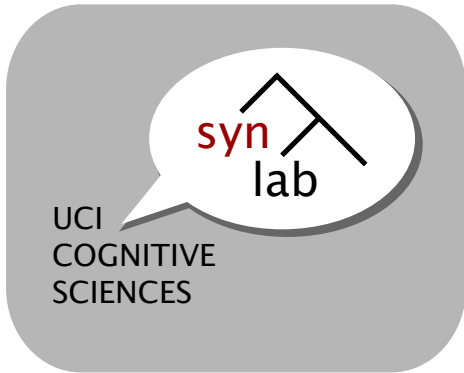


PSYCH 155/LING 155



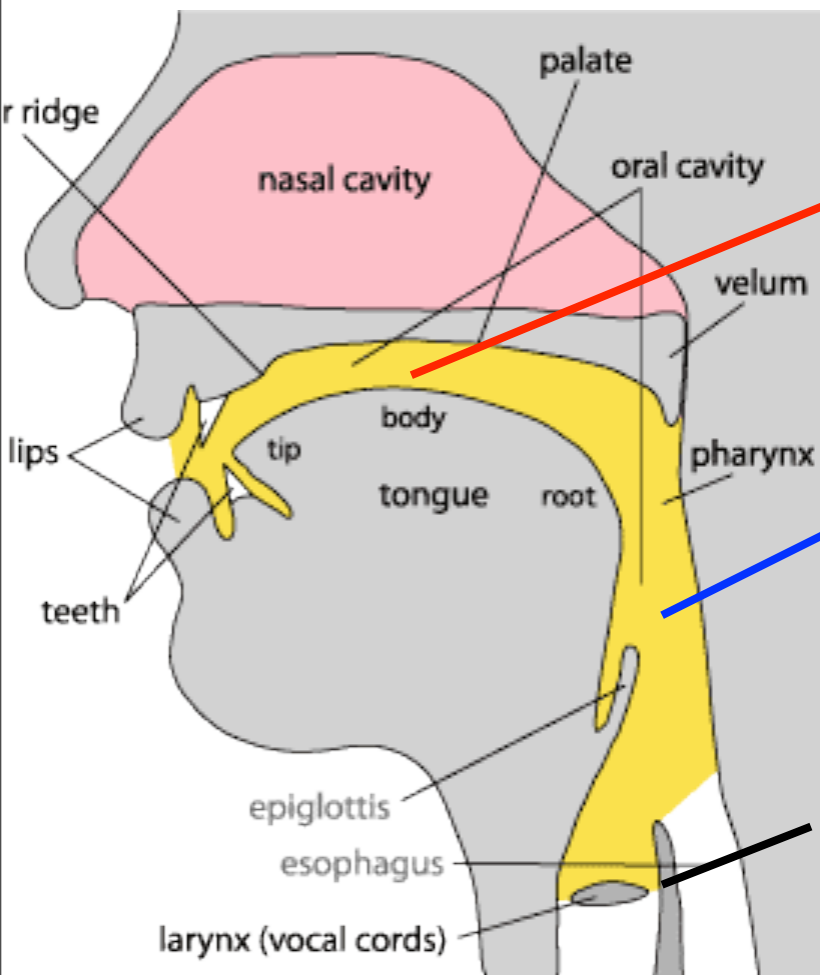
Psychology of Language

Prof. Jon Sprouse

Lecture 4: The Problem of Speech Perception

Quick Review: Formants

We call the lowest resonant band (that isn't the fundamental frequency) created by the trachea F1, and the lowest resonant band created by the oral cavity F2.



F2: the most prominent band in the mouth

The exact frequency of F2 will change based on the shape of the oral cavity

F1: the most prominent band in the trachea

The exact frequency of F1 will change based on the shape of the trachea

F0: the frequency created by the vocal folds

The Problem of Speech Perception

**sensory
representation**



Every property of sound:
-amplitude
-frequency (F0, F1, F2)



IPA CHART
IPA SYMBOLS

VOWELS

ɪ:	ɪ	ʊ	u:
SLEEP	SIT	BOOK	FOOT
e	ə	ɜ:	ɔ:
BED	AFTER	BIRD	PORT
æ	ʌ	ɑ:	ɒ
MAN	CUP	CAR	POT

DIPHTHONGS

ɪə	eɪ	
EAR	FLAY	
ʊə	ɔɪ	əʊ
FEWER	BOY	NO
eə	aɪ	aʊ
HAR	BUY	COW

CONSONANTS

p	b	t	d	tʃ	dʒ	k	g
PEN	BIG	TEN	DOG	CHURCH	JUDGE	CAKE	GUEST
f	v	θ	ð	s	z	ʃ	ʒ
FAIR	VERY	THREE	THIS	BUS	BUZZ	CASH	LEISURE
m	n	ŋ	h	l	r	w	j
MORE	NOW	SONG	HAPPY	LIP	ROSE	WAT	YES

Big Question:

Can we find a set of acoustic properties that **uniquely define** each sound in our language?

Quick Review: Amplitude and F0 don't seem to matter for speech perception

Of the list of properties of sound, it seems that only the **Formants** (the two lowest resonance bands created by the vocal tract) seem to affect speech perception

1. ~~Amplitude (which we perceive as loudness)~~

This just changes loudness

2. Frequency (which we perceive as pitch)

~~a. Fundamental frequency~~

This just changes pitch

~~b. Harmonics~~

This can't be changed alone

c. Resonance Bands

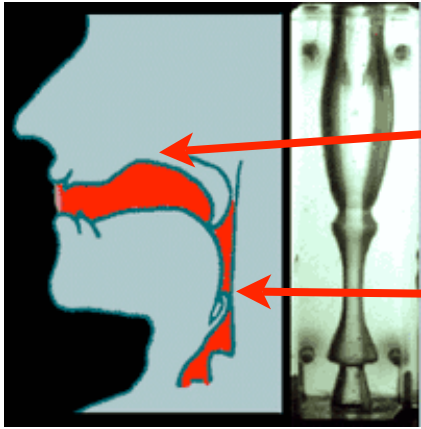
We call these **formants**

We can see that resonance is the crucial property with a simple experiment:

1. Create resonant bodies of different shapes (thus different resonance bands)
2. Ask yourself if the different shapes lead to the perception of different speech sounds

