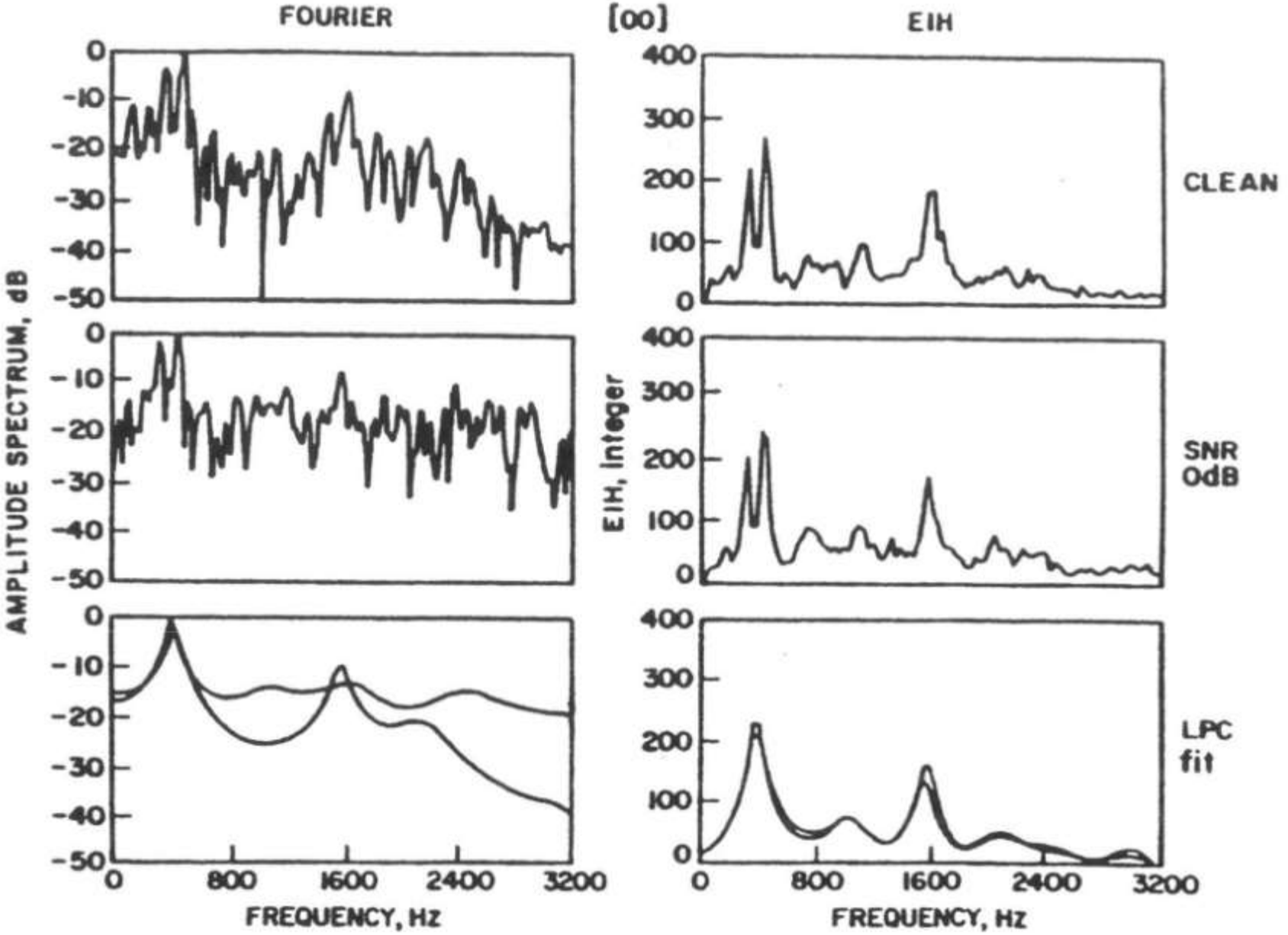
Psychophysical Tuning Curves (PTC)

- Masking for calibrating the auditory nerve response

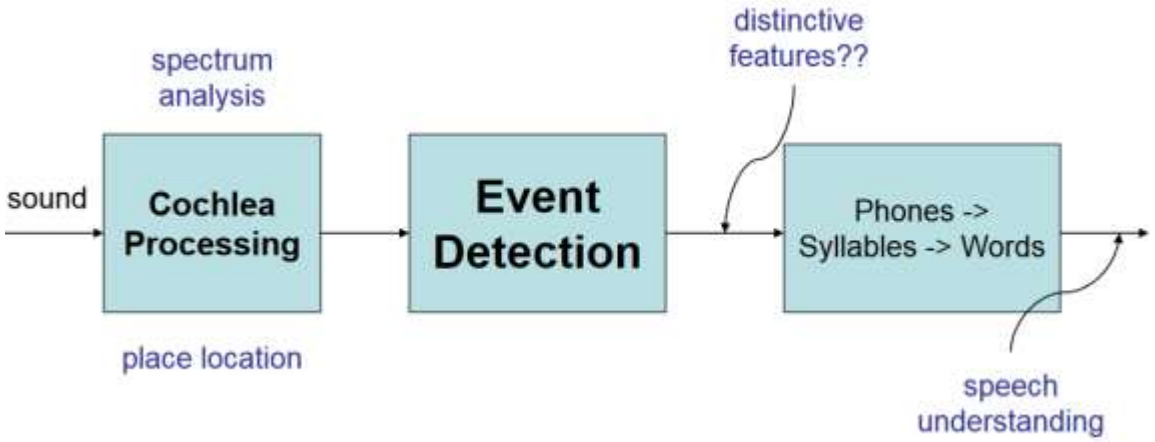


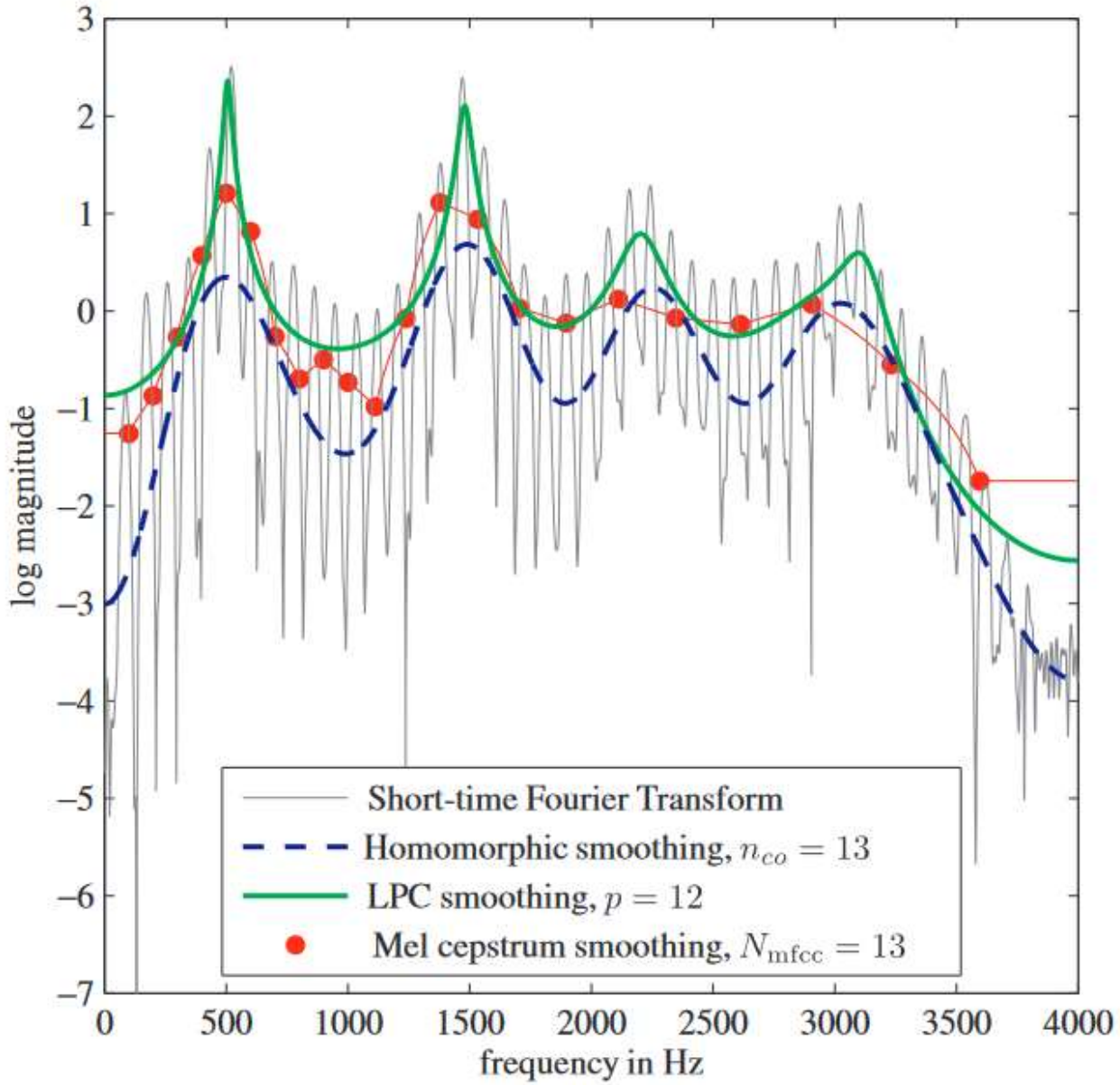
Cochlear Filter Design
耳蜗式过滤器设计




FOURIER
[00]
EIH
AMPLITUDE SPECTRUM, dB
EIH, Integer
0
-10
-20
-30
-40
-50
400
300
200
100
0
CLEAN
SNR
0dB
LPC
fit
0
800
1600
2400
3200
FREQUENCY, Hz

Spectral Analysis

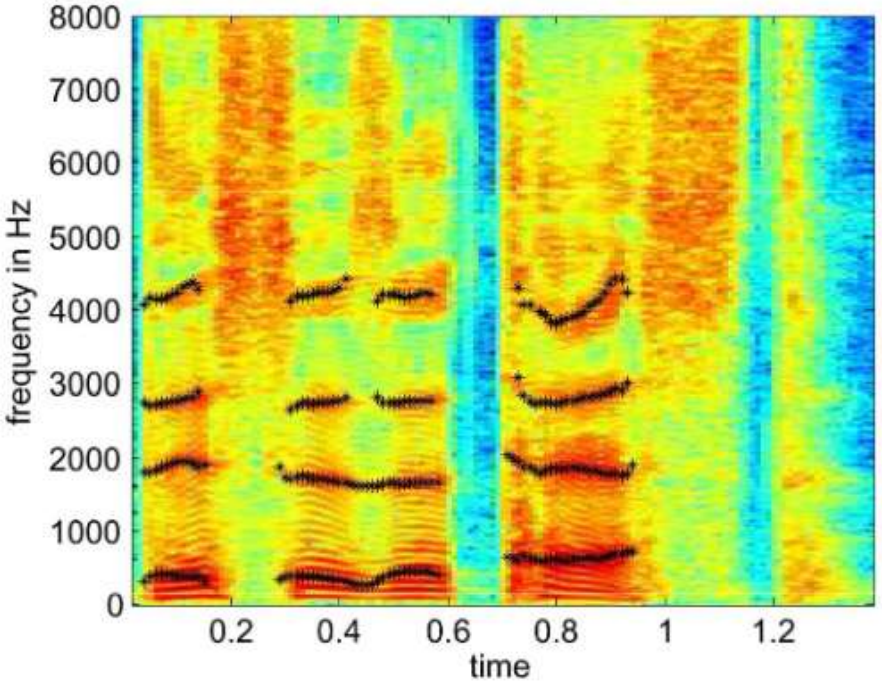


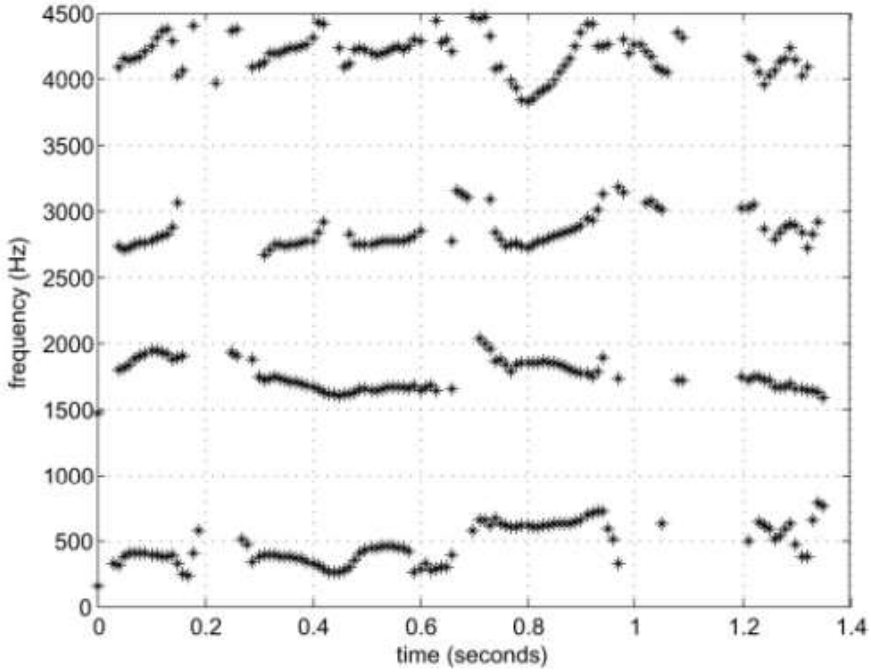


Comparison of spectral smoothing methods.

Compression by LPC

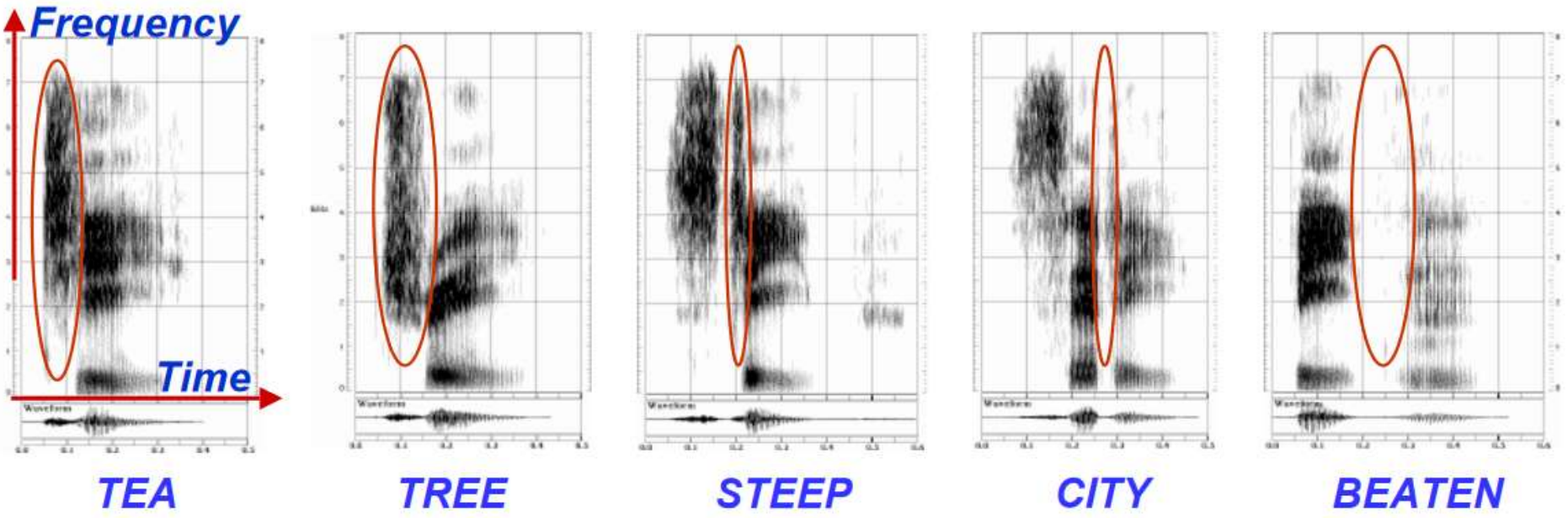
- Linear Predictive Coefficient
- 5--8 polynomials



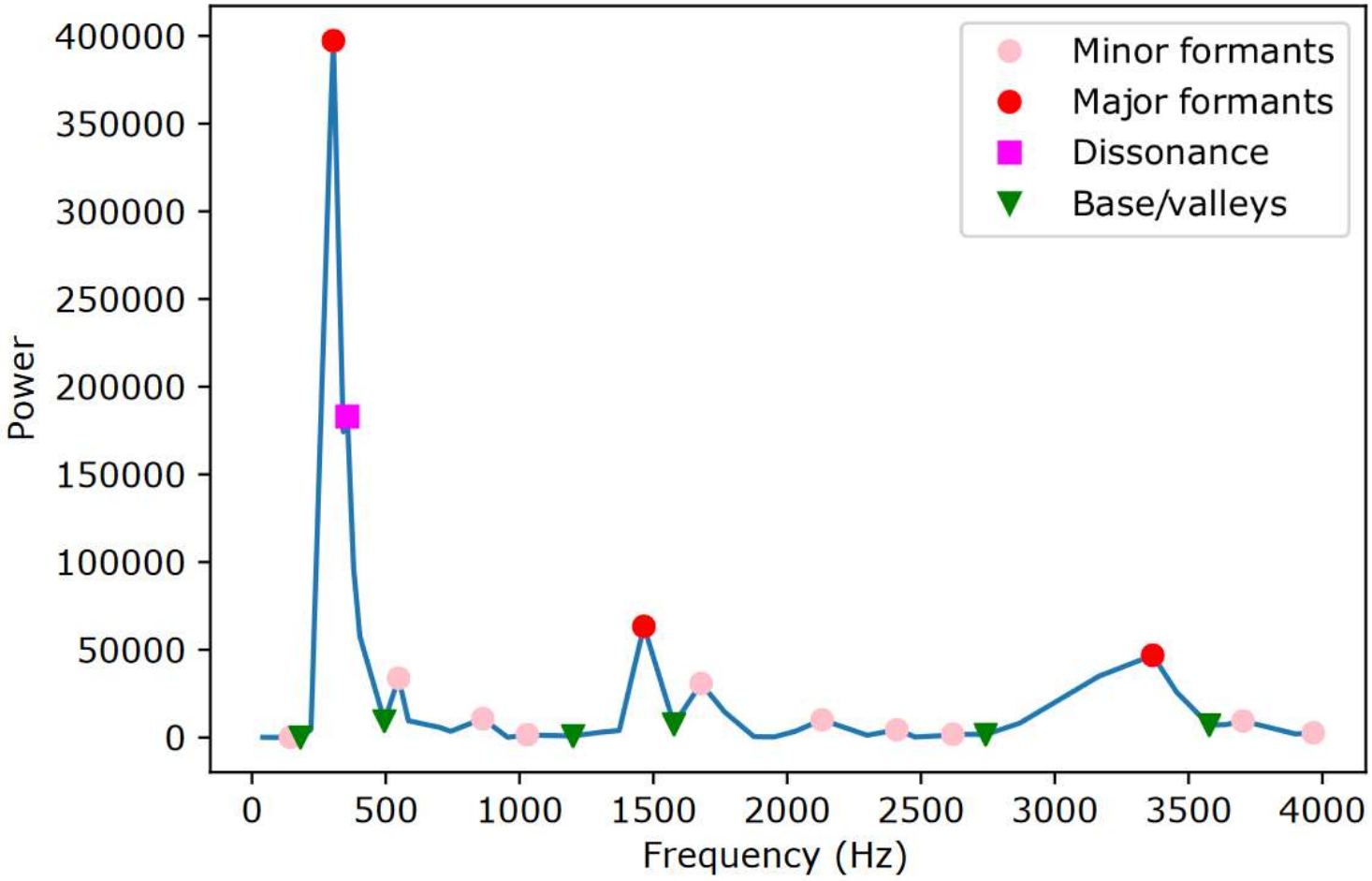


Spectral Context 语境

- The acoustic realization of a phoneme depends strongly on the context in which it occurs



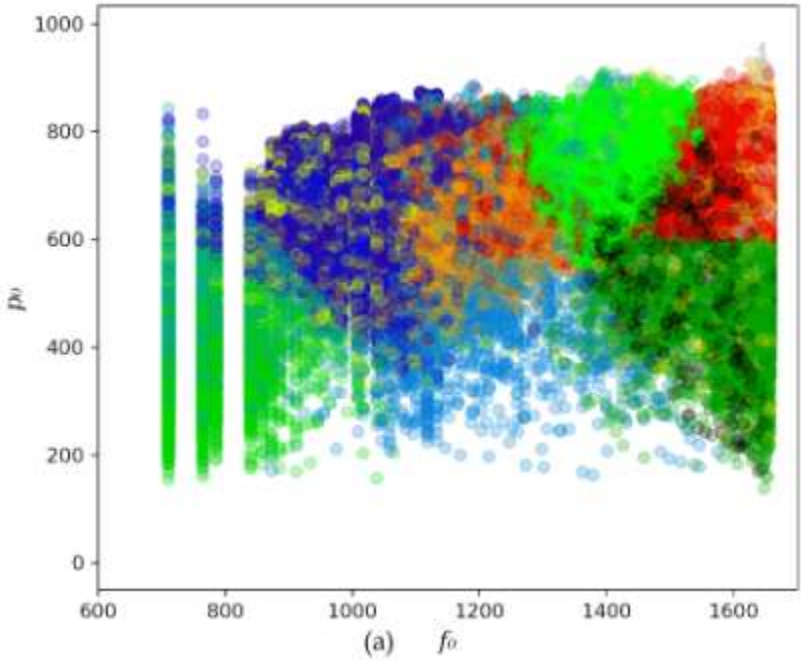
Compression into 12 parameters

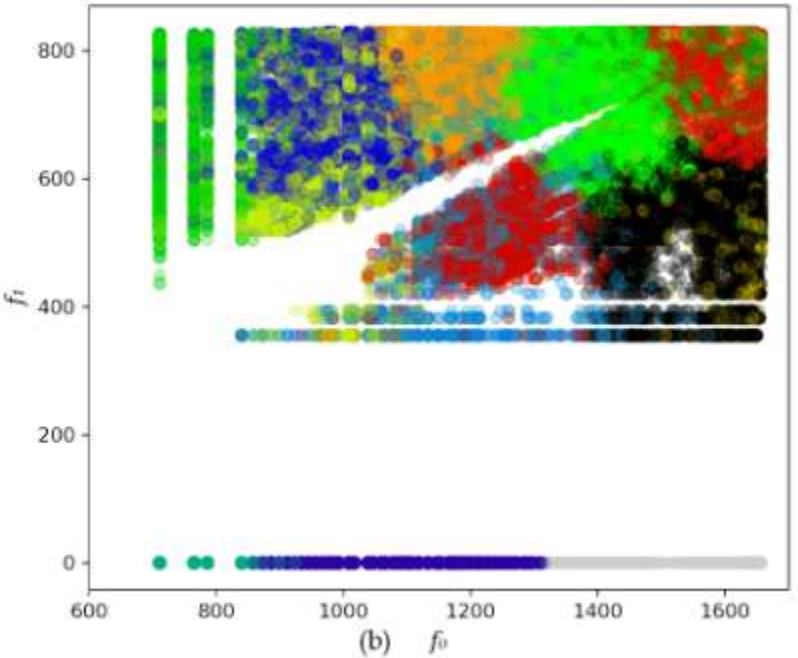


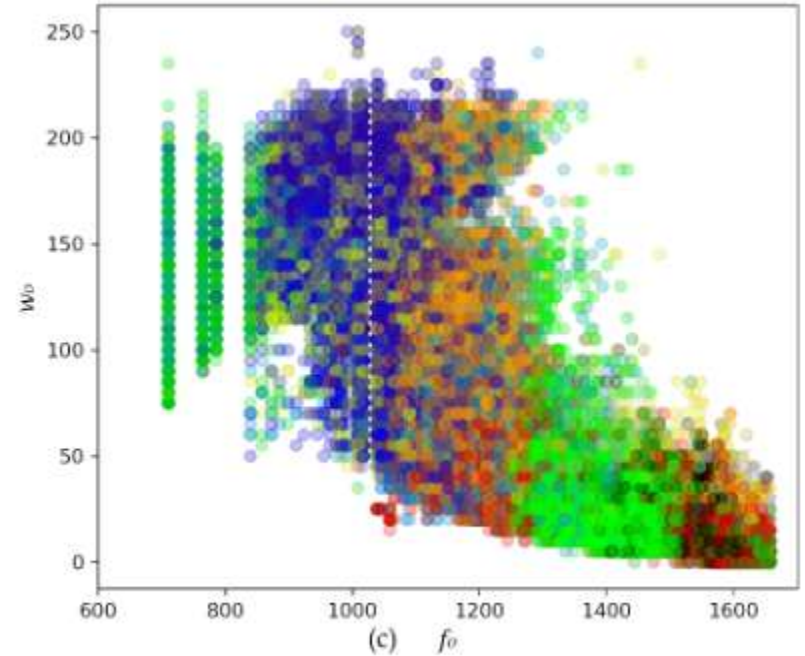
Parameter Discrimination 参数区分

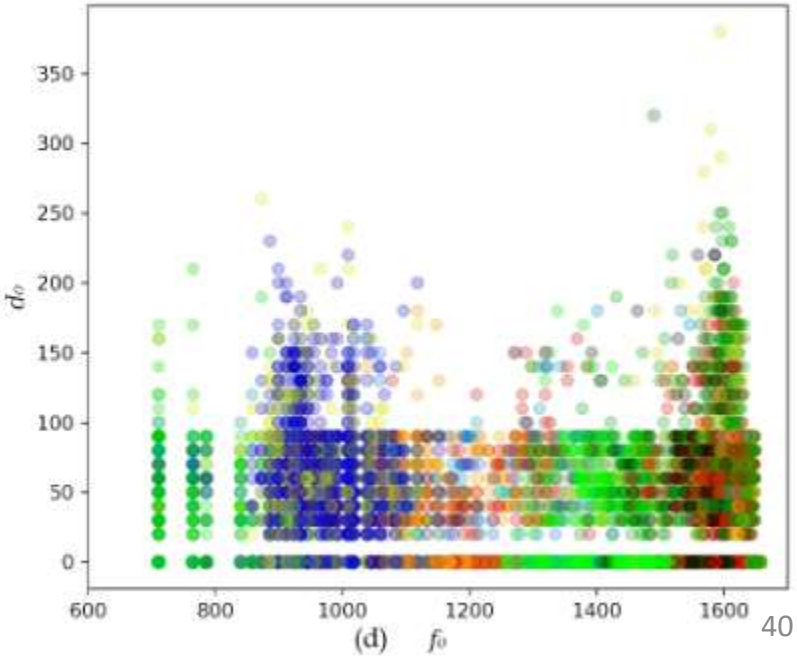
Noticeable Difference by Humans

- Fundamental Frequency: 0.3-0.5%
- Formant Frequencies: 3-5%
- Formant bandwidth: 20-40%
- Overall Intensity: 1.5 dB


1000
800
600
400
200
0
f_0
600
800
1000
1200
1400
1600
(a) f_0


800
600
400
200
0
f_1
600
800
1000
1200
1400
1600
(b) f_0


250
200
150
100
50
0
w_0
600
800
1000
1200
1400
1600
(c) f_0


350
300
250
200
150
100
50
0
d_0
600
800
1000
1200
1400
1600
(d) f_0

Future Research Directions

Time-Domain Analysis 时域分析

- High quality sound synthesis using Time-domain analysis
  https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

1 Second

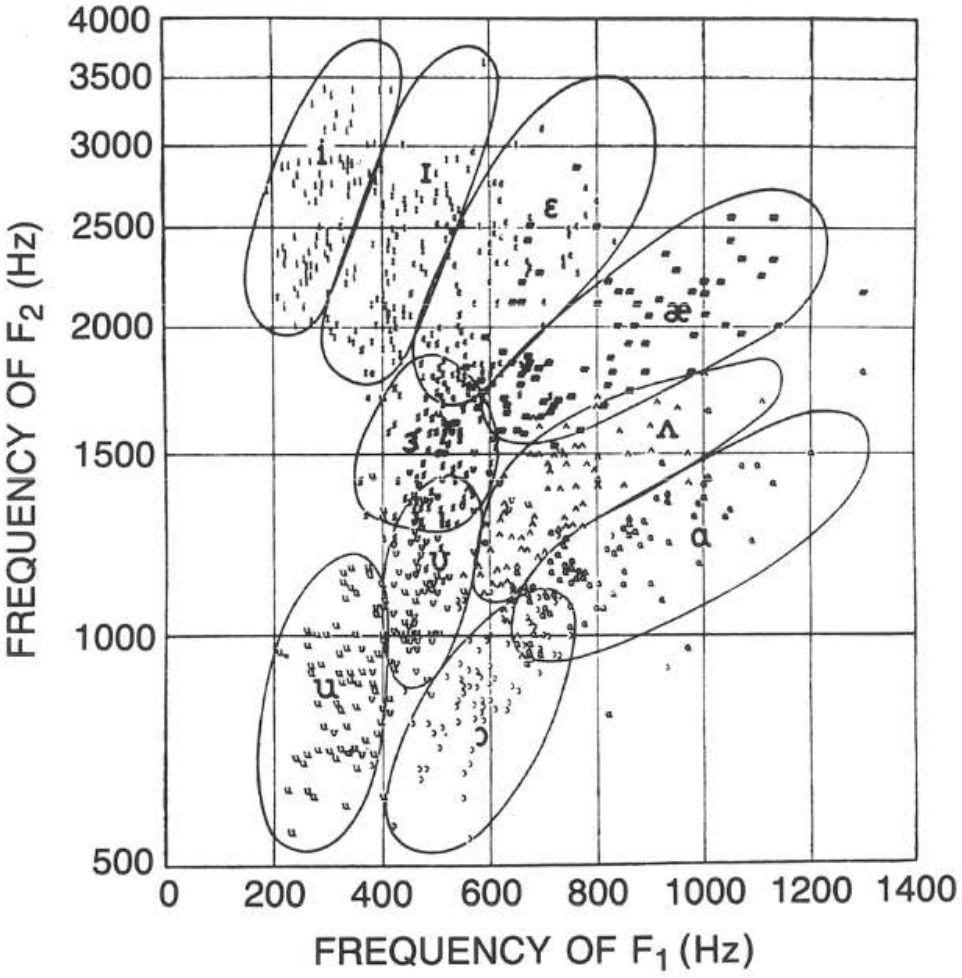
TinyML

- A lot of potential applications in speech interactable medical assistive technology.

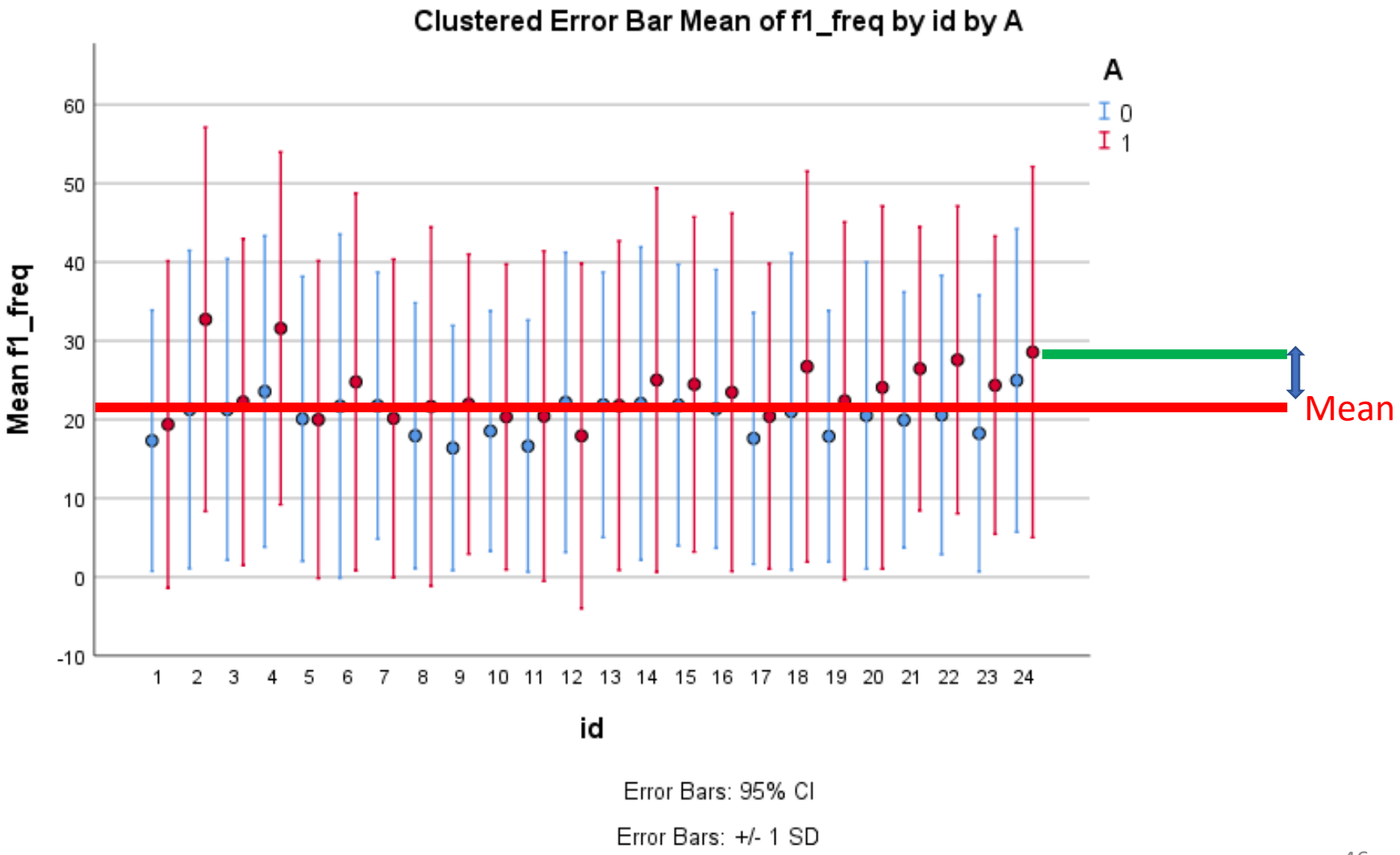
Stereo Analysis 立体分析

- Intensity (DB) is an unreliable source of feature extraction.
- It can be solved by analyzing stereo audio.

Phoneme Sub-Classes



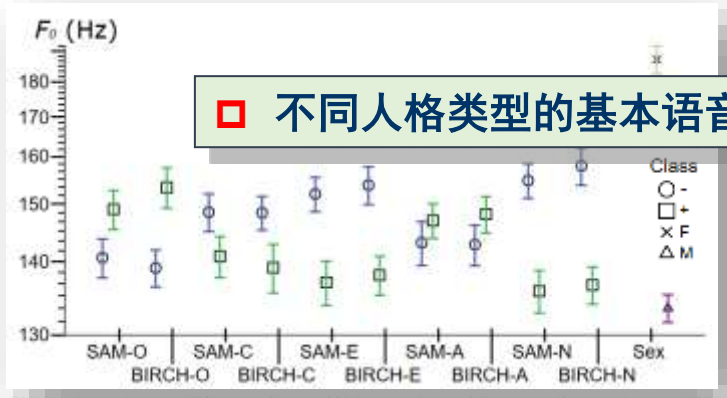
Phoneme Residual



Current Progress

Paper Abstract---个性识别方法

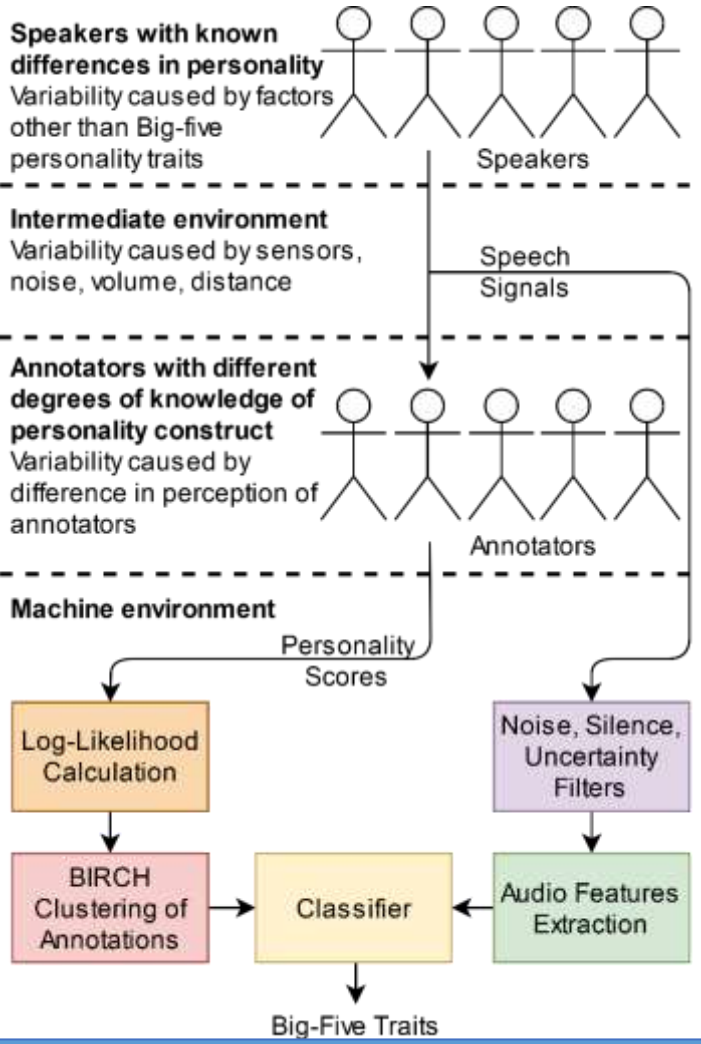
➢ **基于对数似然距离的注释分类的语音个性识别和关键音频特征的提取**



- 不同人格类型的基本语音差异分析

Rater	-O	+O	-C	+C	-E	+E	-A	+A	-N	+N	-Avg.	+Avg.
1	-1.15	0.77	-1.62	0.93	-1.72	1.45	-0.53	0.47	-0.31	0.36	-1.06	0.80
2	1.78	-1.19	-0.80	0.46	-1.45	1.22	-1.85	1.64	-0.84	0.97	-0.63	0.62
3	-1.70	1.14	-1.61	0.93	-1.54	1.30	-1.30	1.14	-1.04	1.21	-1.44	1.14
4	-0.36	0.24	-1.18	0.68	-1.64	1.38	-0.89	0.79	-0.92	1.06	-1.00	0.83
5	-0.95	0.64	-2.08	1.20	-2.23							
6	-0.07	0.05	-0.70	0.40	-2.22							
7	-0.07	0.05	-0.07	0.04	-1.34							
8	-0.54	0.36	-0.27	0.15	-1.40							
9	-0.29	0.20	-0.66	0.38	-0.42							
10	-0.14	0.10	-0.06	0.04	0.01	-0.01	0.00	-0.00	-0.05	0.06	-0.05	0.04
11	-0.19	0.10	-1.22	0.70	-0.65	0.55	-0.72	0.63	-0.99	1.15	-0.75	0.87
Avg.	-0.33	0.22	-0.93	0.54	-1.33	1.12	-0.87	0.77	-0.75	0.87	-0.84	

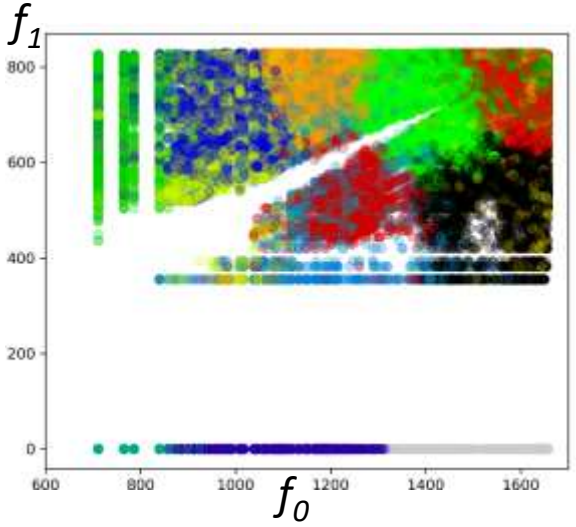
- 识别聆听者对人格的感知模式，以提取客观分数



O(开放性), C(认真的), E(外向性), A(和可亲), N(神经质)