A demonstration that resonance directly influences speech perception

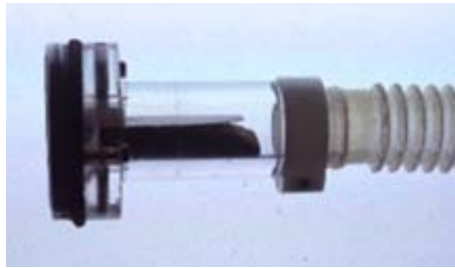
ah



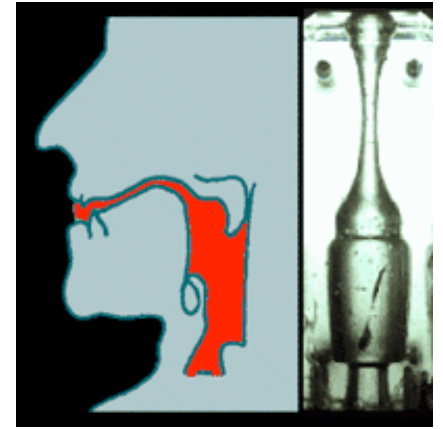
F2: oral cavity

F1: trachea

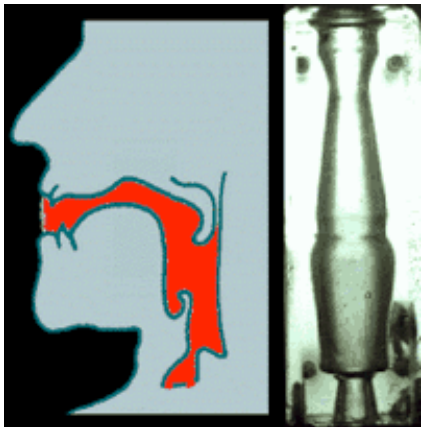
duck call



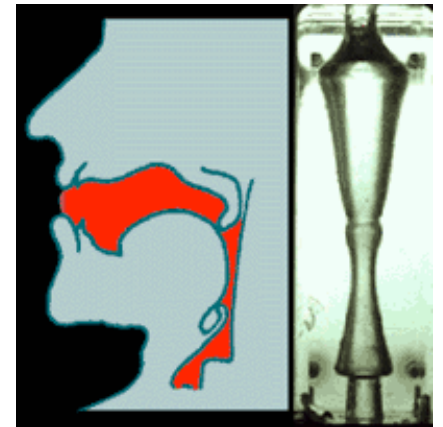
ee



eh



oh



[http://www.exploratorium.edu/
exhibits/vocal_vowels/
vocal_vowels.html](http://www.exploratorium.edu/exhibits/vocal_vowels/vocal_vowels.html)

The Problem of Speech Perception

starting
representation

sensory
representation



Every property of sound:
-amplitude
-frequency (F0, F1, F2)

acoustic
representation

F1: 400 Hz F2: 2700 Hz —

F1: 650 Hz F2: 2200 Hz —

IPA CHART
IPA SYMBOLS

VOWELS

i:	ɪ	ʊ	u:
SLEEP	SIT	BOOK	FOOT
e	ə	ɜ:	ɔ:
BED	AFTER	HERO	PORT
æ	ʌ	ɑ:	ɒ
MAN	CUP	CAR	POT

DIPHTHONGS

ɪə	eɪ	
EAR	FLY	
ʊə	ɔɪ	əʊ
FEWER	BOY	NO
eə	aɪ	aʊ
HIRE	BYE	COW

CONSONANTS

p	b	t	d	tʃ	dʒ	k	g
PEN	BIG	TEN	DOG	CHURCH	JUDGE	CAKE	GUEST
f	v	θ	ð	s	z	ʃ	ʒ
FAIR	VERY	THREE	THIS	BUS	BUZZ	CASH	LEISURE
m	n	ŋ	h	l	r	w	j
MORE	NOW	SONG	HAPPY	LIP	ROSE	WAT	YES

Big Question:

Can we find a set of acoustic properties that uniquely define each sound in our language?

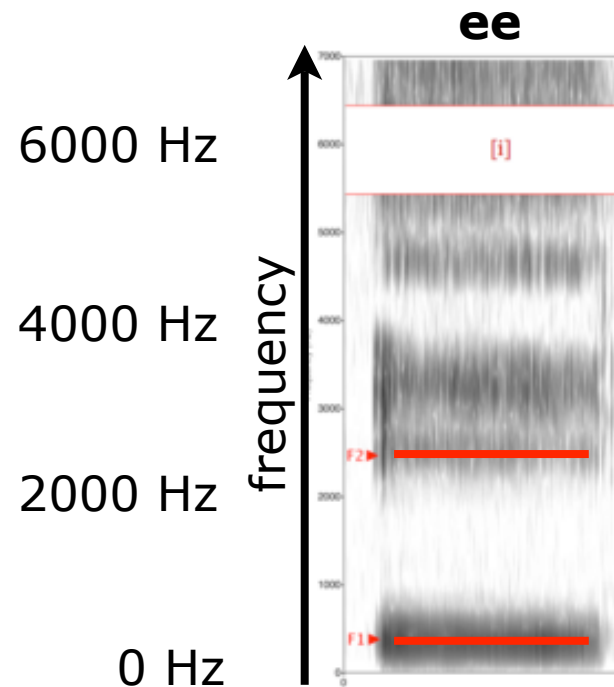
Measuring formants

We can visualize the formants of speech with a type of graph called a spectrogram:

A spectrogram plots **frequency** on the y-axis

And indicates increases in the energy (what we've been calling '**prominence**') with dark shading.

The dark shaded lines represent frequencies that have lots of energy.



Notice that each dark band occurs at a range of frequencies -- this is why we call them **frequency bands**.

The two lowest frequency bands are **F1 and F2**, which we can highlight in red.

A strong claim that we should test

Big Question: Can we find a set of acoustic properties that **uniquely define** each sound in our language?

Possible Answer: Each speech sound can be **uniquely defined** by the first two formants (**F1** and **F2**).

F1: 400 Hz **F2:** 2700 Hz —

F1: 650 Hz **F2:** 2200 Hz —

IPA CHART

IPA SYMBOLS

VOWELS

i: SLEEP	ɪ SIT	ʊ BOOK	u: FOOT
e BED	ə AFTER	ɜ: BIRD	ɔ: FORT
æ MAN	ʌ CUP	ɑ: CAR	ɒ POT

DIPHTHONGS

ɪə EAR	eɪ FLY	
ʊə FERRY	ɔɪ BOY	əʊ NO
eə HAR	aɪ BUY	aʊ COW

CONSONANTS

p PEN	b BO	t TEN	d DO	tʃ CHURCH	dʒ JUDGE	k CAKE	g QUEST
f FAR	v VERY	θ THREE	ð THIS	s BUS	z BUZZ	ʃ CASH	ʒ LEISURE
m MORE	n NOW	ŋ SING	h HAPPY	l LIP	r ROSE	w WAR	j YES

So if we figured out the formants for each sound, we will have solved the speech perception problem.