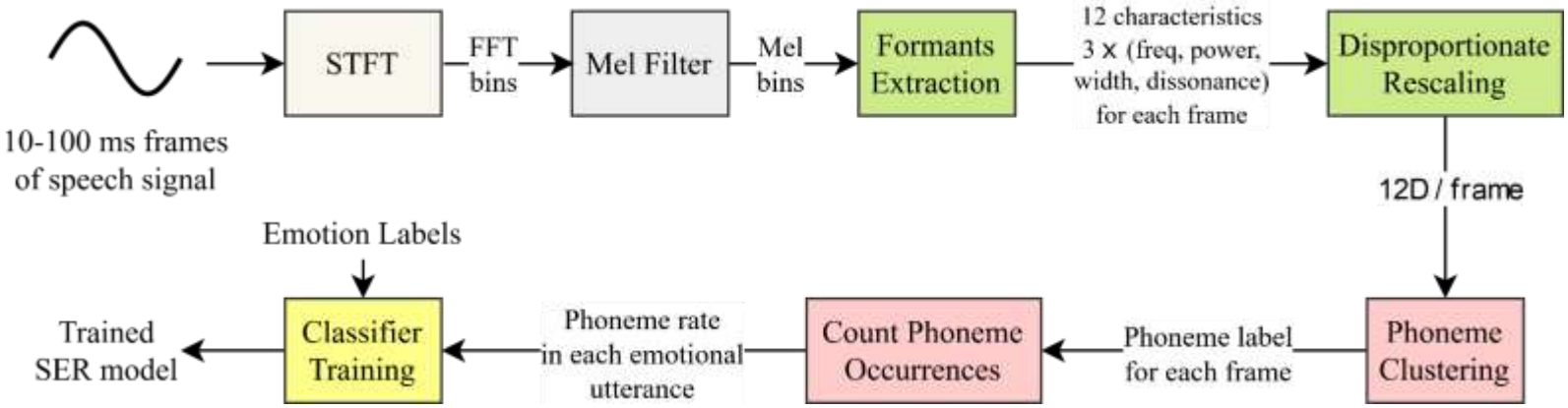
✓ Zhen-Tao Liu, Abdul Rehman, Min Wu, Wei-Hua Cao, and Man Hao. Speech Personality Recognition Based on Annotation Classification Using Log-likelihood Distance and Extraction of Essential Audio Features. *IEEE TRANSACTIONS ON MULTIMEDIA*. (SCI, T1）

情感识别方法



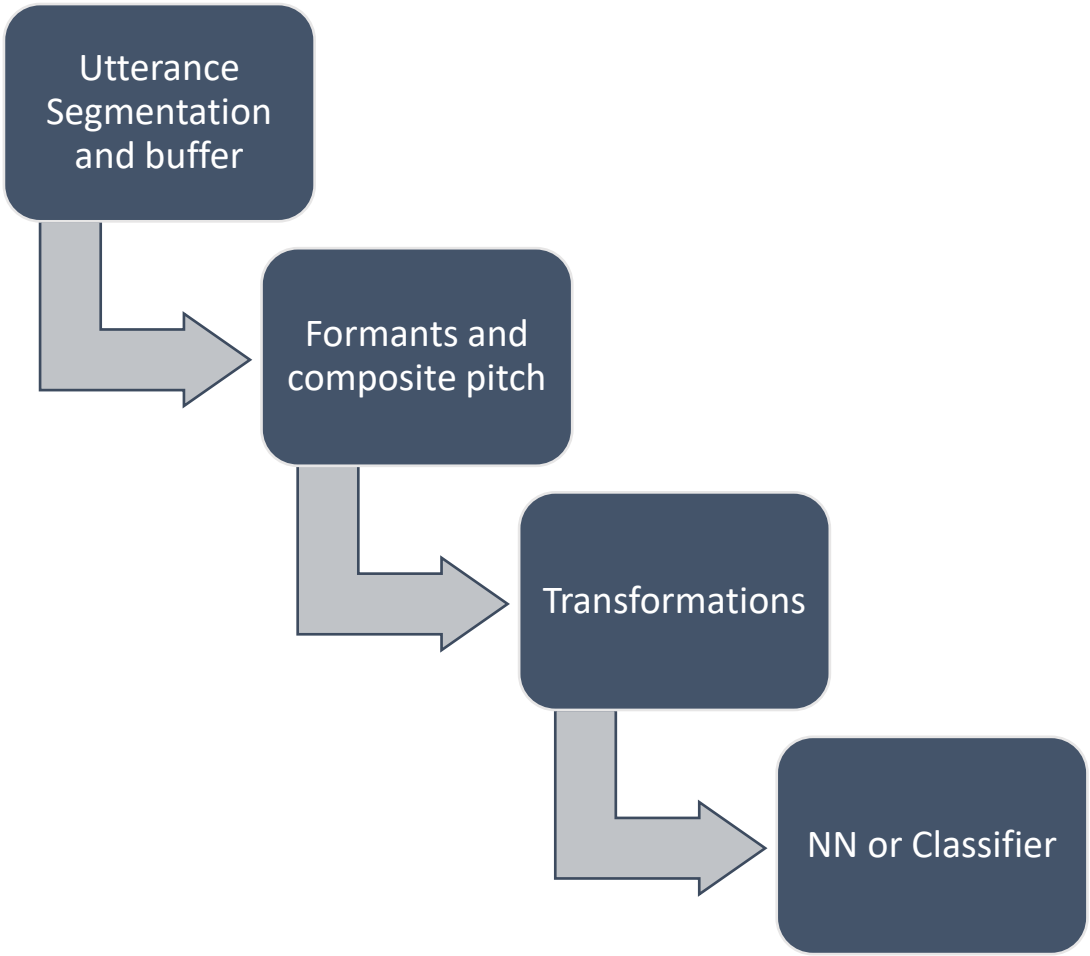
➢ 基于共振峰特征提取和音素的语音情感识别

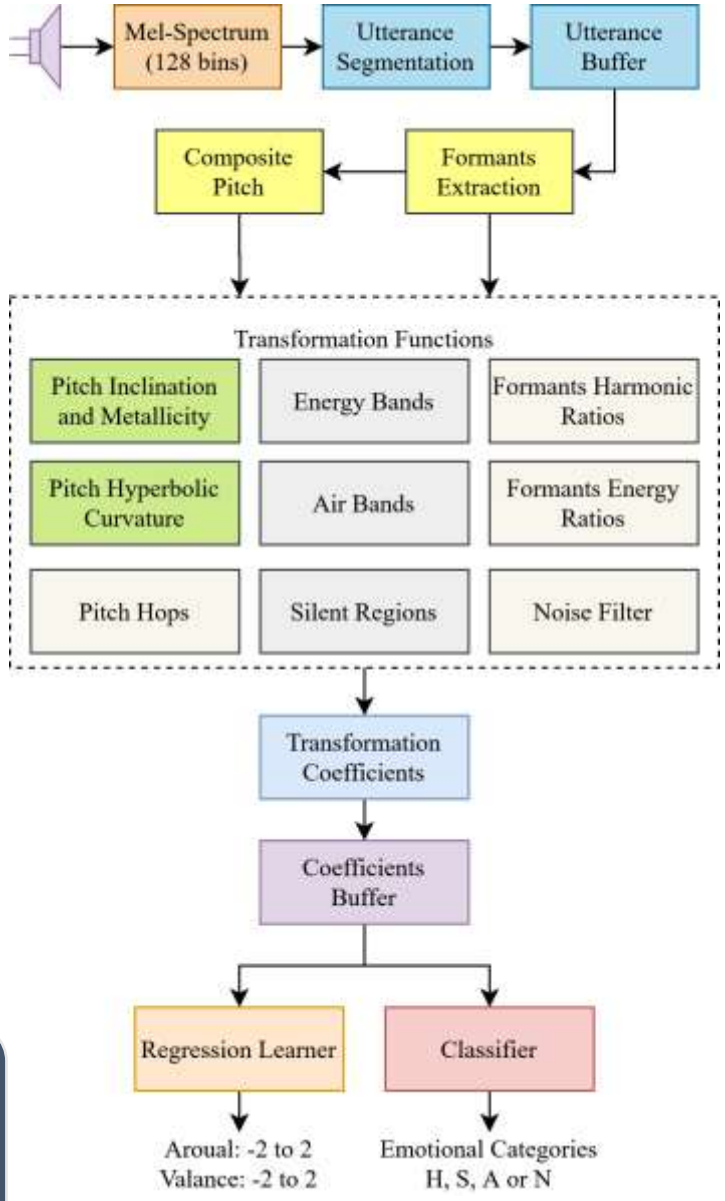
- 使用自动检测到的语音单位识别出情绪
- 根据共振峰特征的相似性，音素聚在一起
- 实验结果表明，降低了计算成本，提高了鲁棒性



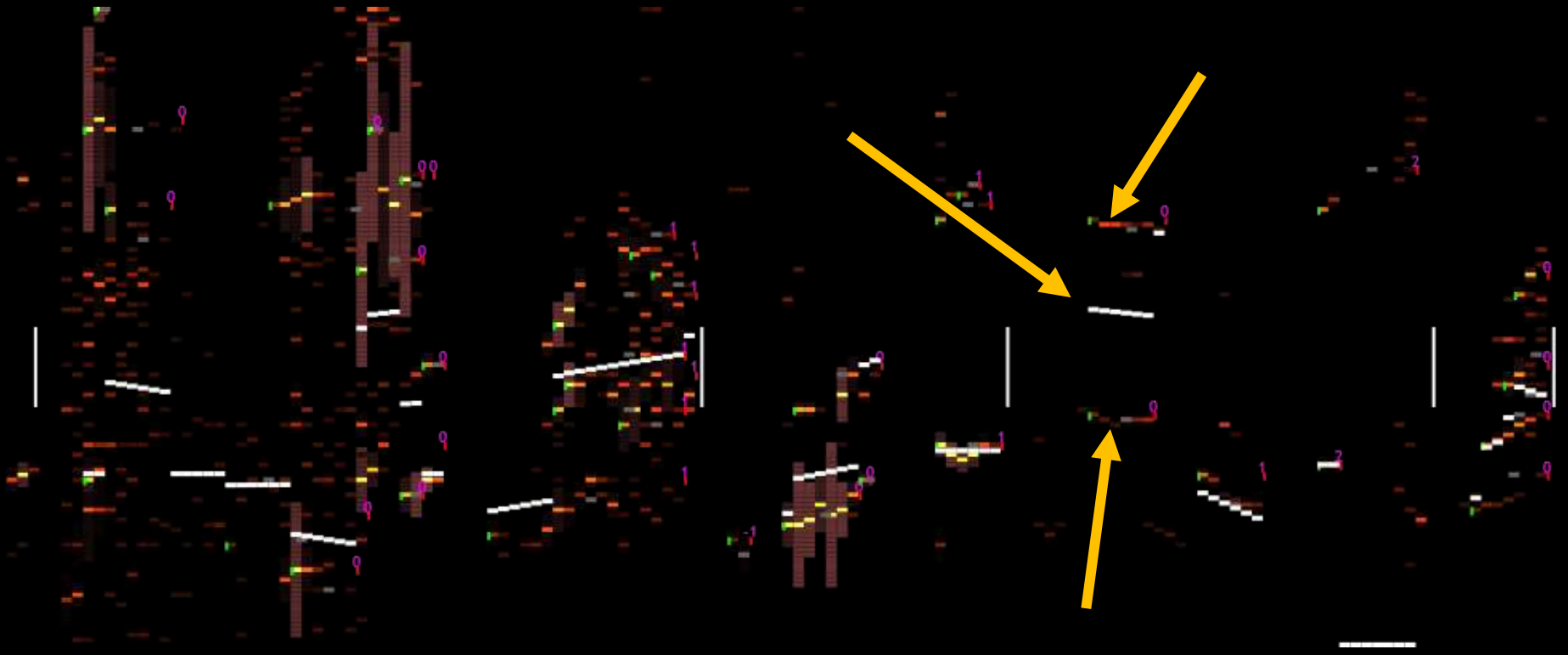
✓ Zhen-Tao Liu, Abdul Rehman, Min Wu, Wei-Hua Cao, Man Hao. Speech Emotion Recognition Based on Formant Characteristics Feature Extraction and Phoneme. Preprint

The ongoing work

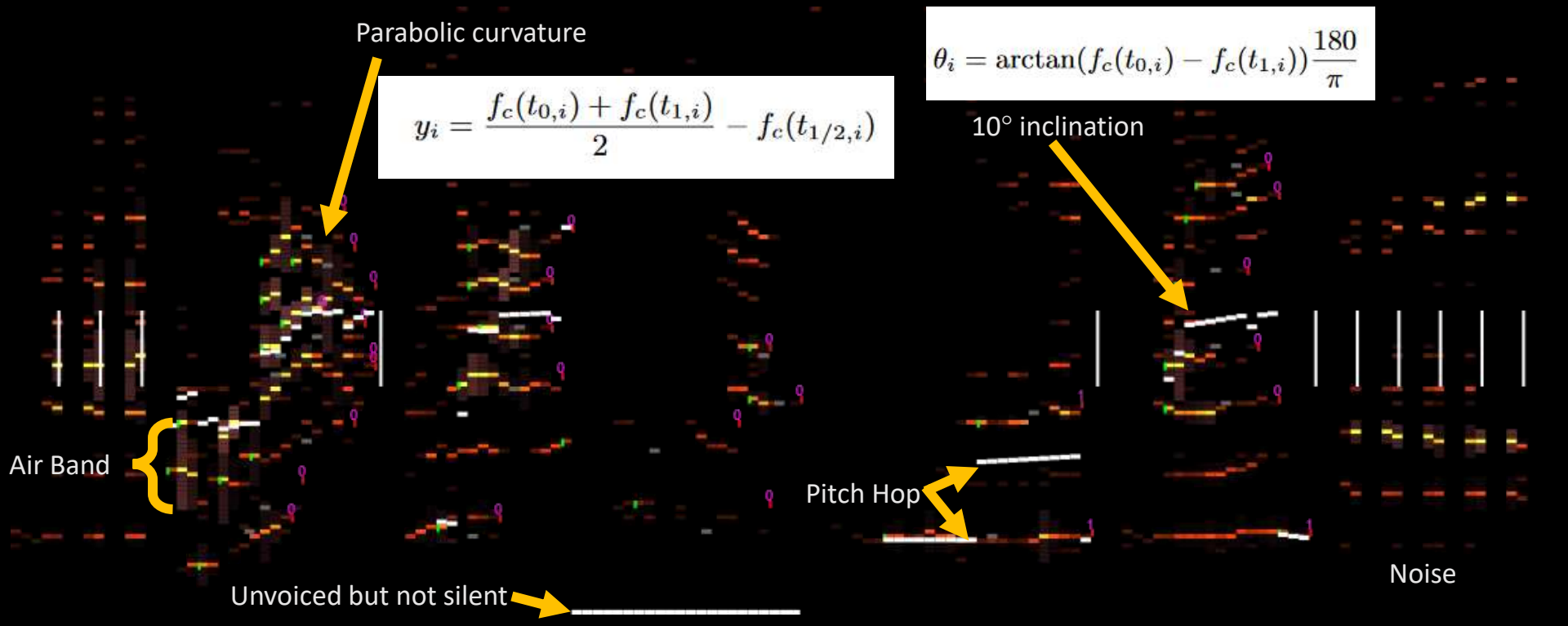




Formants and Composite Pitch
共振峰和复合音高



Transformation Examples



Experiments

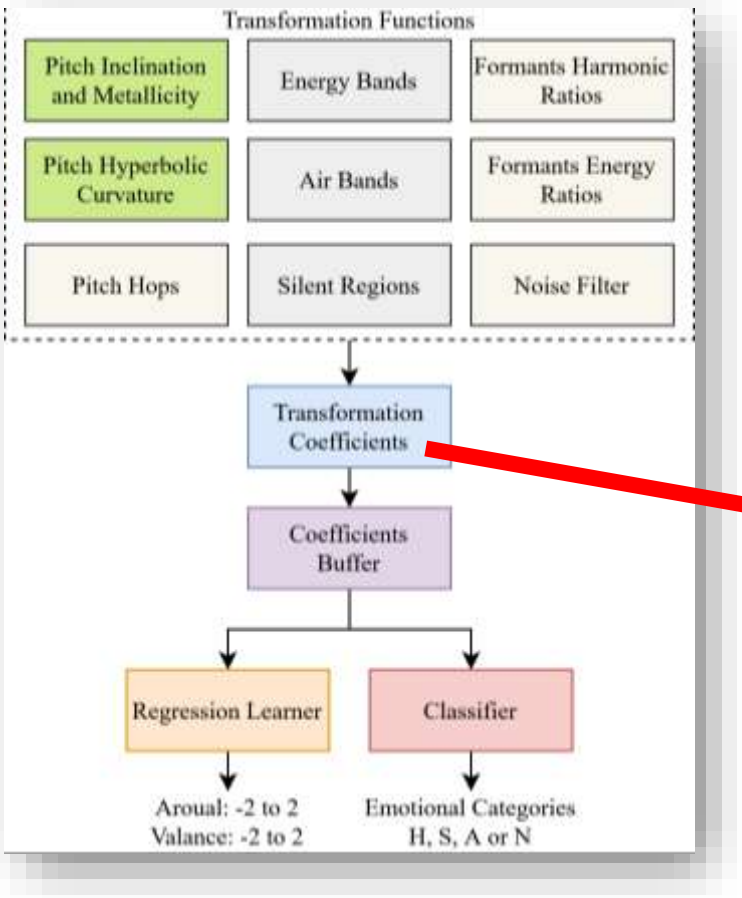


Table 1 Input vectors and output dimensions of transformers.

Transformer	Input	Output
Inclination and Metallicity	$f_c(t)$	5x6
Hyperbolic Curvature	$f_c(t)$	5x6
Pitch Hops	$f_c(t)$	10
Energy Bands	$f_h(t), p_h(t)$	12
Air Bands	$f_h(t), s_h(t)$	6
Silence (Pauses)	$f_h(t)$	4
Formant Harmonic Ratios	$f_h(t)$	4
Formant Energy Ratios	$f_h(t), p_h(t)$	4
Noise Filter	$f(t,l), p(t,l)$	1
All coefficients		101

Experiments

- Feature means for 4 emotions, 2 sexes

	F-N	F-S	F-H	F-A	M-N	M-S	M-H	M-A
Energy	2414.76	2169.04	2354.58	2528.63	2257.95	2381.12	2247.09	2586.46
bin_lens	88117.7	83954.2	89938.3	80715.3	85589.5	95583.9	86197.6	88715.9
bins_span	41539	34746.4	35638.2	34011.4	37253.9	43486.6	37720.4	35087.4
bins_diff	12799.2	7240.13	10643	11794.5	12333.5	11236.4	10899.8	11818.8
bins_fire_up	14844.1	1188.51	5651.53	8130.73	10182.3	5231.98	8229.54	6723.99
bins_fire_dn	6334.94	2330.97	5761.19	7143.05	7801.94	5438.85	5857.64	5871.46
bins_shape_0	16153.9	10940.8	12212.4	13050.1	14655.9	13968.8	13238.5	13276.3
bins_shape_1	11667.5	3530.01	12823.6	19157.4	19310.9	10252.6	19503.9	18262.4
bins_shape_2	15120.8	6591.2	17985.1	20723.5	20806.8	11870.2	23209.5	19607.8
bins_shape_3	3626.91	5620.03	8265.22	10104.3	15435.7	8004.41	16256.8	4971.35
bins_shape_4	9137.12	7008.73	14733.3	15546	17033.4	5846.61	16546.6	14515.2
bins_shape_5	4070.47	2073.77	3345.48	2913.55	5073.67	2585.37	4386.32	2412.39

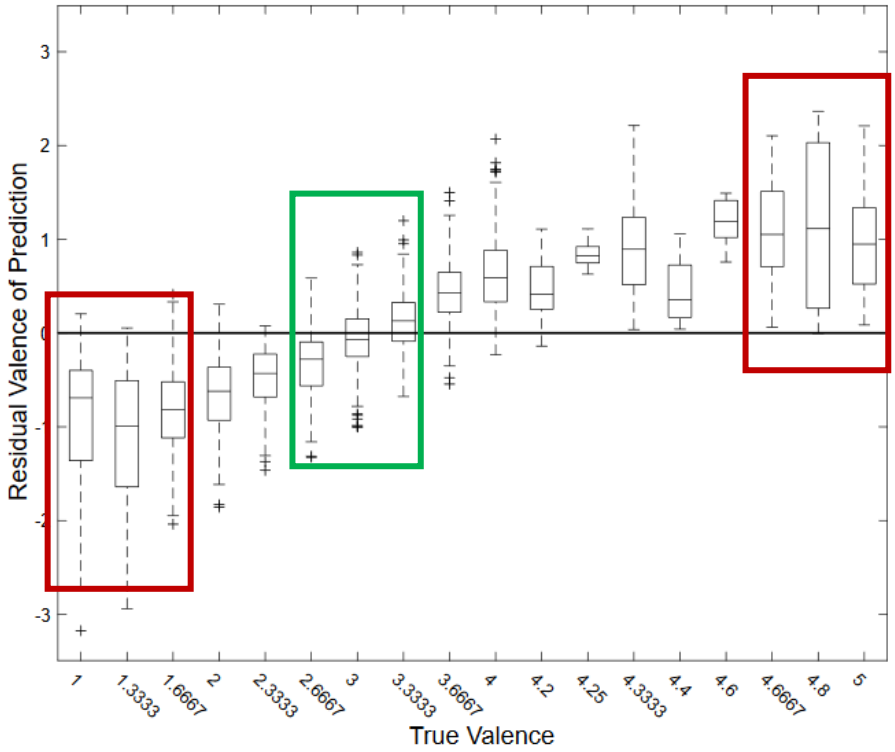
Preliminary results

Table 2 RMSE and R^2 of Gaussian Process Regression prediction when trained with annotation on continuous (-2 to 2) Arousal and Valence scales.

	IEMOCAP		MSP-Improv	
	Arousal	Valence	Arousal	Valence
RMSE	0.50	0.75	0.71	0.49
R^2	0.54	0.47	0.56	0.58

Table 3 UAR% of the SVM prediction of 3 classes (High, Low and Medium) of Arousal (A) and Valence (V) for different training and testing sets.

Train set \ Test set	IEMOCAP		MSP-Improv	
	A	V	A	V
IEMOCAP	71.0	65.3	44.2	48.9
MSP-Improv	54.6	41.6	69.7	67.4



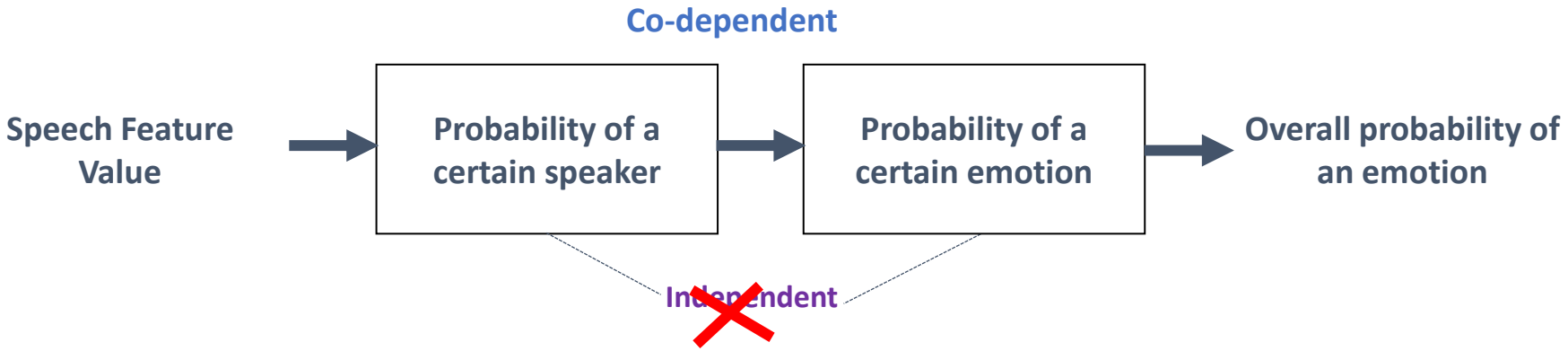
Current Challenges

- Holistic speaker modeling
  - Consider other states and traits that temporarily impact on the voice production.
  - In other words, “one is not only emotional, but also potentially tired, having a cold, is alcohol intoxicated, or, sounds differently because being in a certain mood.”

Current Challenges

- Robustness across cultures and languages as one of the major white spots in the literature.
  - A number of studies show the downgrades one may expect when going cross language.
  - Cross-cultural studies are still particularly sparse.

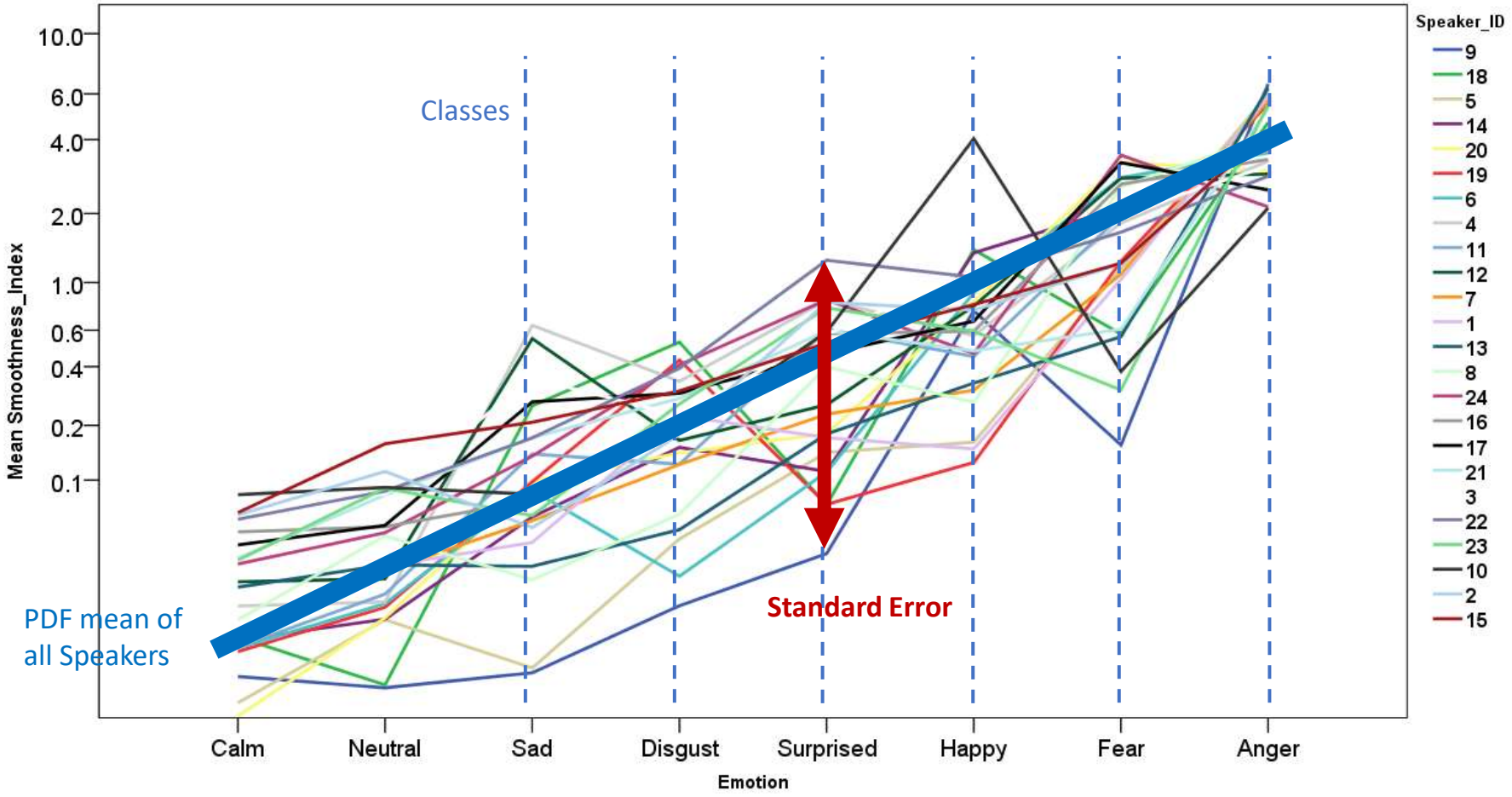
Fallacy of Generalization



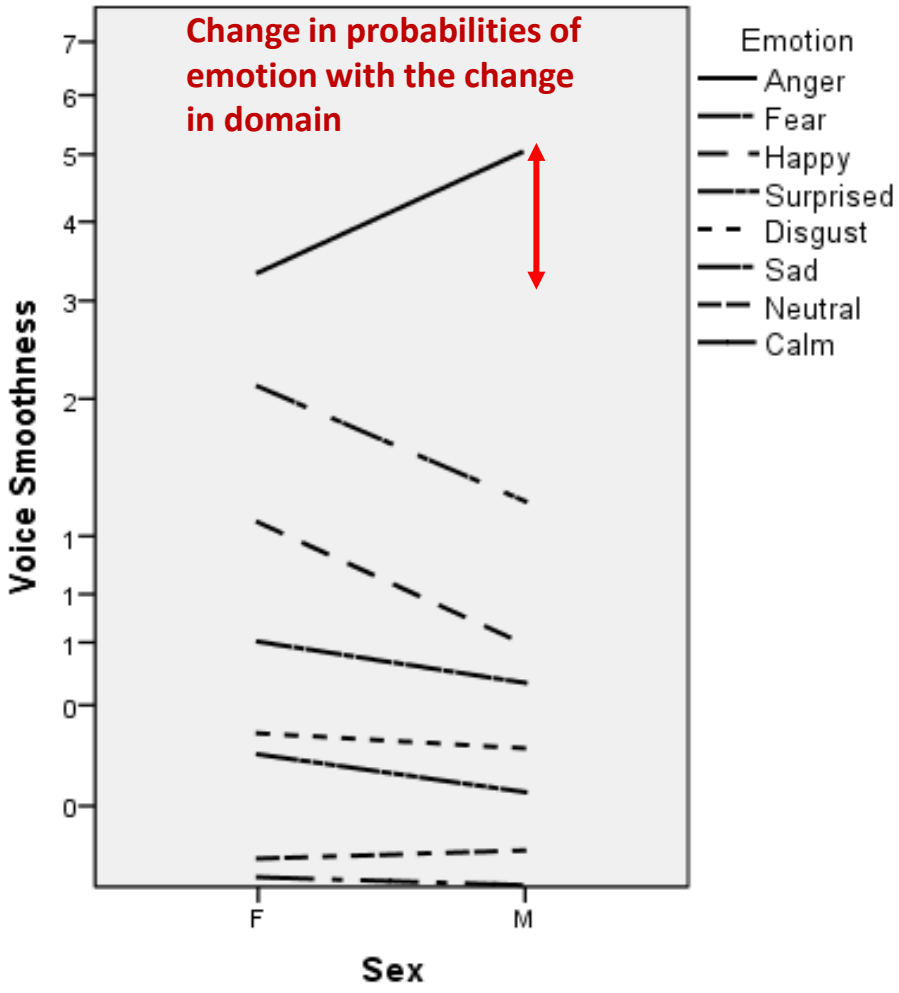
Changing the speaker (or domain) changes the probability of any given emotion within it.

Assuming the Naïve Bayesian assumption makes the task easy but overly generalized.
As its difficult to estimate the **Non-Naïve** Bayesian Bias with just 5 seconds of speech.