Prediction 1: Only F1 and F2 matter

If all speech sounds can be defined using F1 and F2, then the prediction is that **all we really need to perceive speech is F1 and F2.**

In other words, all of the other information in the auditory signal is irrelevant to speech perception. **We can test this by subtracting everything but F1 and F2 from a sentence and asking whether we can still perceive it.**

First, let's listen to each of the formants alone:

1. F1 of a complete sentence
2. F2 of a complete sentence

Notice that F1 and F2 in isolation don't really sound like speech. Instead, they sound like a series of changing tones

3. F3 of a complete sentence

Notice that F3 doesn't really change much at all

OK, before we combine them together, what do you think the sentence is?

Testing the prediction

Now listen to F1, F2 (and F3) played together:

“A bird in the hand is worth two in the bush”

Speech that is reduced to formants is sometimes called **sine-wave speech** because it can be created by synthesizing the simple sine-waves of each of formants.

The fact that sine-wave speech is intelligible is can be taken as evidence that the formants (F1, F2, F3) are critical for speech perception.

The converse fact is also true: if you take away the formants, the remaining sound is nothing at all like speech!

<http://hctv.humnet.ucla.edu/departments/linguistics/VowelsandConsonants/vowels/chapter7/abirdinthehand.html>

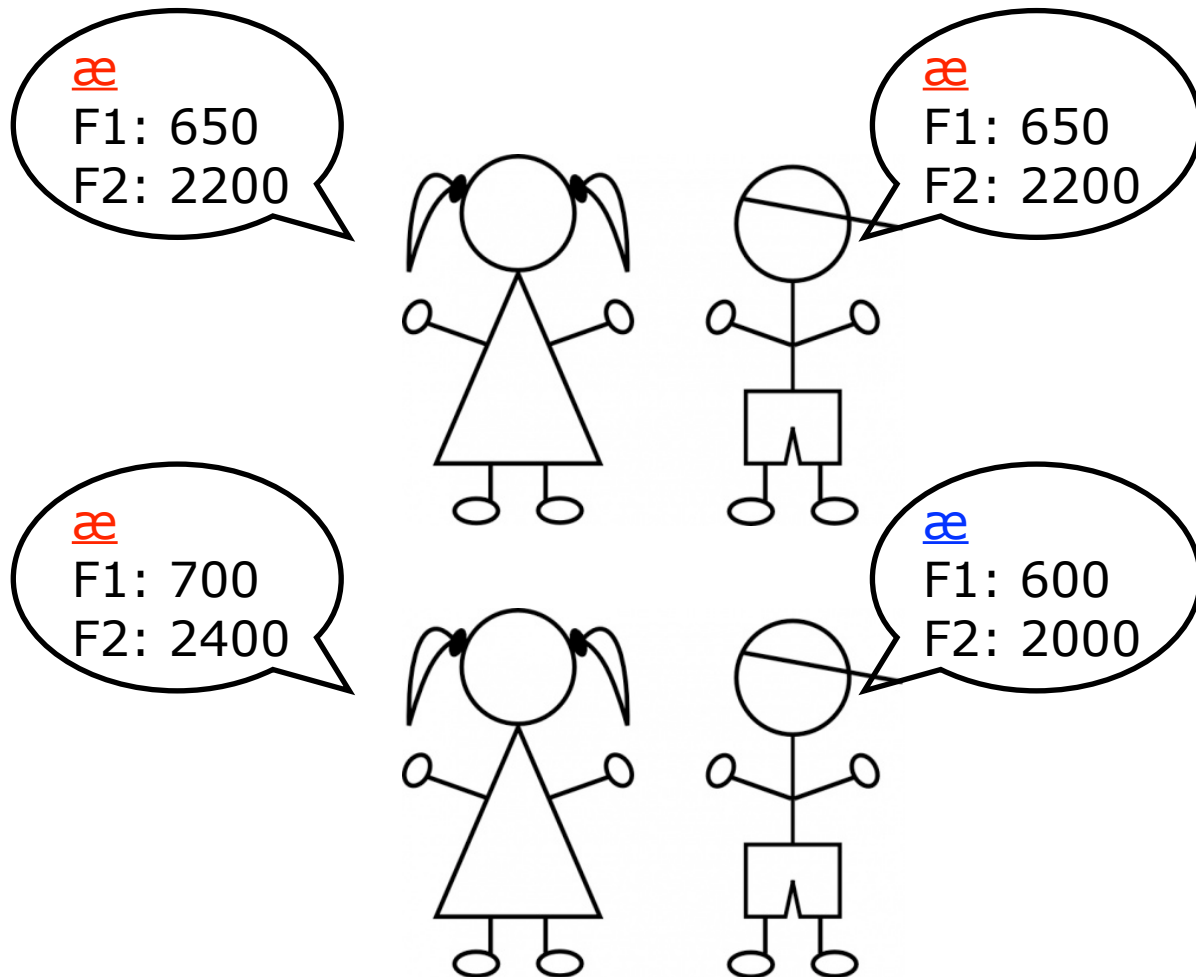
Let's make our definition more specific so we can test more predictions

The acoustic definition of speech sounds (F1 and F2) does **not** require a single definition for every speech sound.