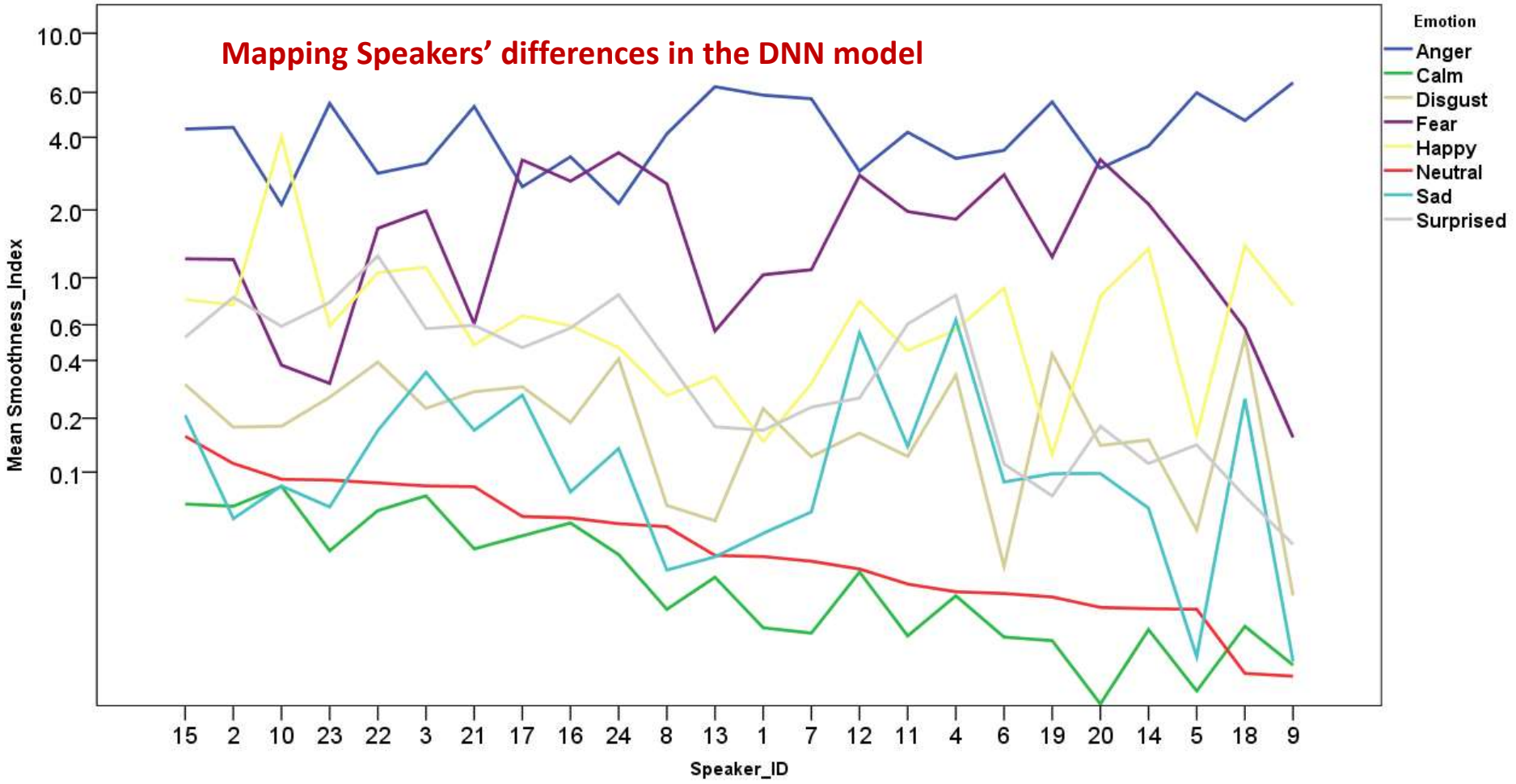
Generally: Ignoring Speaker Differences



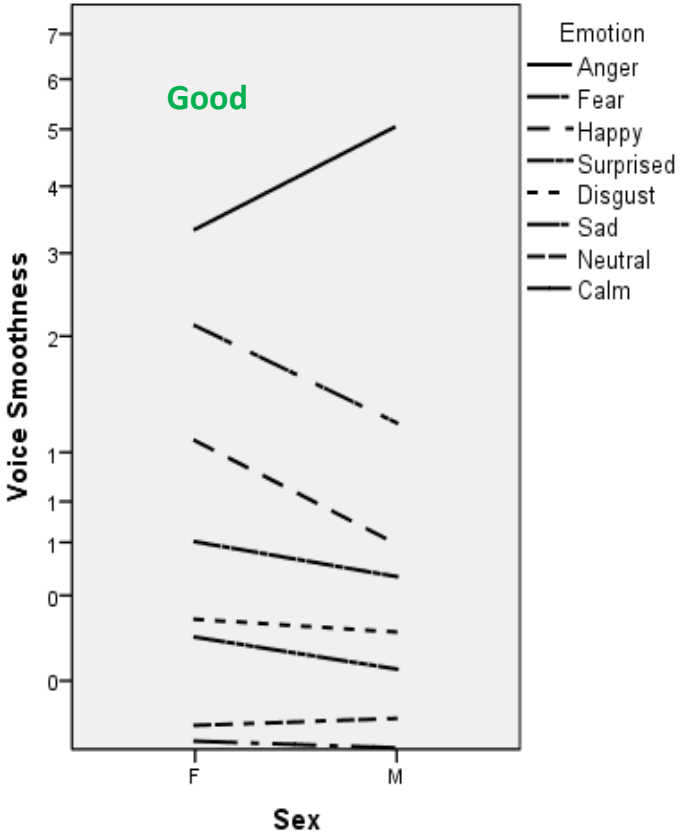
Learning the domain-emotion covariance

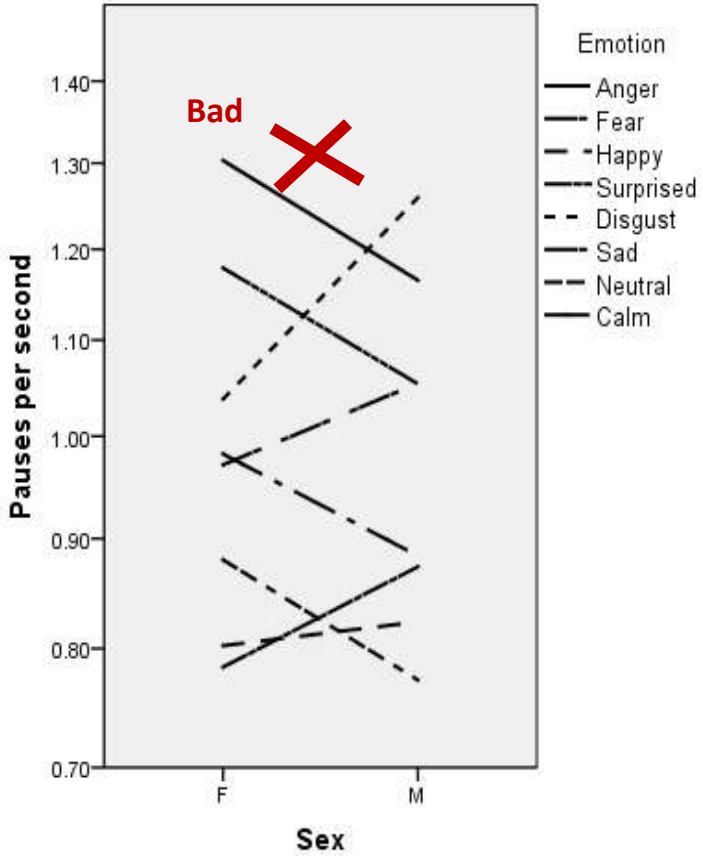


Speakers' Personality Terrain

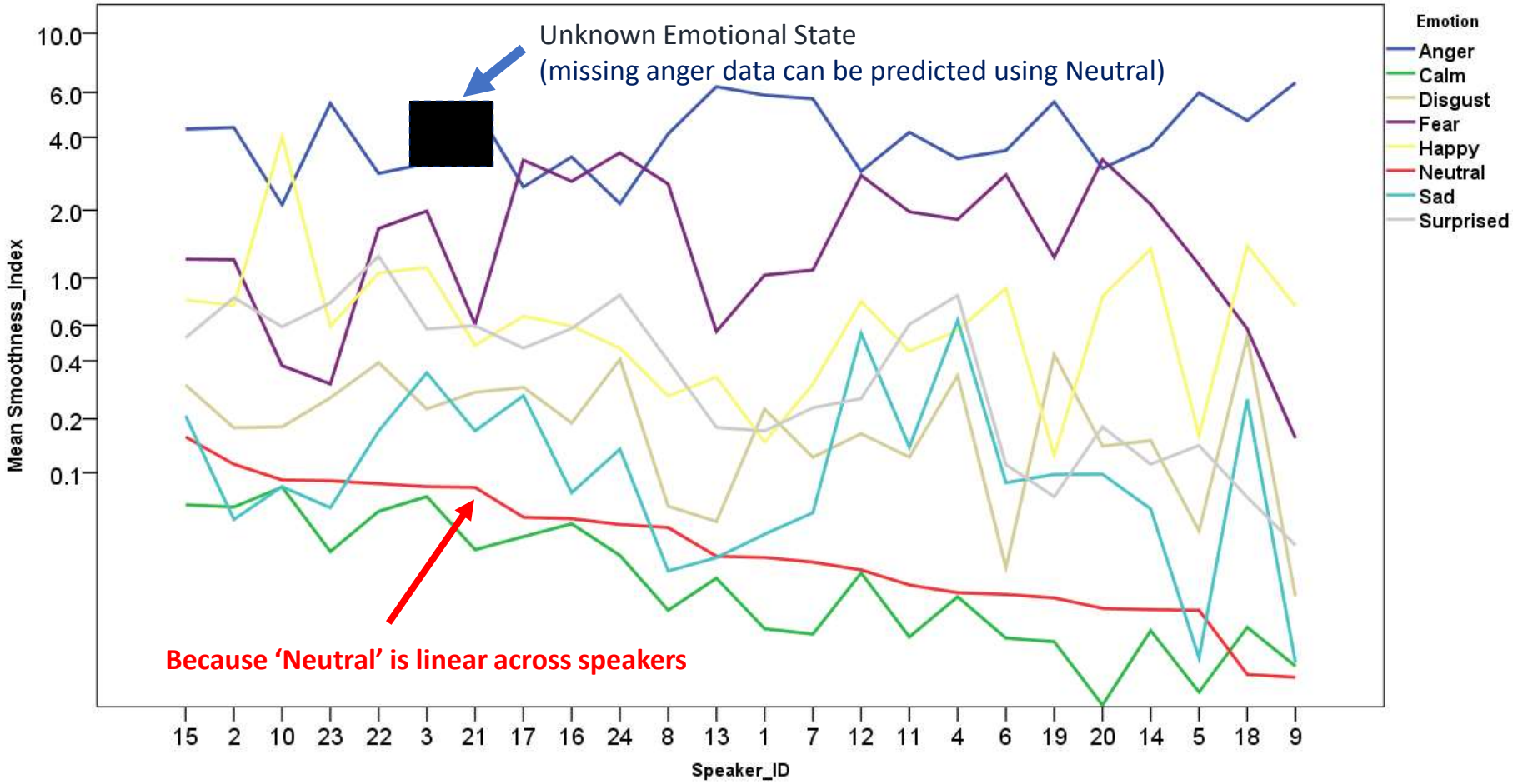


Trouble with the traditional generalization strategy



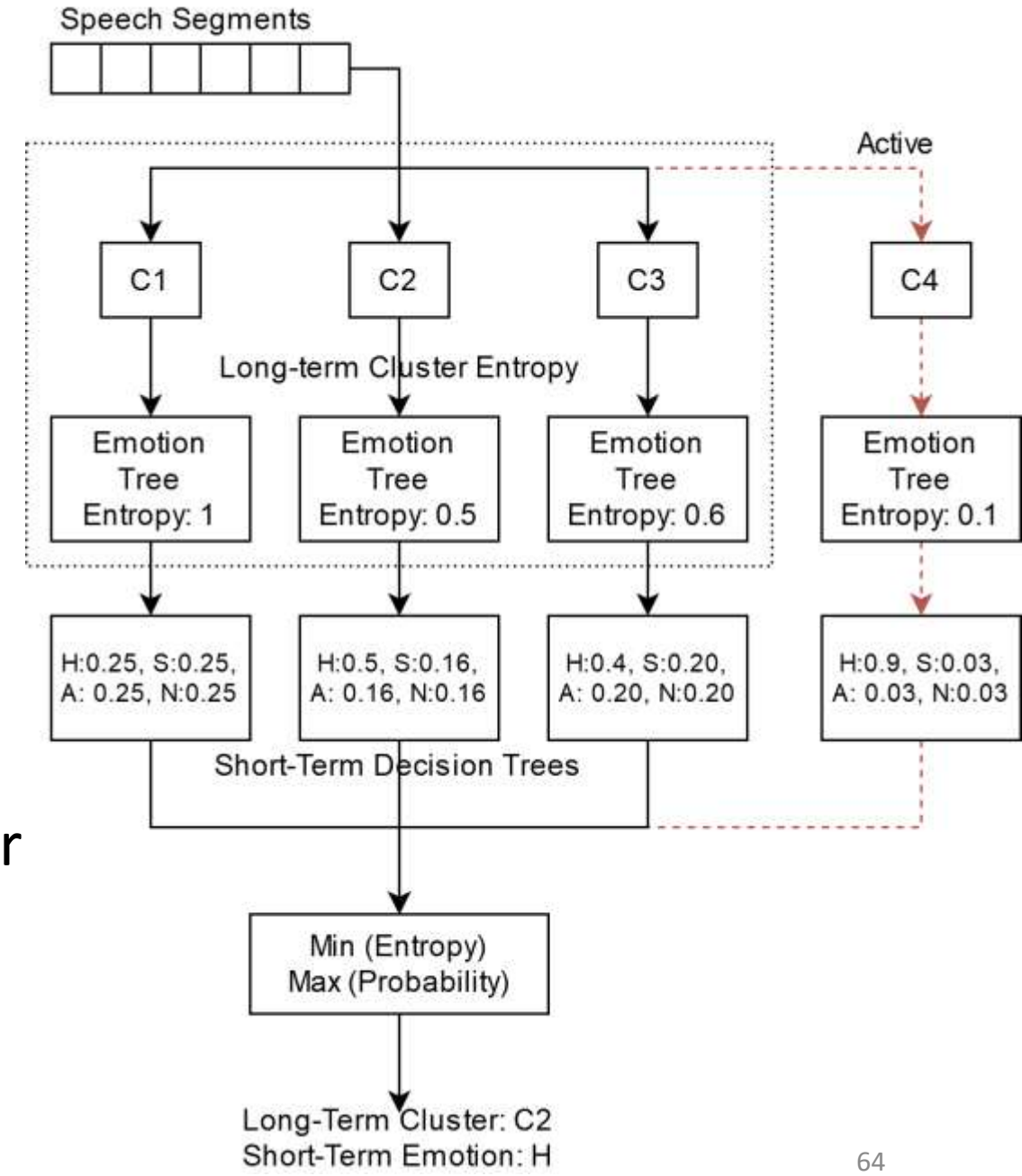


Sparse Autoencoding (Emotion-to-emotion)



Active learning solutions

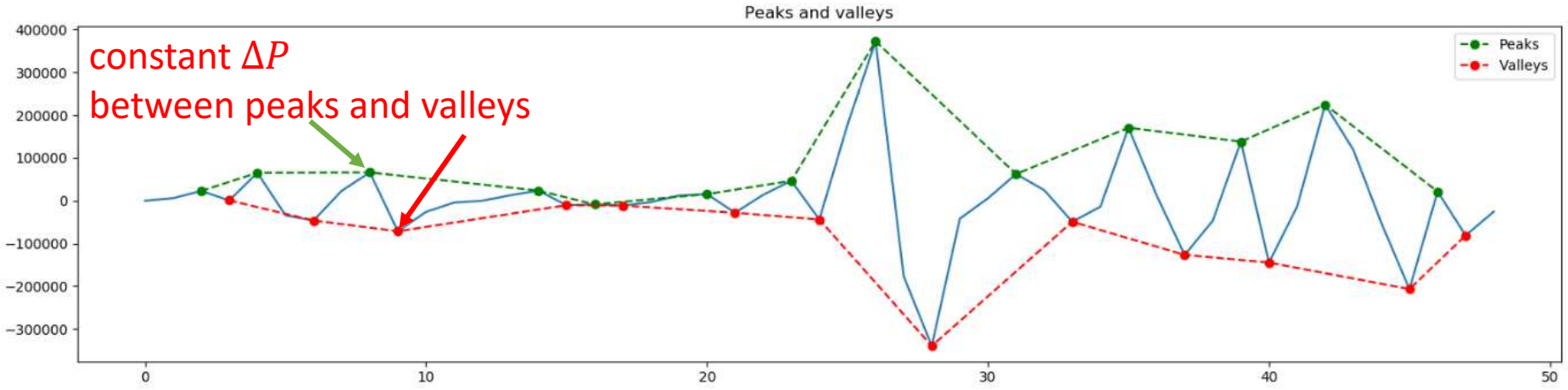
- Long-term Short-term Decision Trees.
- Long term Entropy
- Inference engine for cross-tree pattern recognition



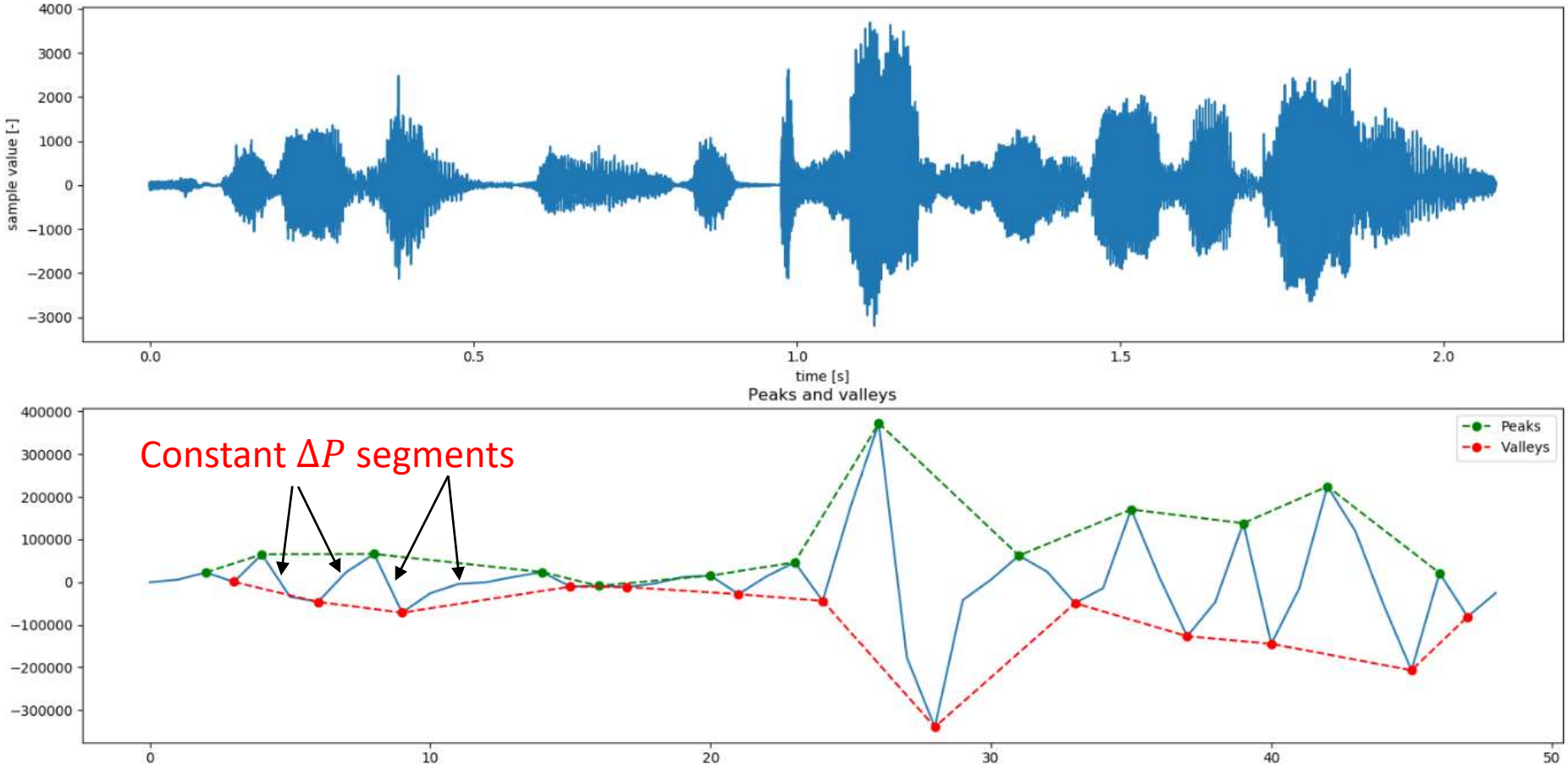
Finding Explainable Patterns

Quick Syllable Segmentation

- *Change in power variance:*
  - $\boldsymbol{dV(t) = d\,\mathrm{Var}\big(\mathrm{Power(t)}\big)\,dt}$
- Peaks = relative maxima of $\boldsymbol{dV(t)}$
- Valleys = relative minima of $\boldsymbol{dV(t)}$
- Instant changes in power mark the change of syllable:

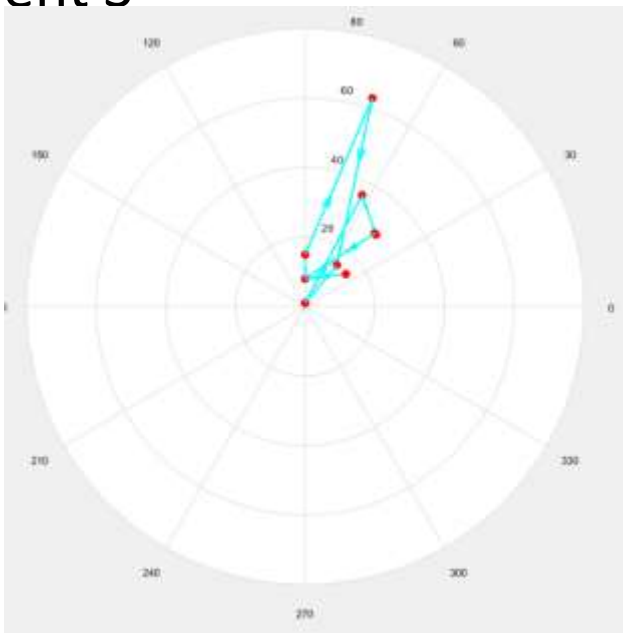


Quick Syllable Segmentation

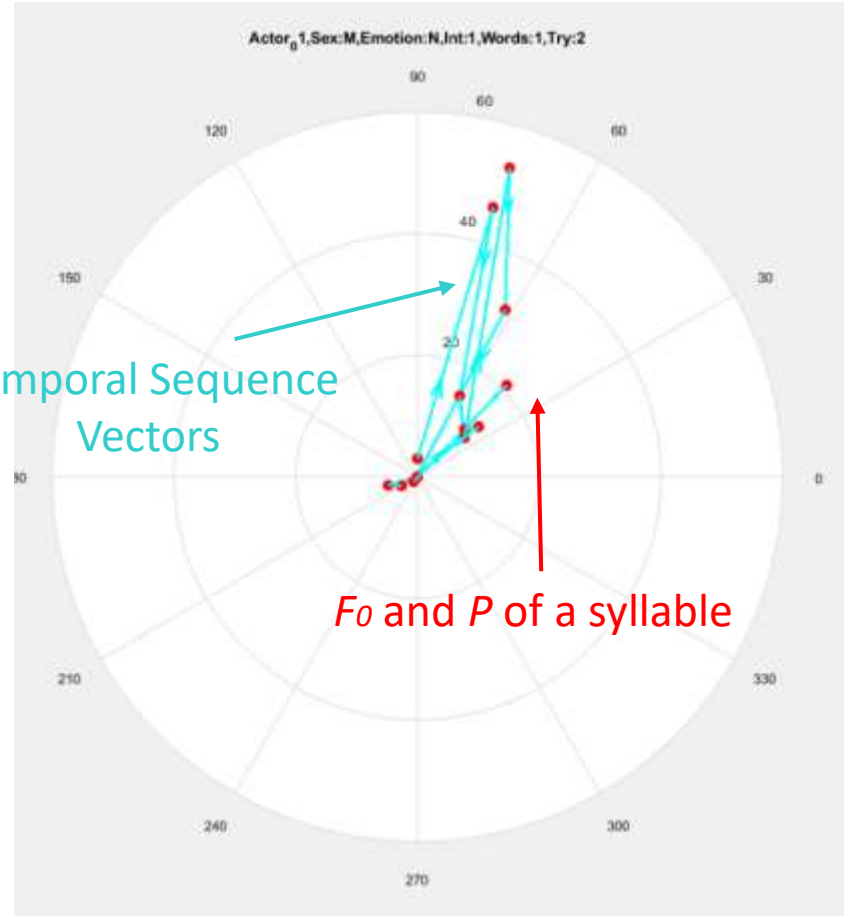


Polar Coordinates

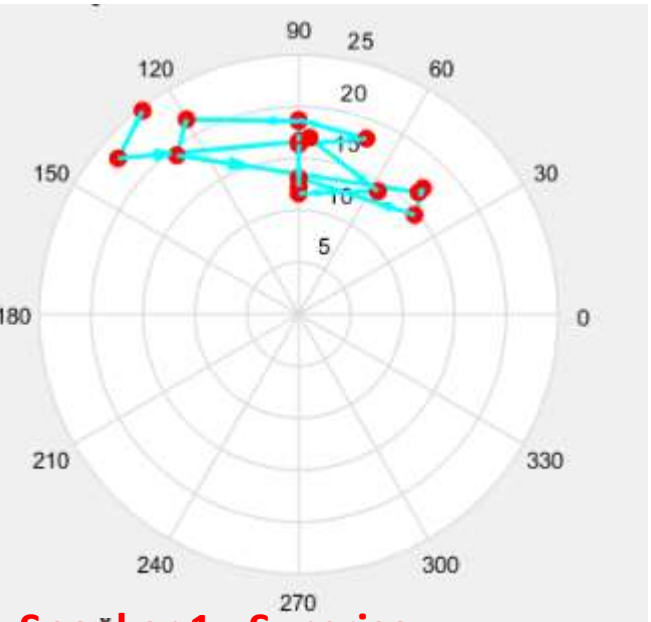
- $z = x + iy = P(S).e^{i(f_0(S)-60)}$
- Where
  - (S) is the iterator of segment/syllable
  - f_0 is the most fundamental frequency of the syllable S
  - $P(S)$ is the average power of the segment S



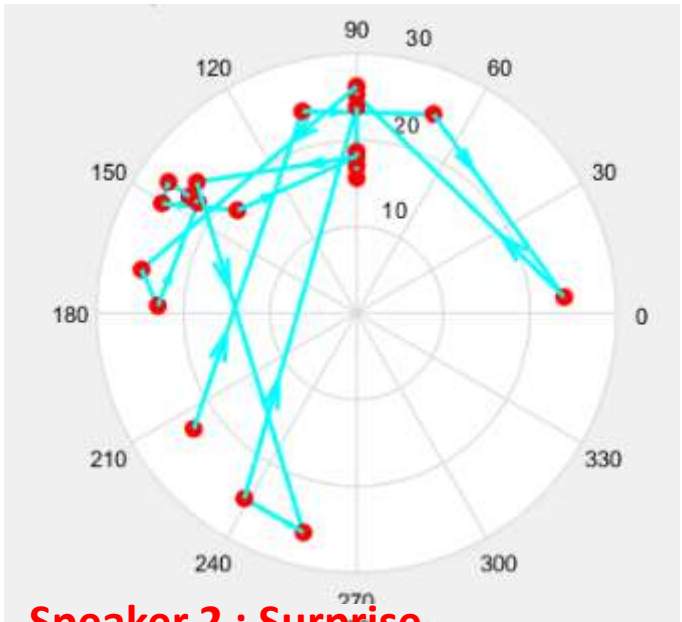
Temporal Sequence



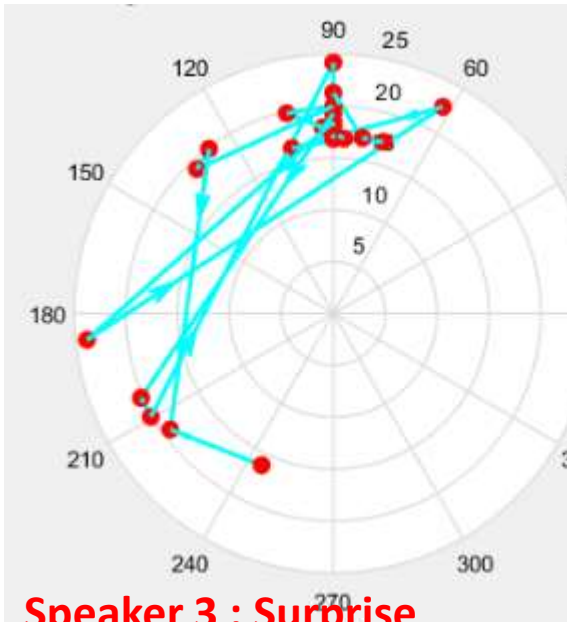
Explainable Patterns: Surprise



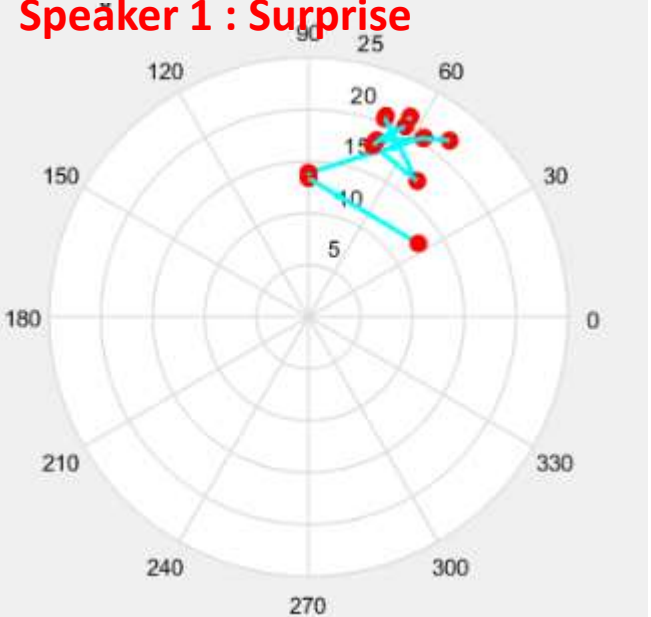
Speaker 1 : Surprise

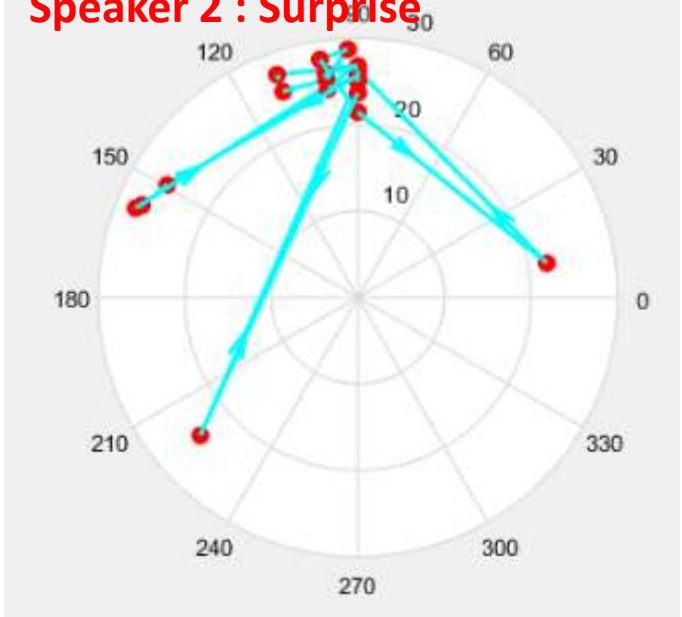


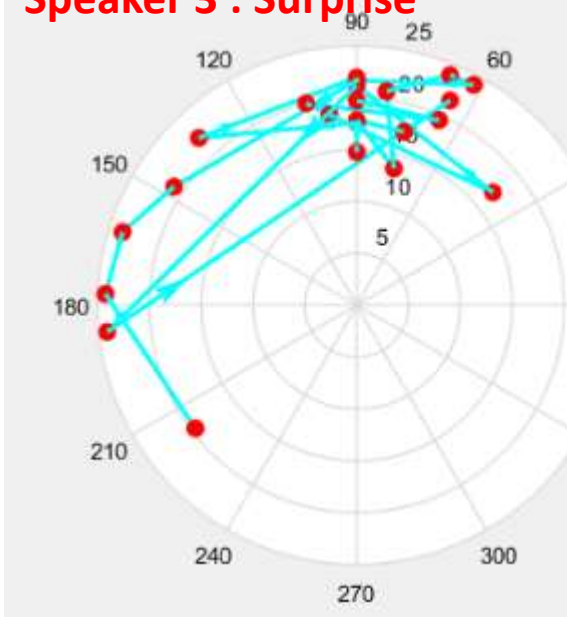
Speaker 2 : Surprise



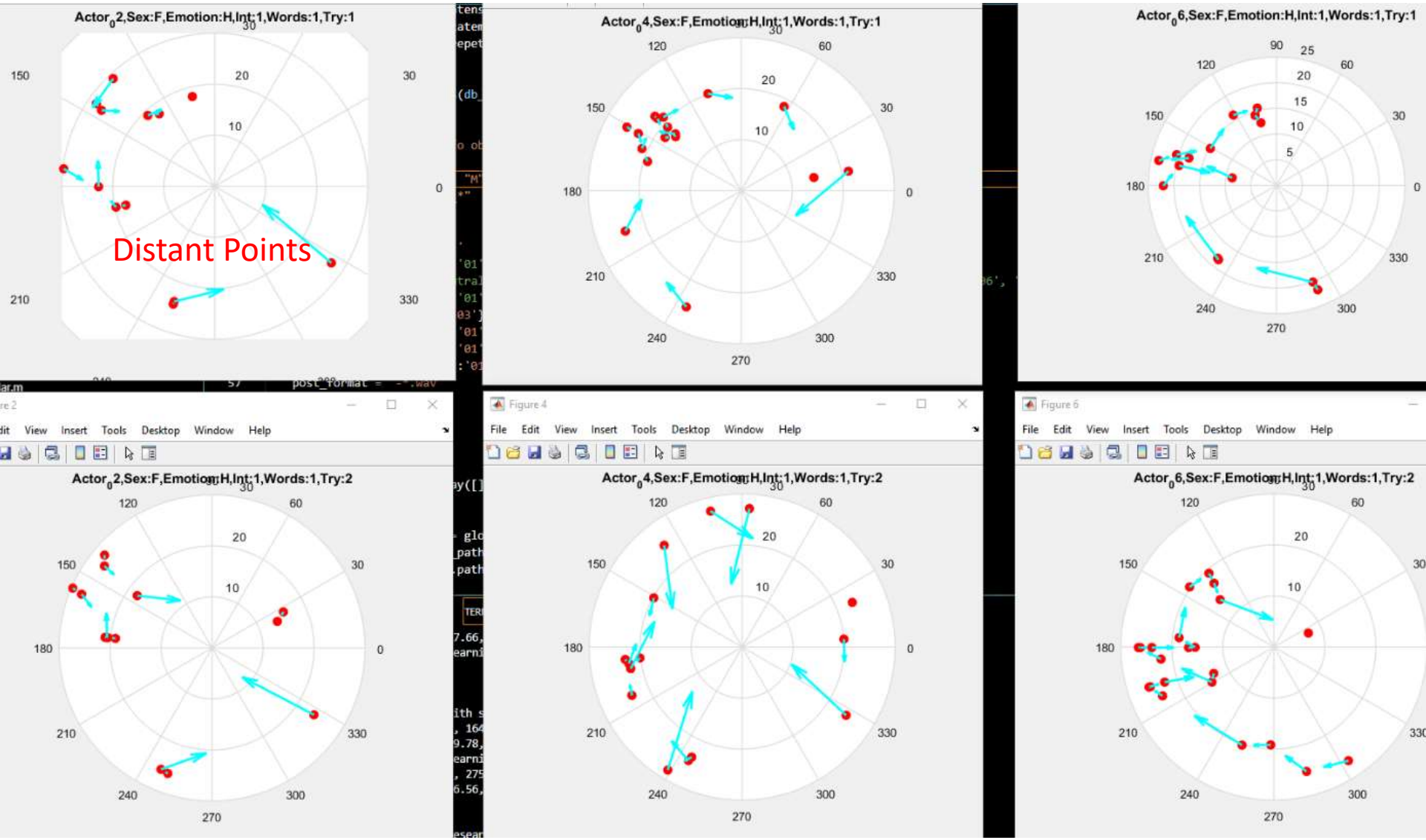
Speaker 3 : Surprise



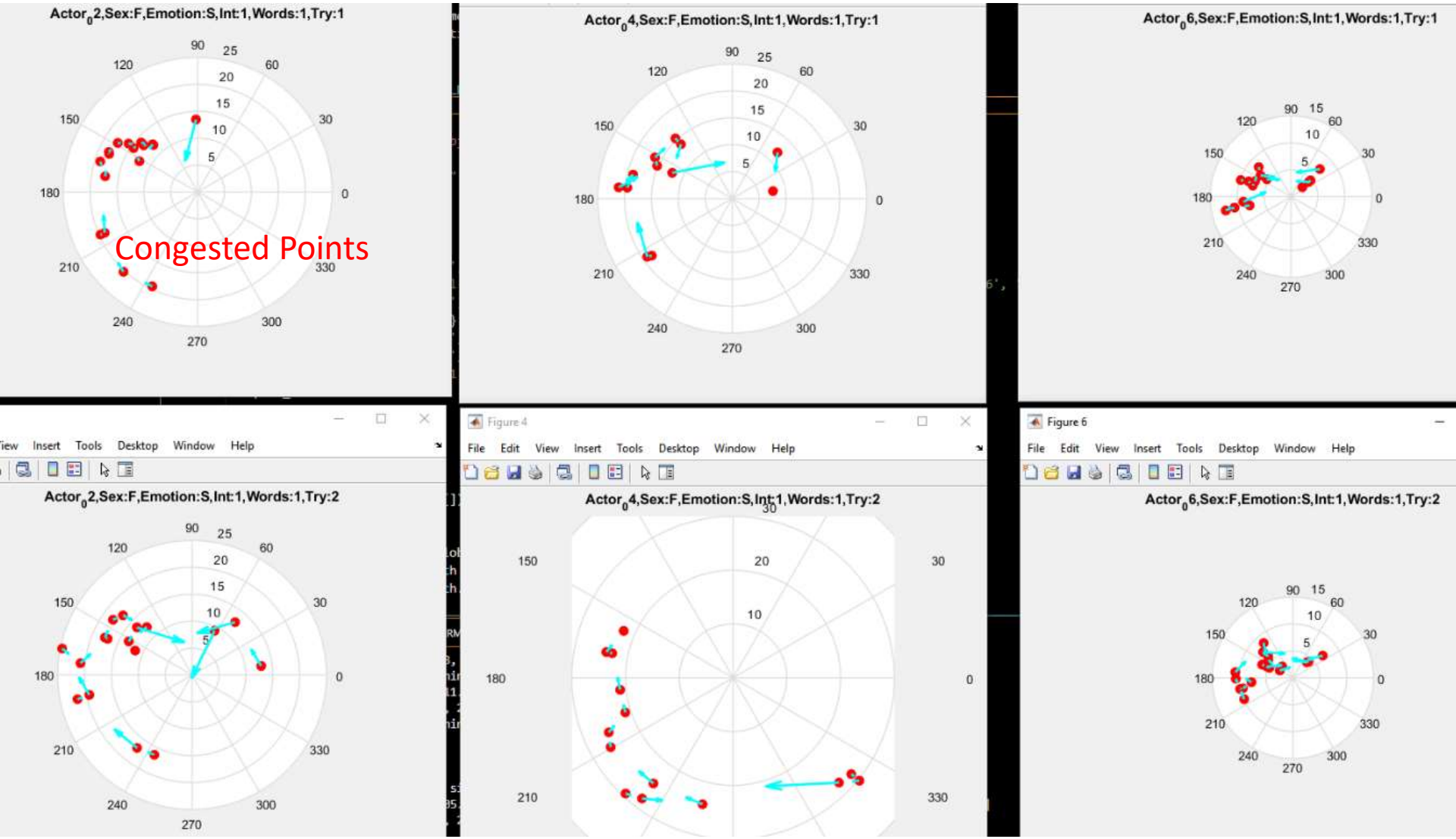




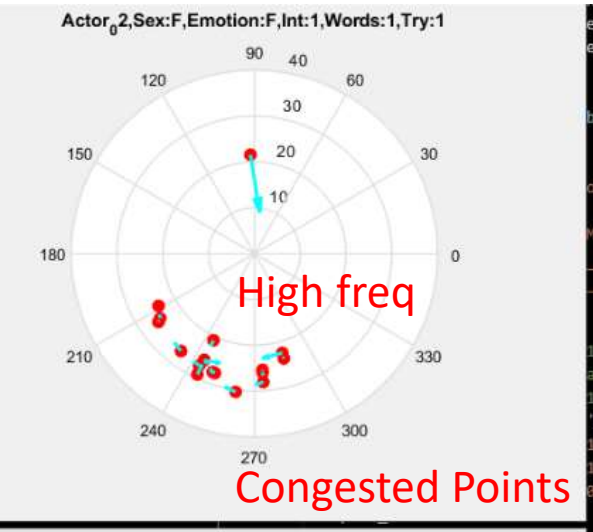
Explainable Patterns: Happy

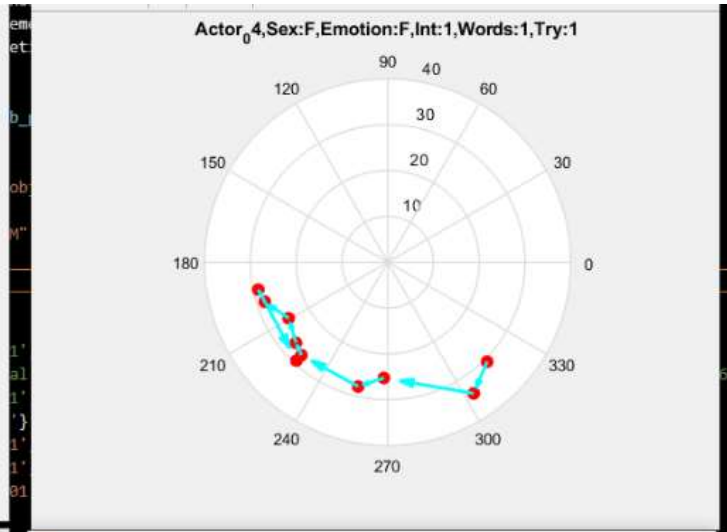


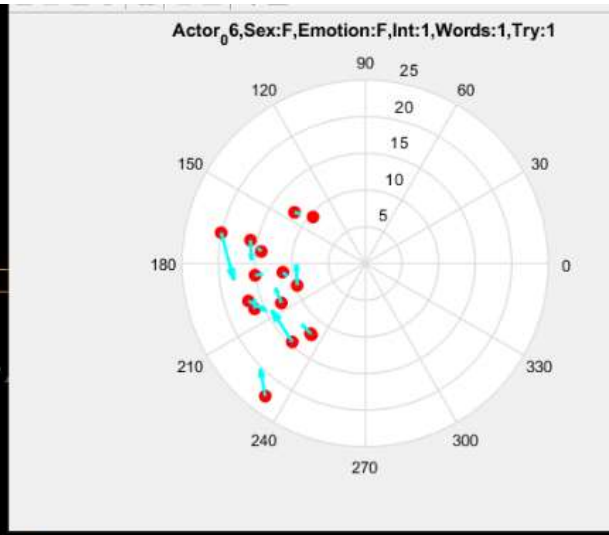
Explainable Patterns: Sad

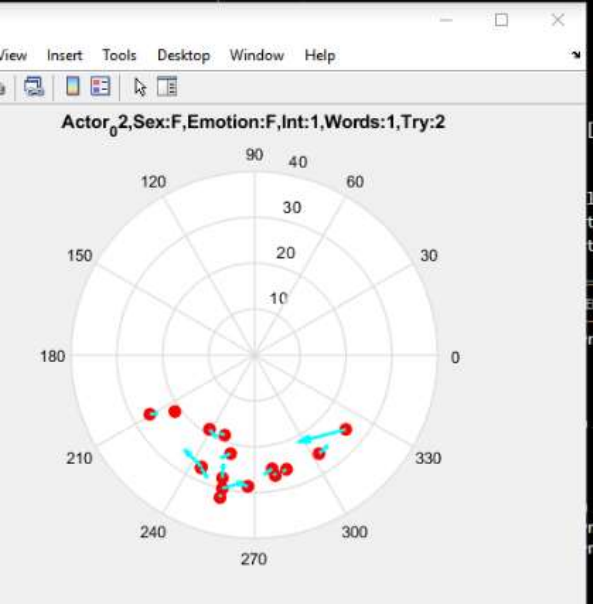


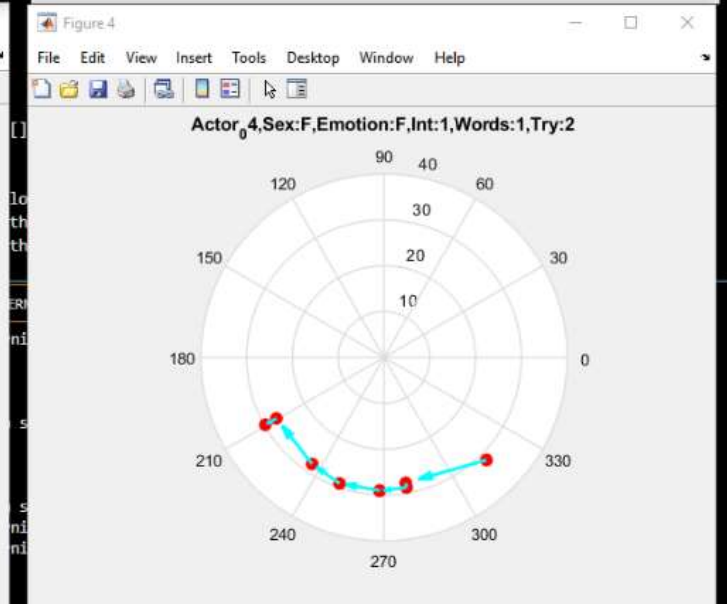
Explainable Patterns: Fear

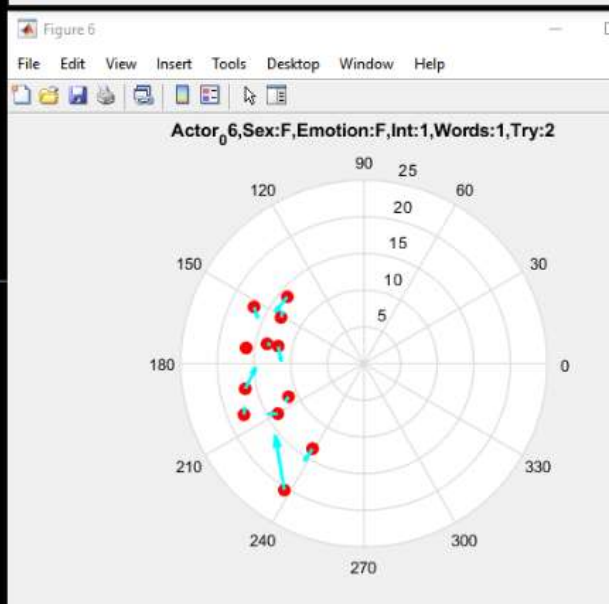