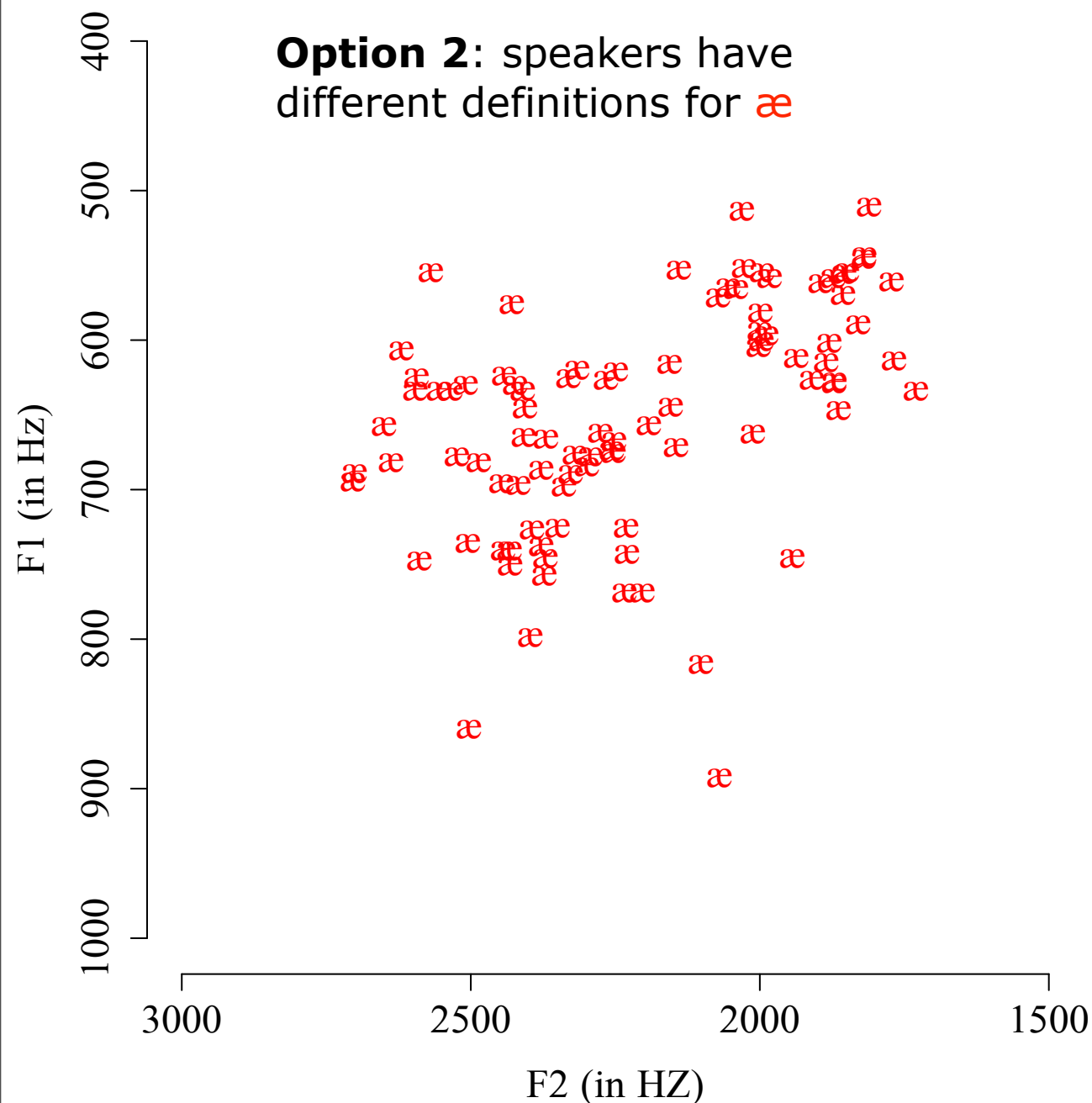
For example, the sound æ (as in "had") could be the same for all speakers, or it could vary a bit:

Option 1: all speakers have the same definition for æ

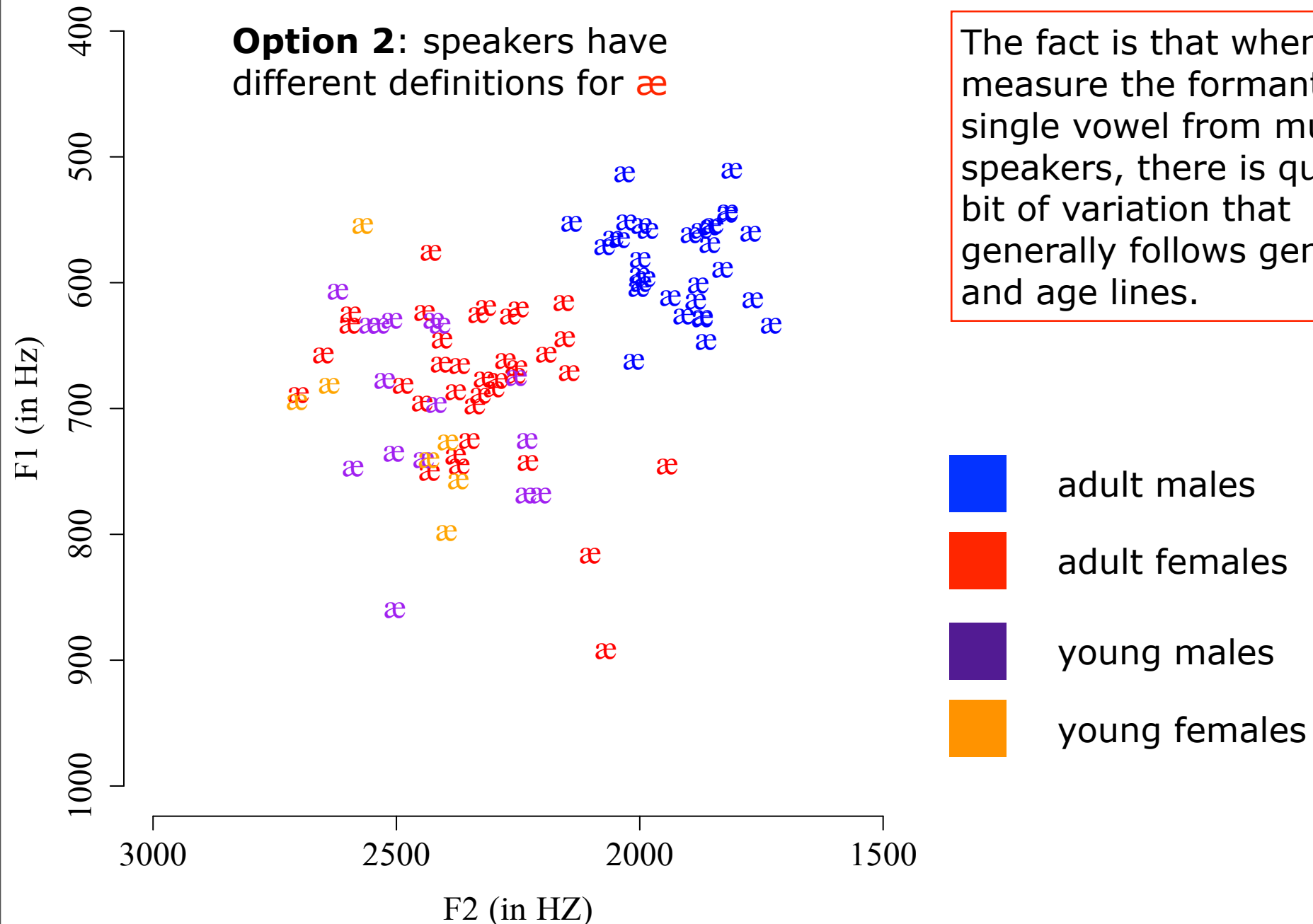
Option 2: speakers have different definitions for æ