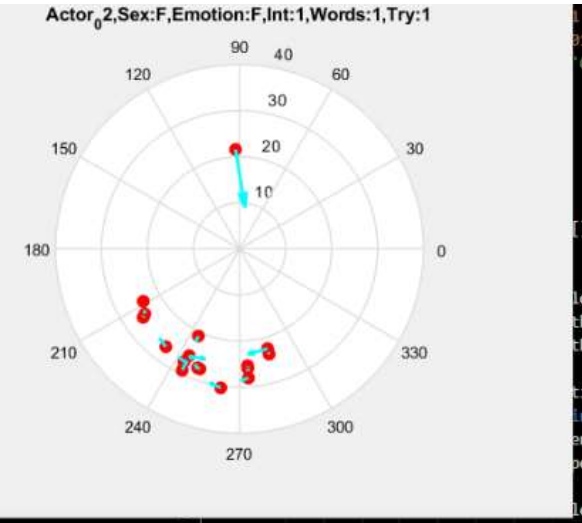


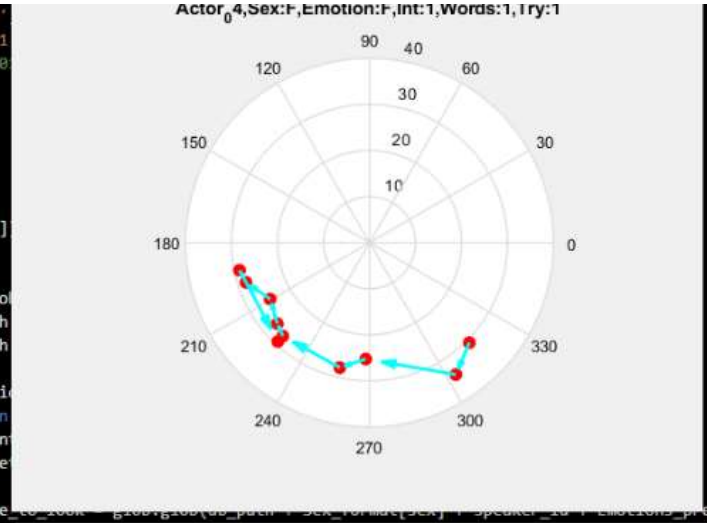


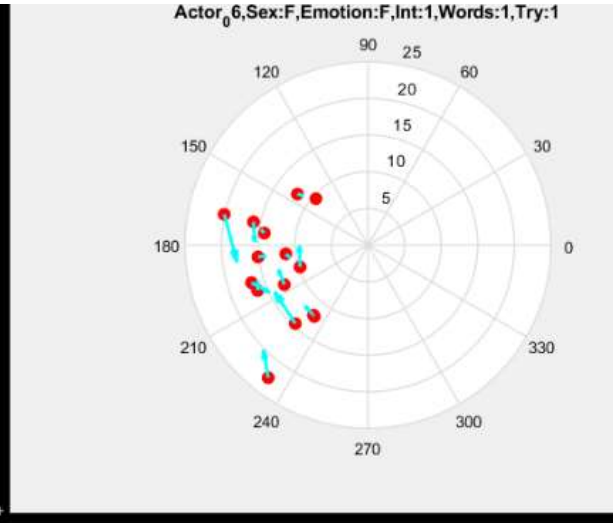


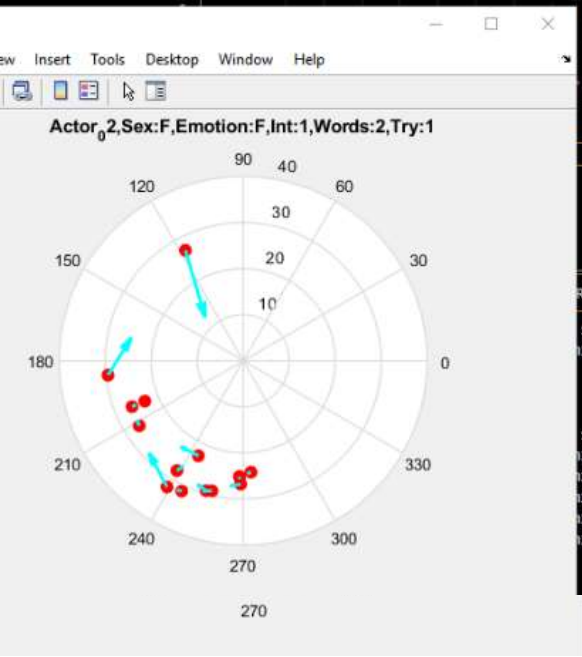


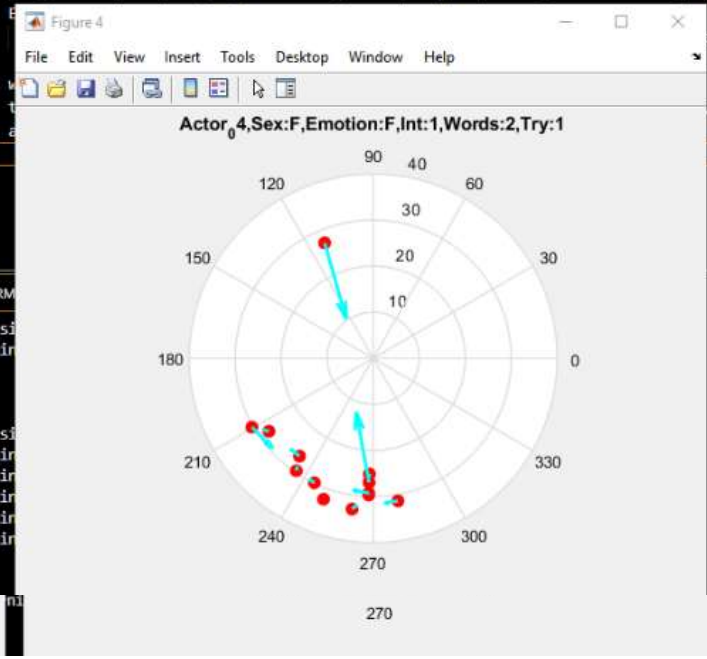
Explainable Patterns: Different words

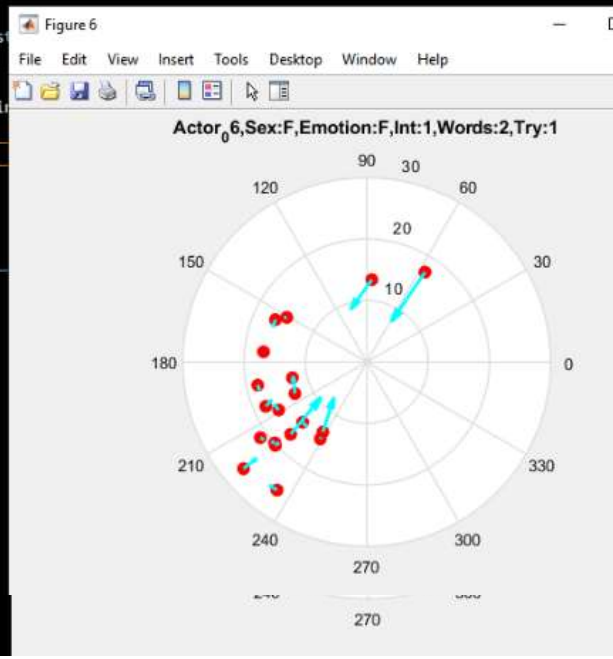


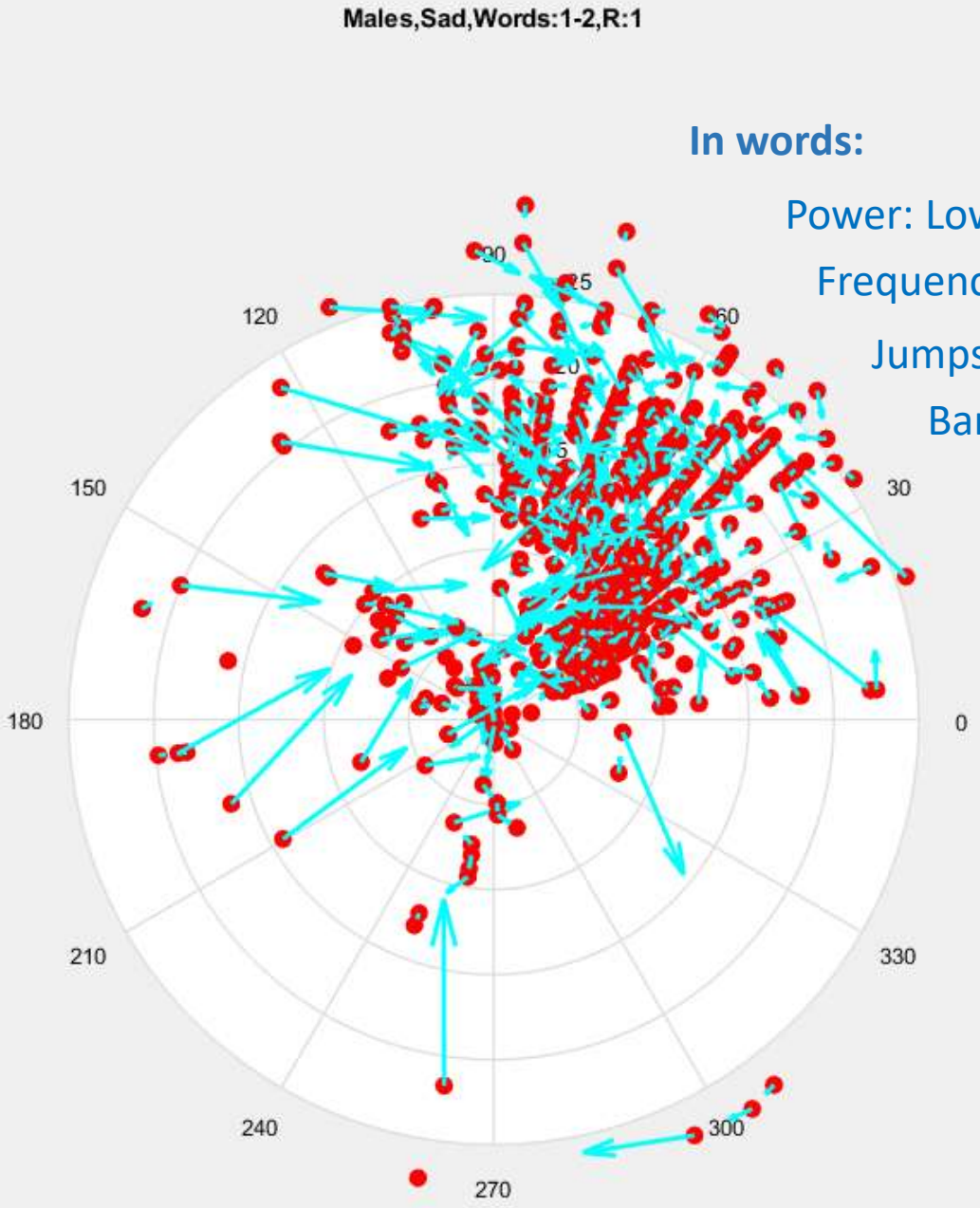












In words:

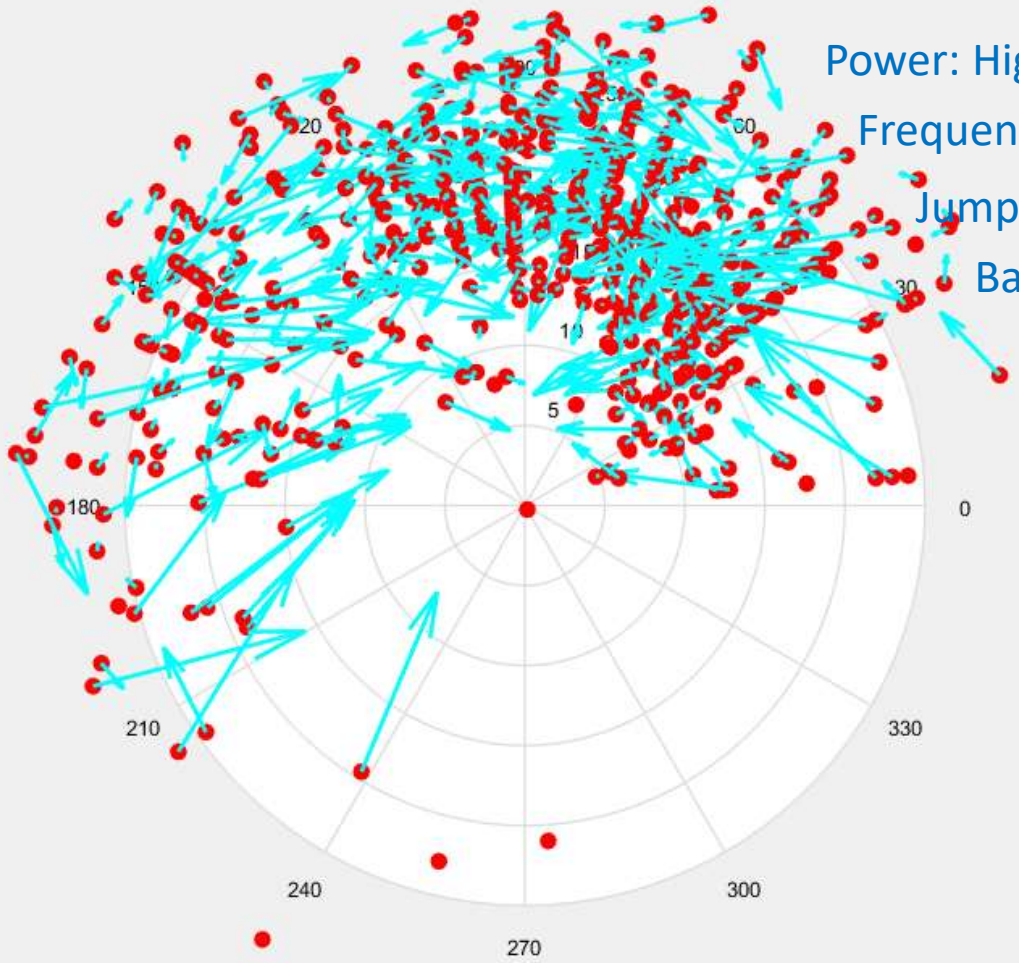
Power: Low

Frequency: Low

Jumps: Small

Bandwidth: Narrow

Males,Happy,Words:1-2,R:1

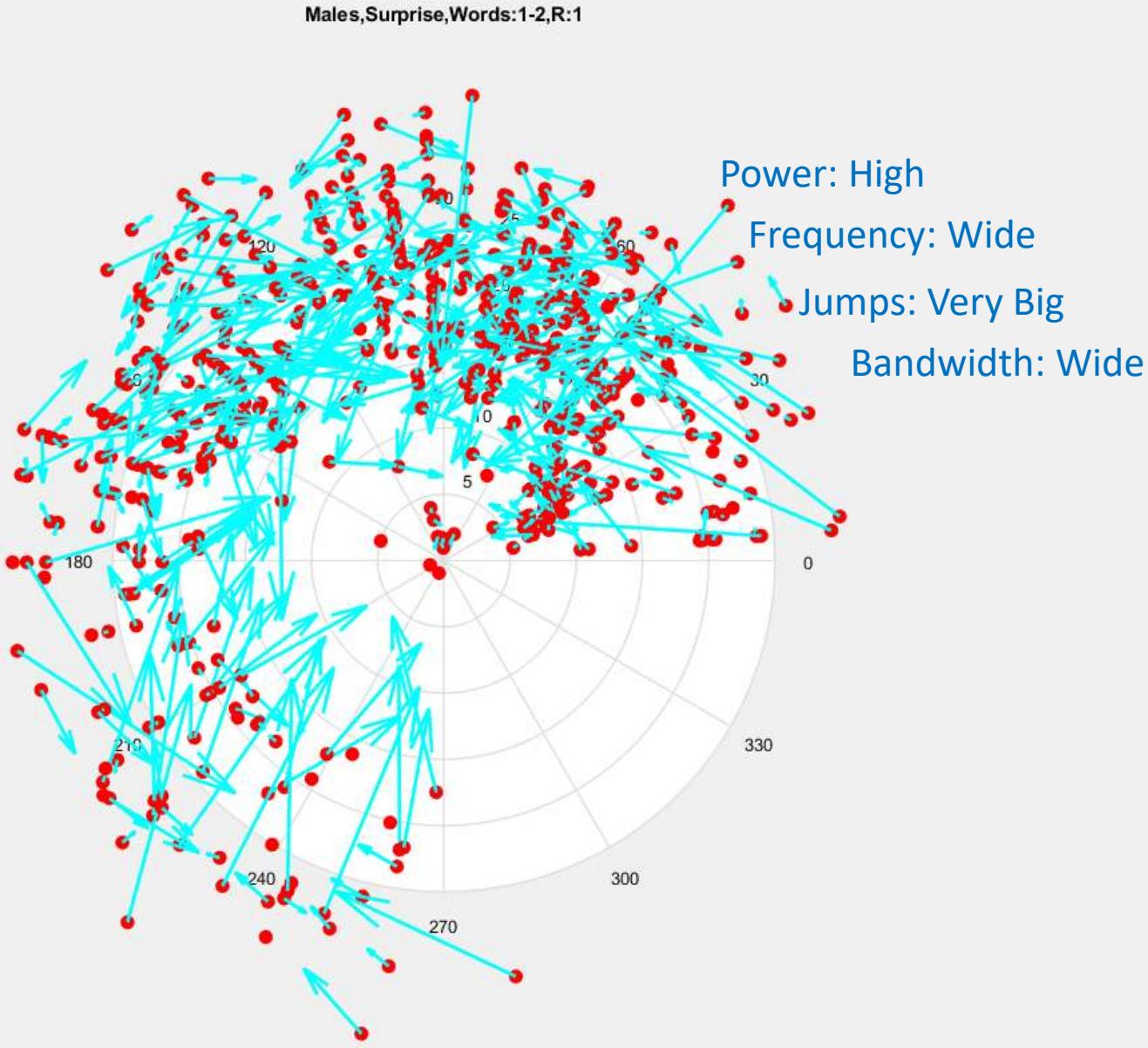


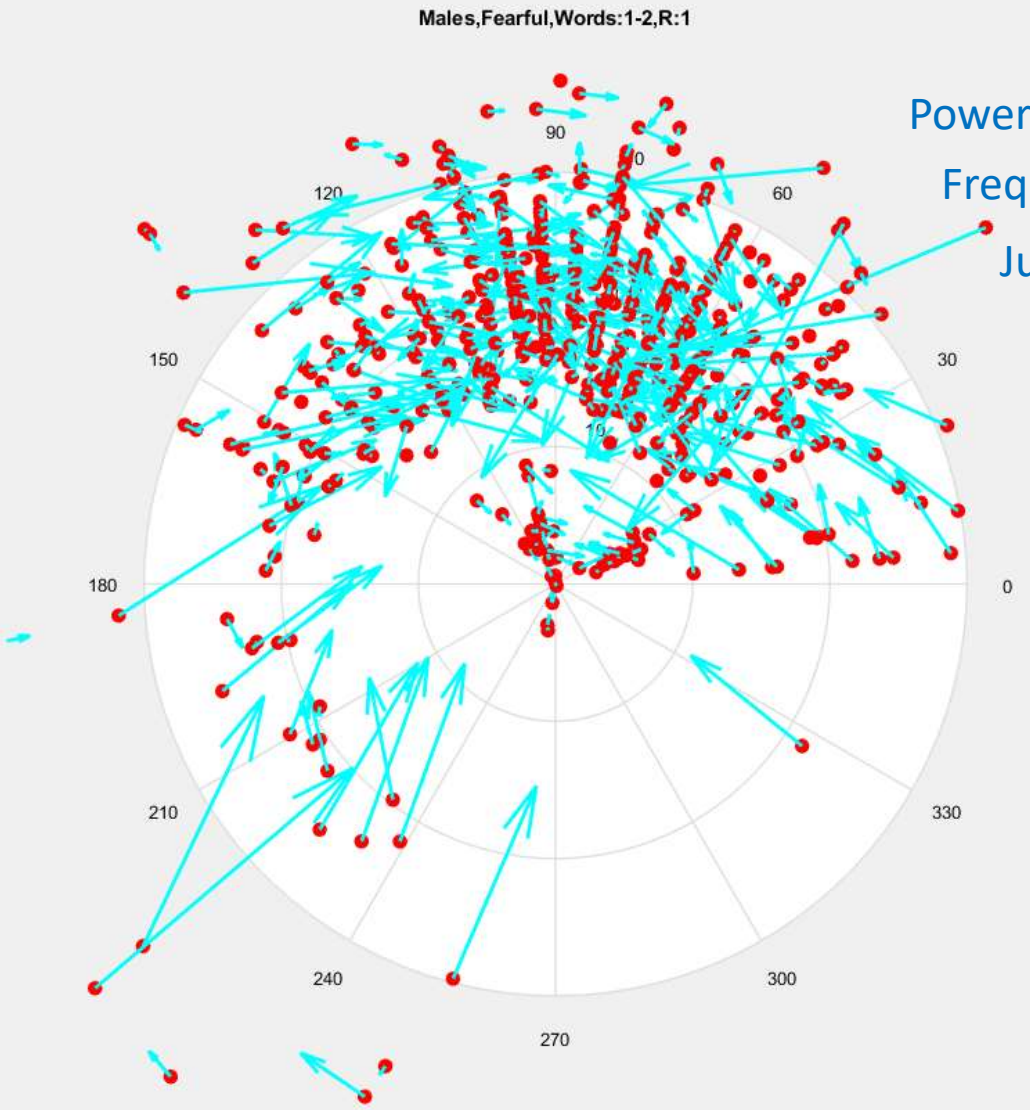
Power: High

Frequency: Wide

Jumps: Big

Bandwidth: Wide


Males,Surprise,Words:1-2,R:1
Power: High
Frequency: Wide
Jumps: Very Big
Bandwidth: Wide
120
60
30
0
330
300
270
240
210
180
5
10

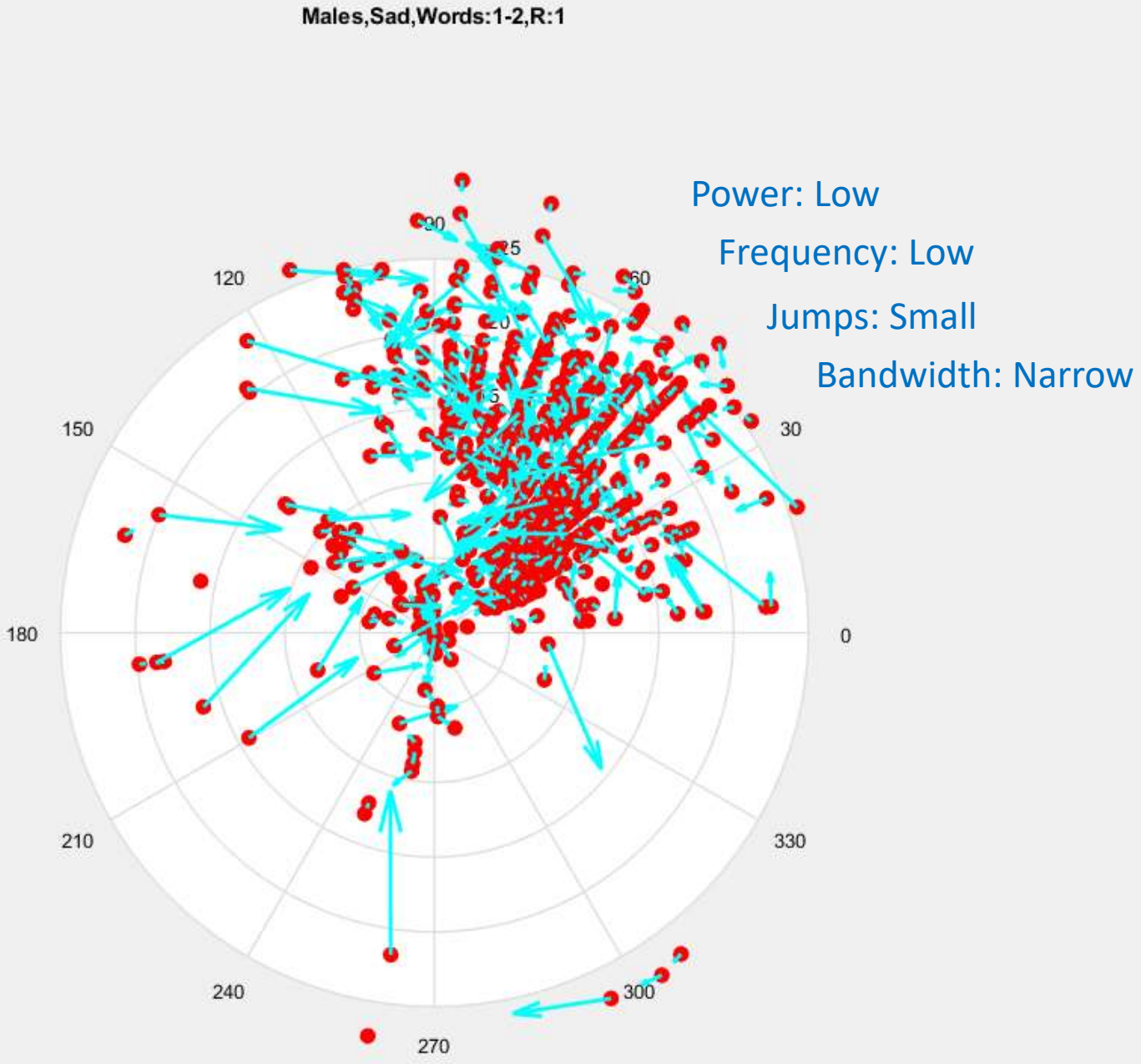


Power: Low

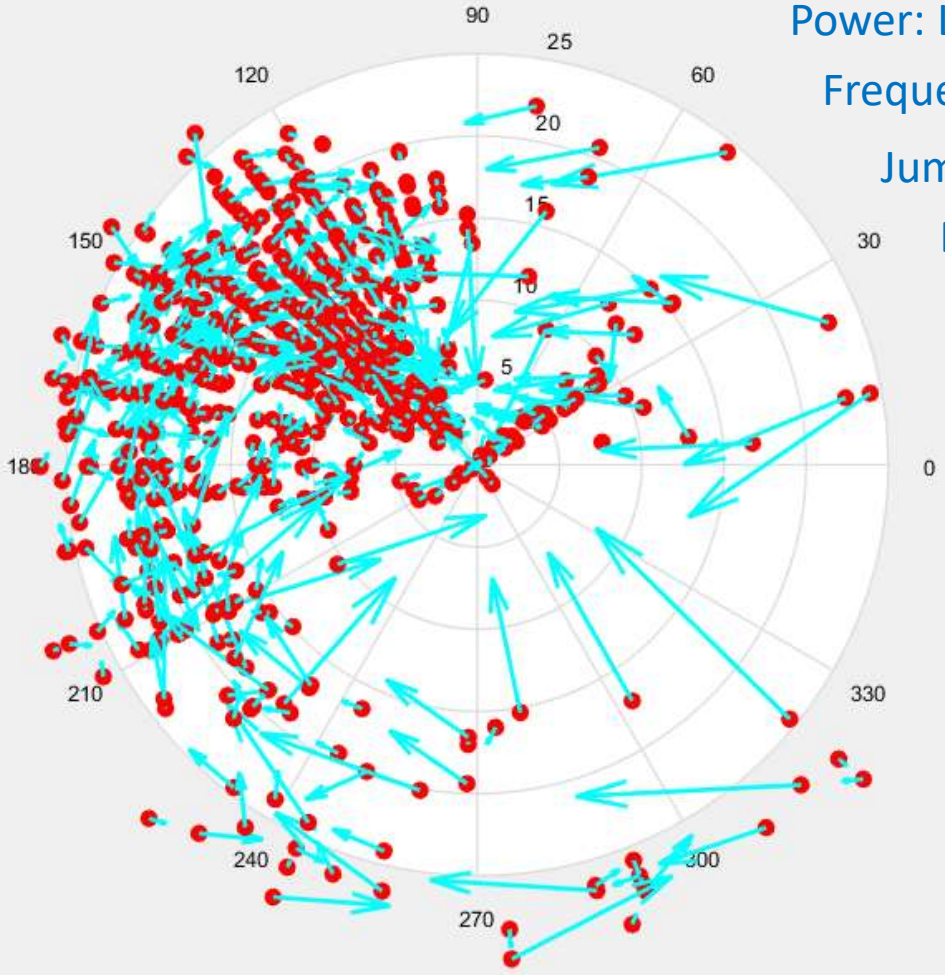
Frequency: Low

Jumps: Big

Bandwidth: Wide


Males,Sad,Words:1-2,R:1
90
120
60
150
30
180
0
210
330
240
300
270
Power: Low
Frequency: Low
Jumps: Small
Bandwidth: Narrow

Females,Sad,Words:1-2,R:1

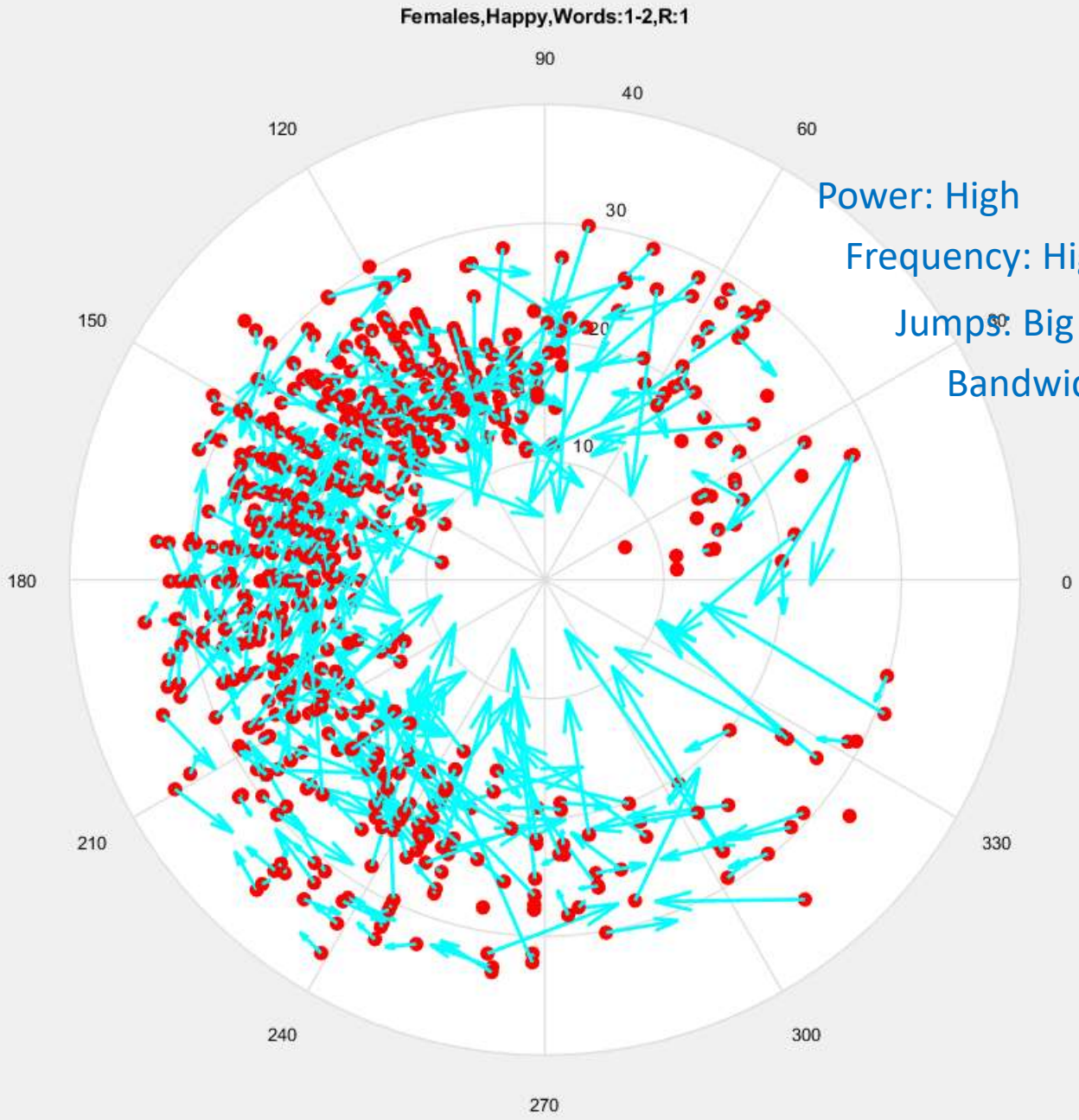


Power: Low

Frequency: High

Jumps: Small

Bandwidth: Narrow

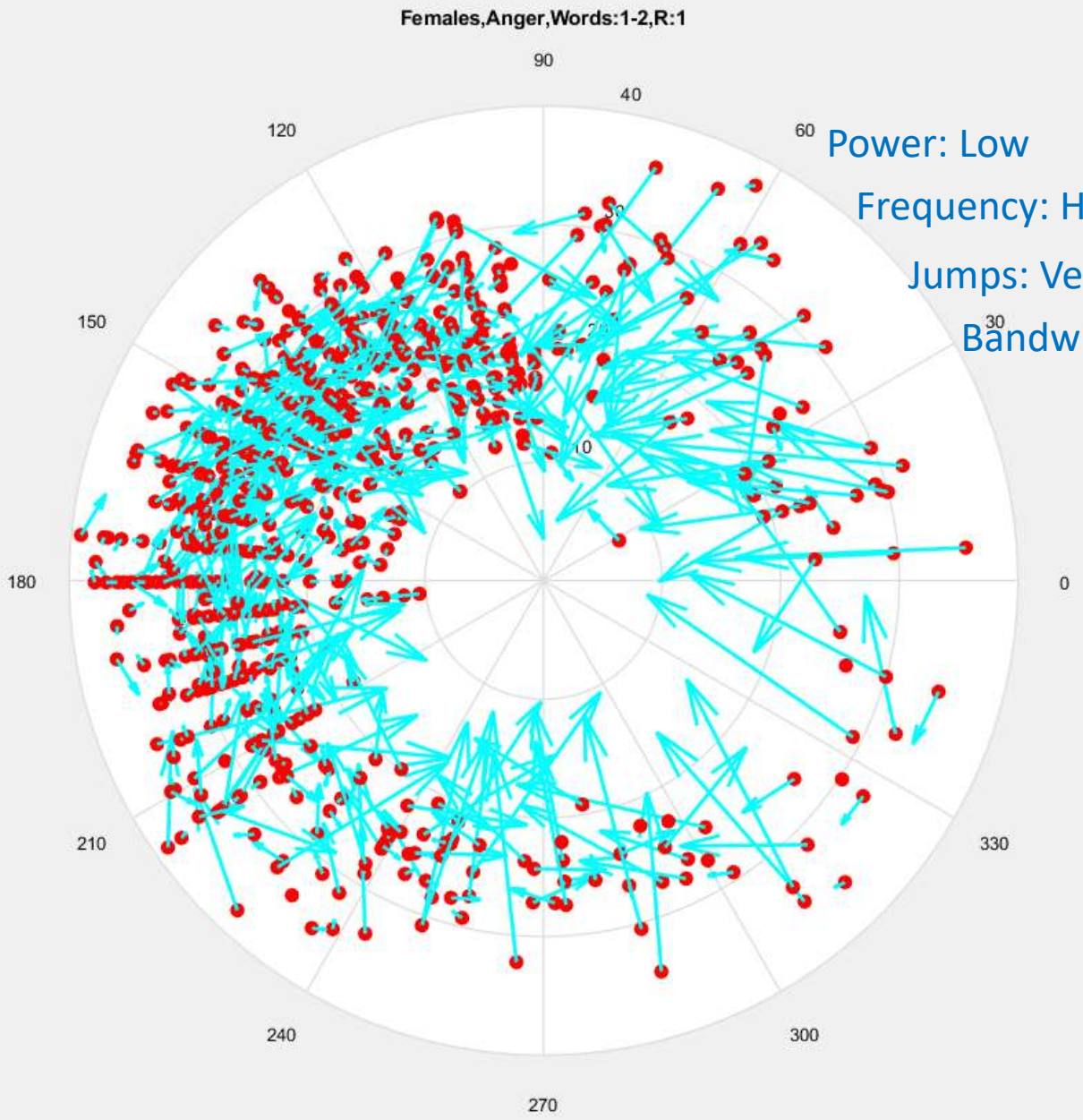


Power: High

Frequency: High

Jumps: Big

Bandwidth: Wide

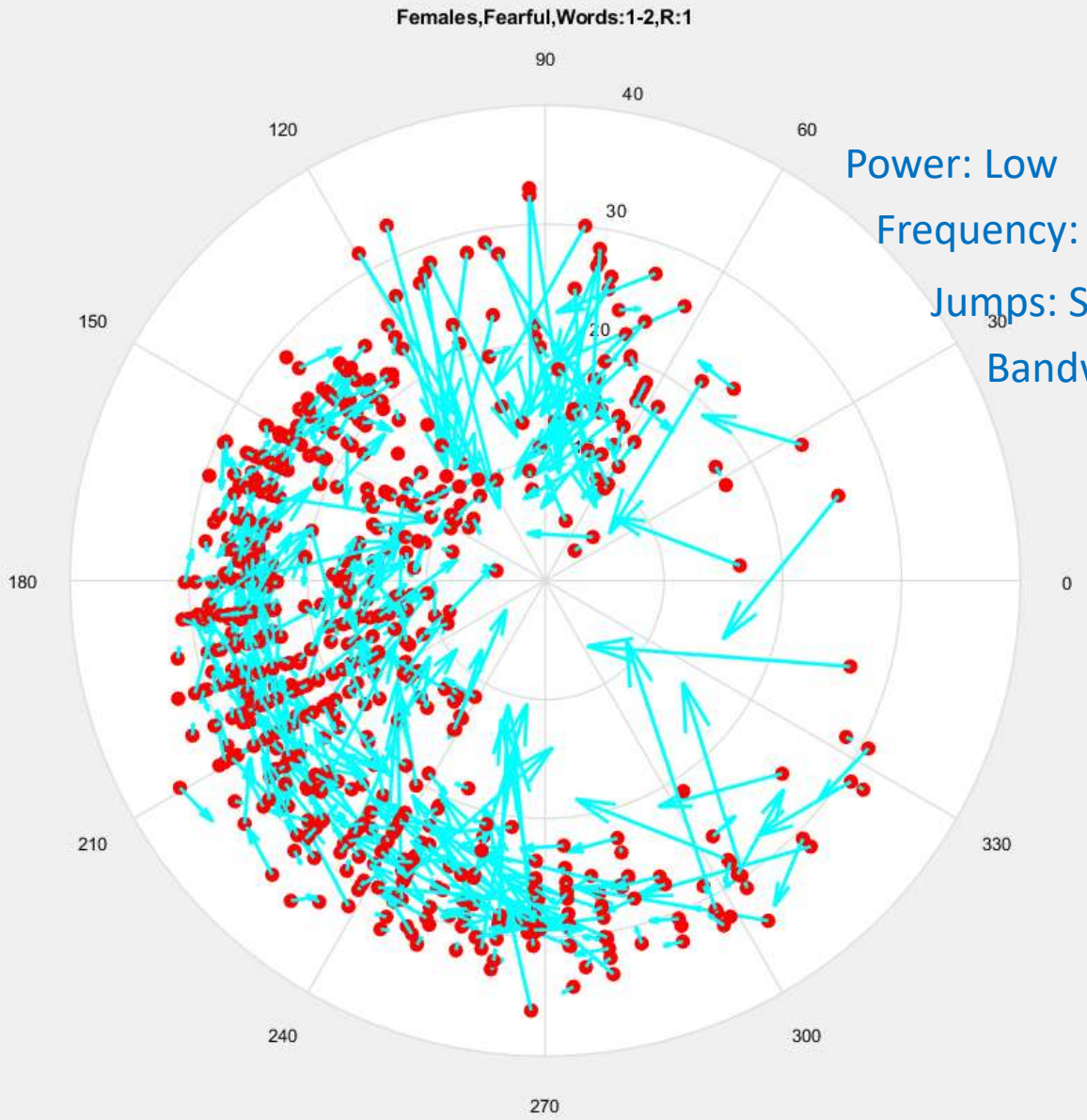


Power: Low

Frequency: High

Jumps: Very Big

Bandwidth: Wide

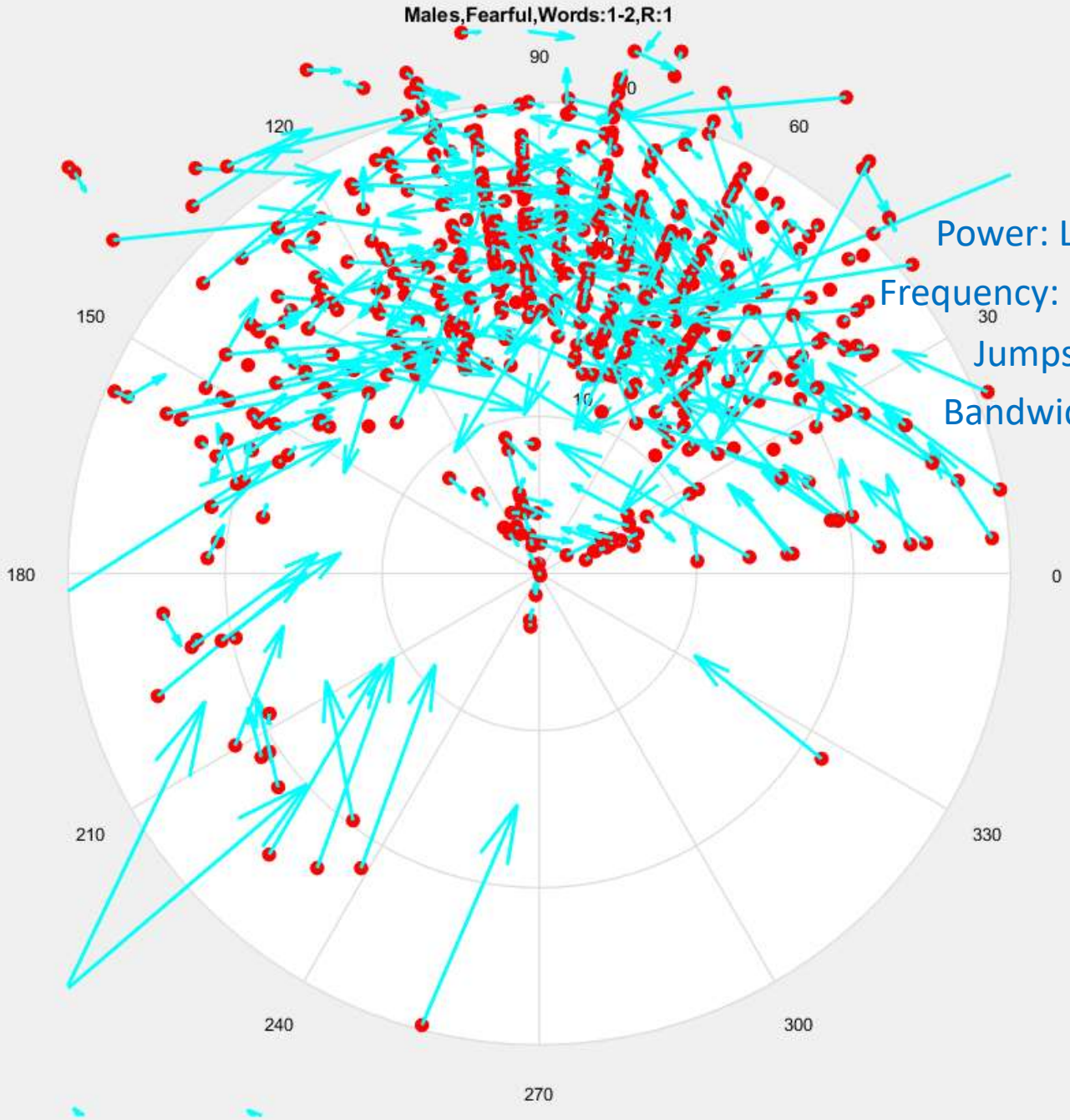


Power: Low

Frequency: Very High

Jumps: Small

Bandwidth: Wide


Males,Fearful,Words:1-2,R:1
90
120
60
150
30
180
0
210
330
240
300
270
Power: Low
Frequency: Low
Jumps: Big
Bandwidth: Narrow

Thanks

- Try codes:
  - https://github.com/tabahi/SpeechNotebooks/tree/main/Chapter01
  - https://github.com/tabahi/formantfeatures

- Image sources for ears and auditory perception:
  - https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/
  - http://www.speech.cs.cmu.edu/15-492/
  - https://www.inf.ed.ac.uk/teaching/courses/asr/lectures-2020.html
  - https://www.cse.iitb.ac.in/~pjyothi/cs753/index.html
  - https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20recognition%20course.html