Speakers have different definitions



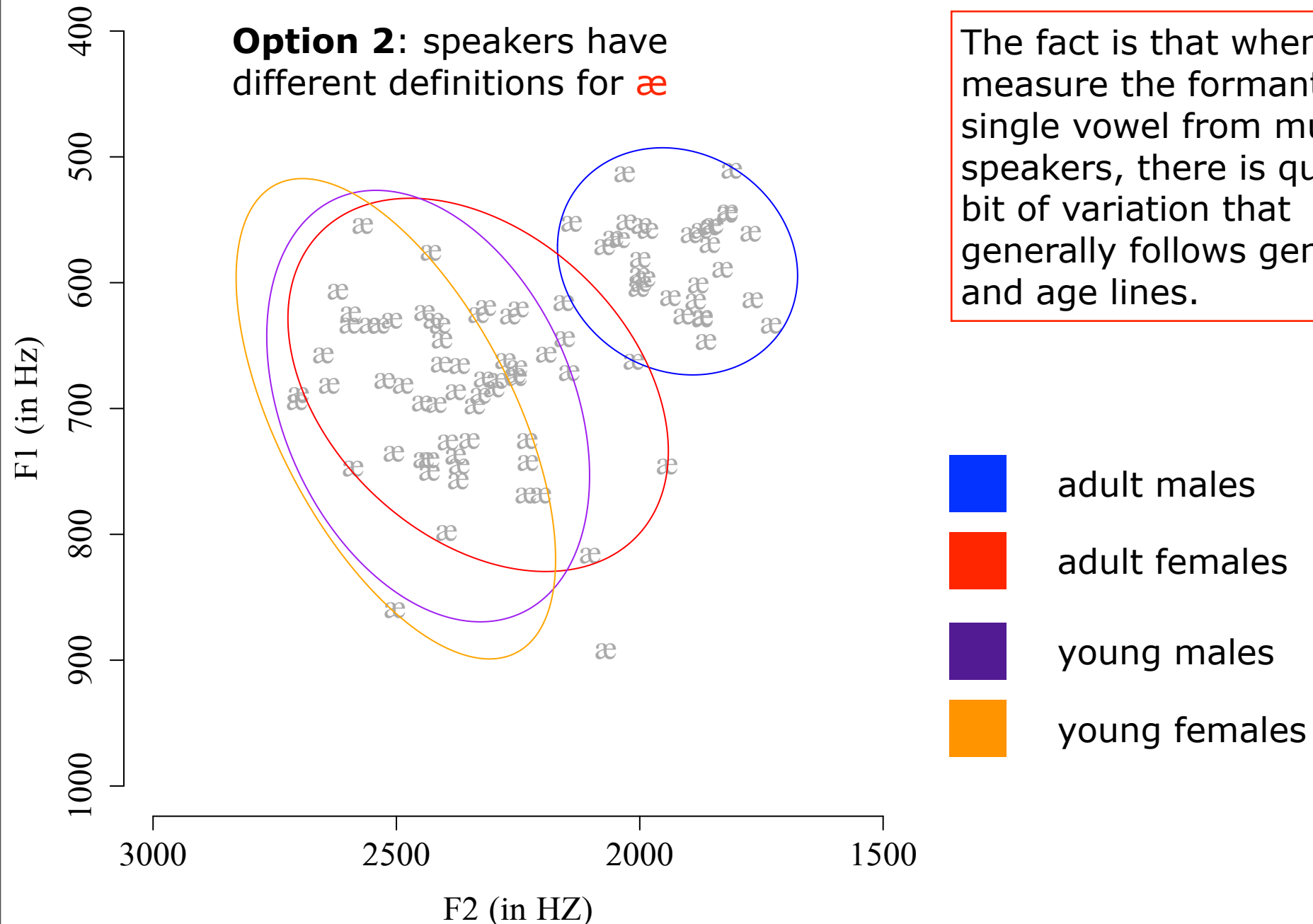
Speakers have different definitions



Speakers have different definitions

Option 2: speakers have different definitions for æ

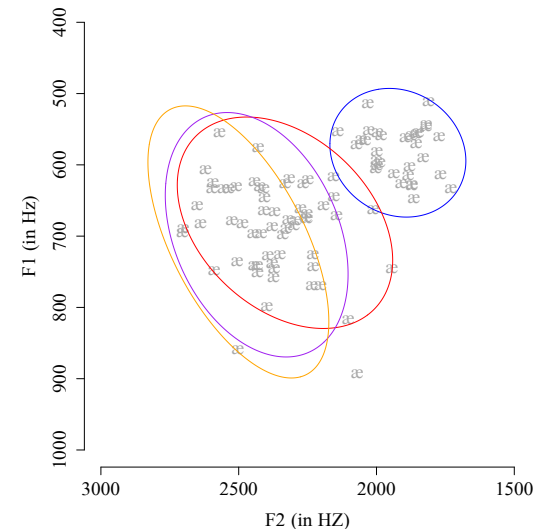
The fact is that when we measure the formants for a single vowel from multiple speakers, there is quite a bit of variation that generally follows gender and age lines.



We don't want two definitions for one vowel!

Remember that our goal is to find a **single definition** for each vowel based on its acoustic properties.

Before we abandon F1 and F2 as our definition, let's see if we can find a way to fix this problem.



To do this we need to know a bit more about what the auditory cortex can do... so that we know what types of solutions are at our disposal.

So now I am going to tell you about an interesting ability of the auditory cortex that may suggest a way to solve the gender/age problem above.

Restoration of the Missing Fundamental

Recall that vibrating objects have a **fundamental frequency** and an associated set of **harmonics** that are **integer multiples** of the fundamental frequency:

F0	200 Hz
<hr/>	
1st	400 Hz
2nd	600 Hz
3rd	800 Hz
4th	1000 Hz
5th	1200 Hz

We call a tone that contains multiple frequencies a **complex tone**, and a tone that contains a single frequency a **simple tone**.

In the case of complex tones that have a harmonic structure, we perceive the **pitch** of the tone as being equal to the pitch of the **fundamental frequency**.

We perceive the harmonics as **overtones**, which give the tone a **richer sound**.

Now, let's ask ourselves why we treat the fundamental differently than the harmonics (i.e., we perceive the pitch of the complex tone as equal to the fundamental, and not equal to the harmonics)?

Restoration of the Missing Fundamental

~~**Hypothesis 1:** It is simply because the fundamental has the highest amplitude (i.e., the loudest)~~

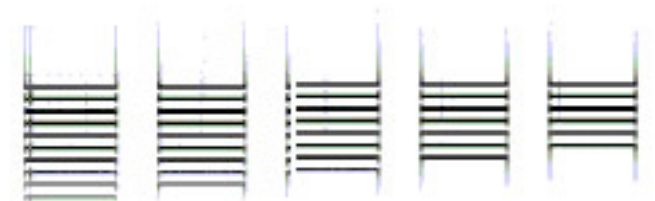
F0	200 Hz	100 dB
1st	400 Hz	50 dB
2nd	600 Hz	25 dB
3rd	800 Hz	12 dB
4th	1000 Hz	6 dB
5th	1200 Hz	3 dB

This hypothesis makes an interesting prediction:

If the crucial property is amplitude, then taking away the fundamental should change the **pitch** of the tone: the pitch should now be the frequency of the **1st harmonic**!

Similarly, if we take away both the F0 and the 1st Harmonic, the pitch should be based on the 2nd Harmonic.

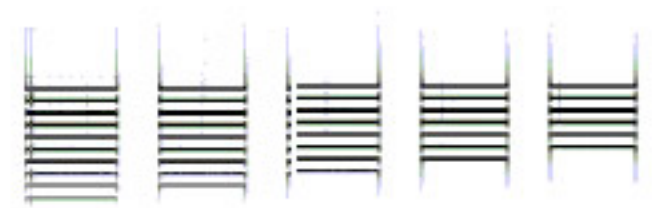
Here is a schematic of this test: each successive complex tone has the lowest frequency removed.



Restoration of the Missing Fundamental

Surprisingly, removing the lowest tone in these complexes **does not** change the **pitch** that we perceive. How can this be?

The answer seems to be that the brain **restores** the missing fundamental from a complex tone if that tone appears to have harmonic structure.



If this is just an illusion, it isn't very helpful.

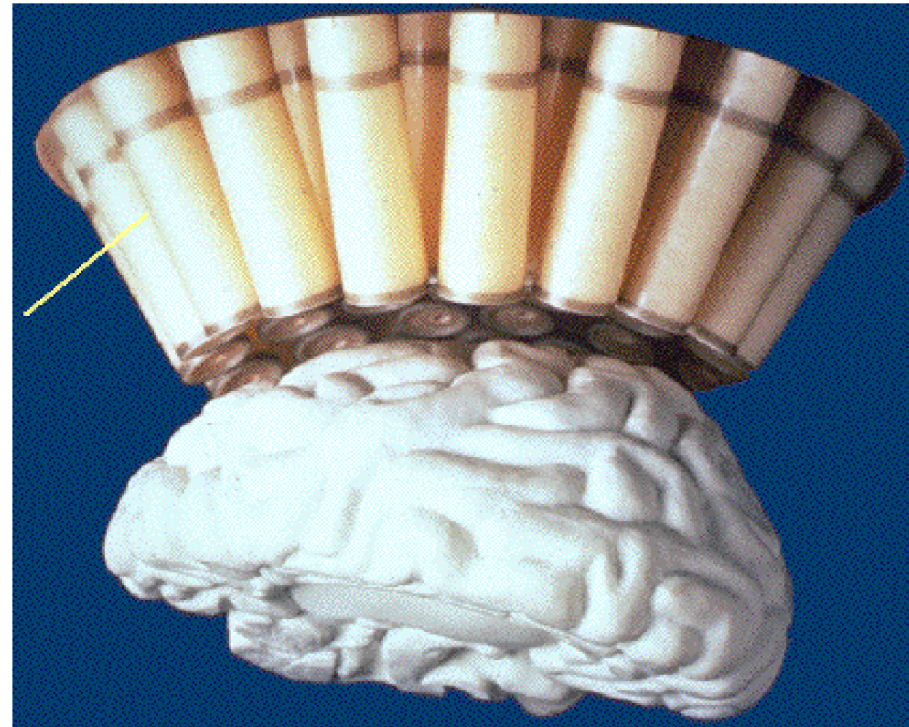
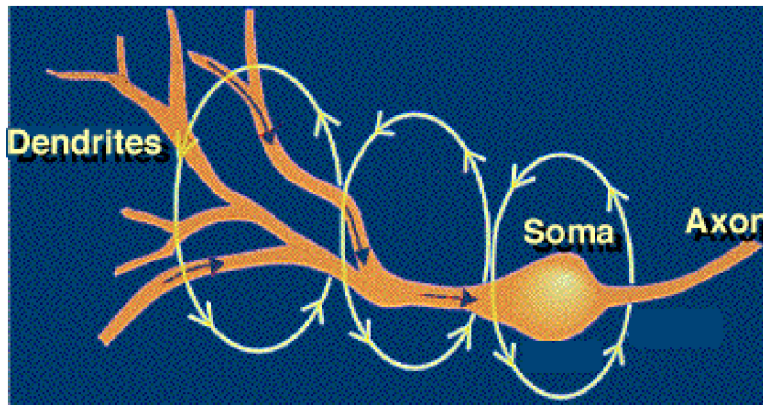
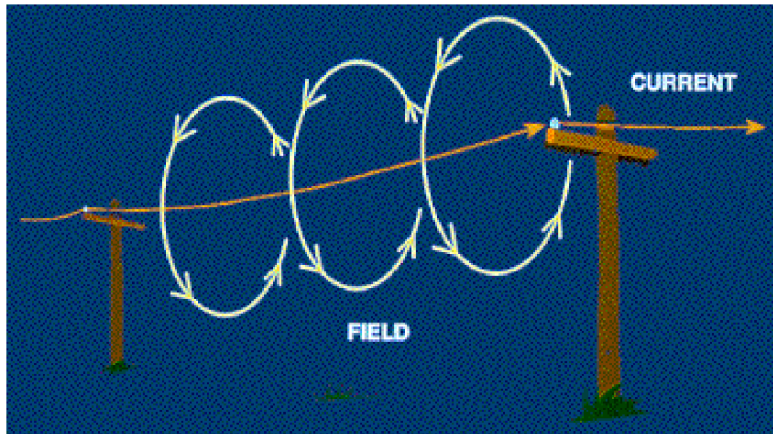
But if the auditory cortex can actually reconstruct the fundamental from the harmonics, then it tells us something about the abilities of the auditory cortex:

1. The auditory cortex may be able to perform calculations on the incoming signal in order to create new information that is not transparently available in the signal (i.e.,
2. The auditory cortex may be able to do some sort of mathematical factoring (or perhaps division) to figure out the common denominator in the tone complexes

Magnetoencephalography



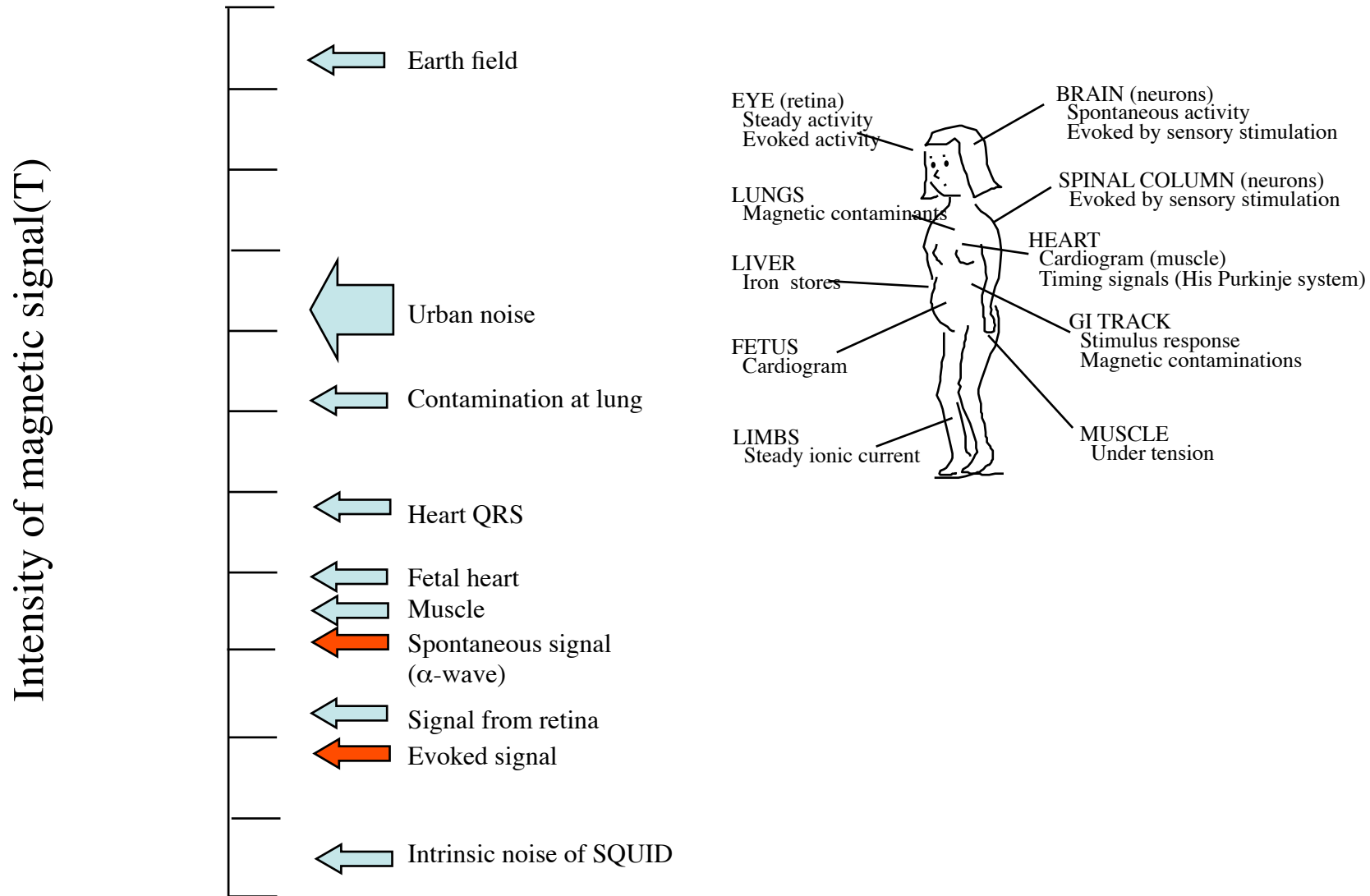
Magnetoencephalography



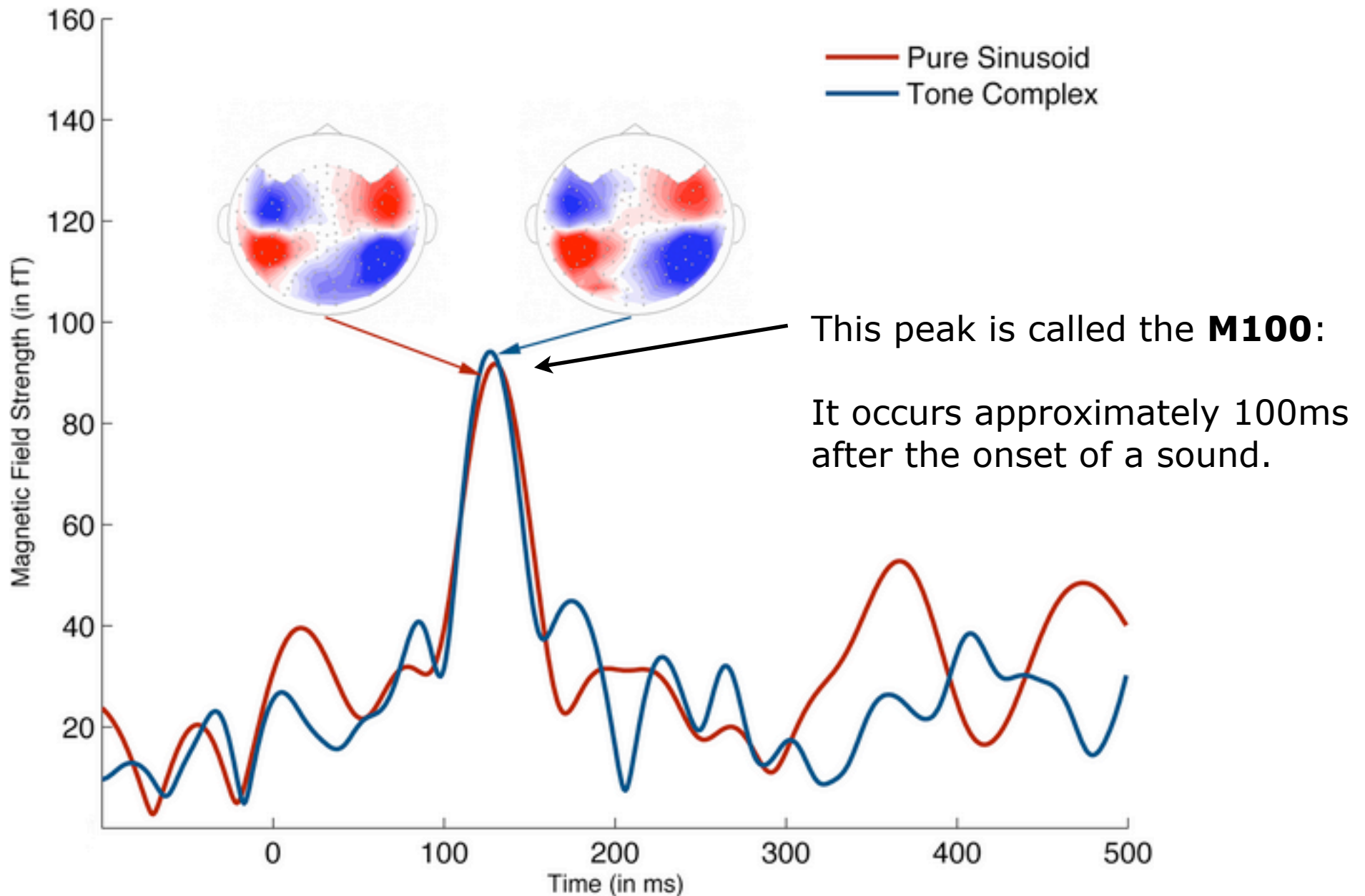
Brain magnetic fields recorded
fully non-invasively by arrays of
SQUID* detectors

[*Superconducting QUantum Interference Device]

Magnetoencephalography



Magnetoencephalography

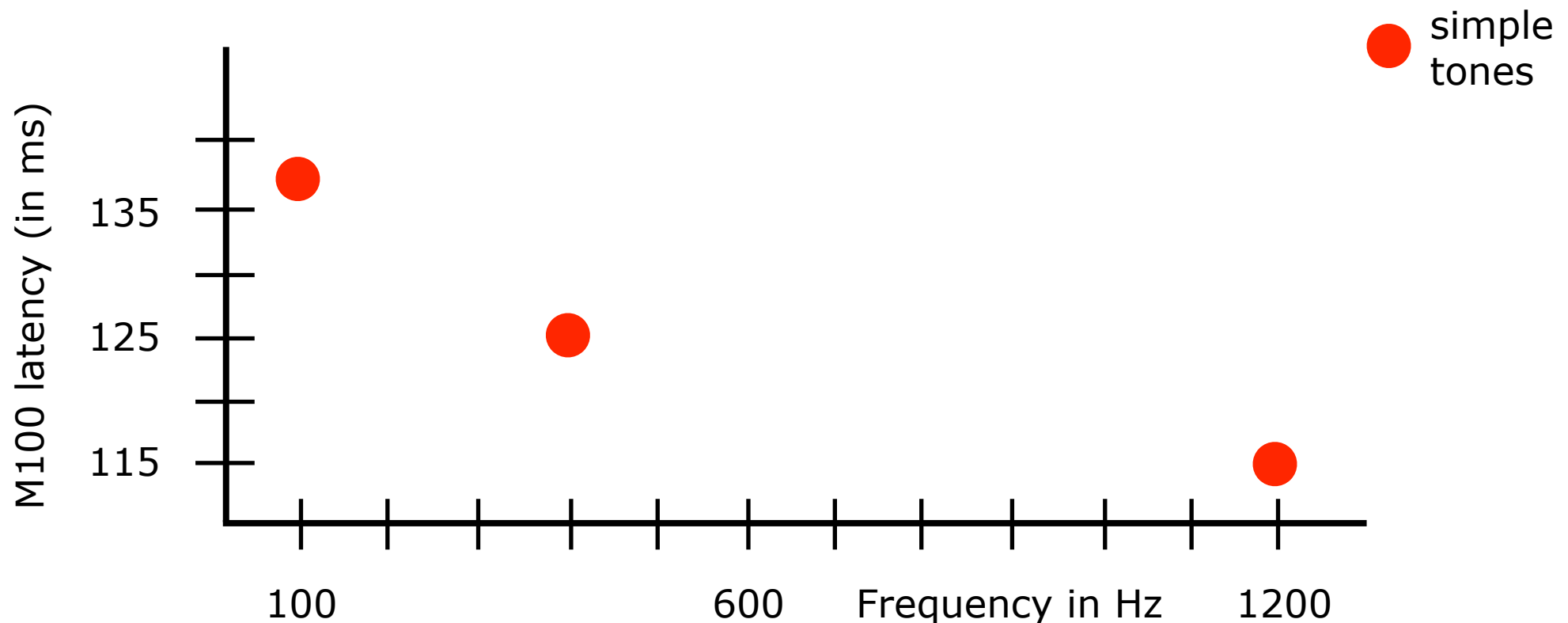


Brain evidence for F0 restoration

One interesting fact about the **M100** is that its **latency** (the time at which it occurs after the onset of a sound) changes based on the **frequency** of the sound.

Low frequencies have **longer** latencies (the M100 happens **later**)

Higher frequencies have **shorter** latencies (the M100 happens **earlier**)



Brain evidence for F0 restoration

Monahan et al. 2008 created complex tones with harmonic structure to test whether the M100 latency responded to the missing fundamental.

The lowest frequency in each complex tone was 1200Hz.

The highest frequency in each complex tone was 2400Hz.

The middle frequencies in each complex tone led to different restored F0s.

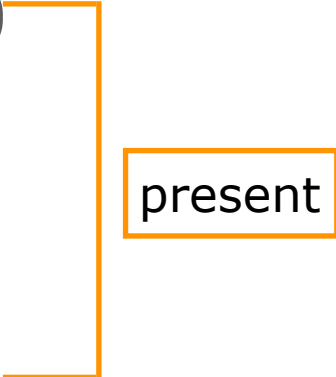
	Tone 1	Tone 2	Tone 3	
Restored F0:	100	400	1200	
lowest F:	1200	1200	1200] present
middle F:	1300	1600	--	
middle F:	2300	2000	--	
highest F:	2400	2400	2400	

not present

present

Brain evidence for F0 restoration

	Tone 1	Tone 2	Tone 3
Restored F0:	100	600	1200
lowest F:	1200	1200	1200
middle F:	1300	1800	--
middle F:	2300	1800	--
highest F:	2400	2400	2400



No restoration hypothesis:

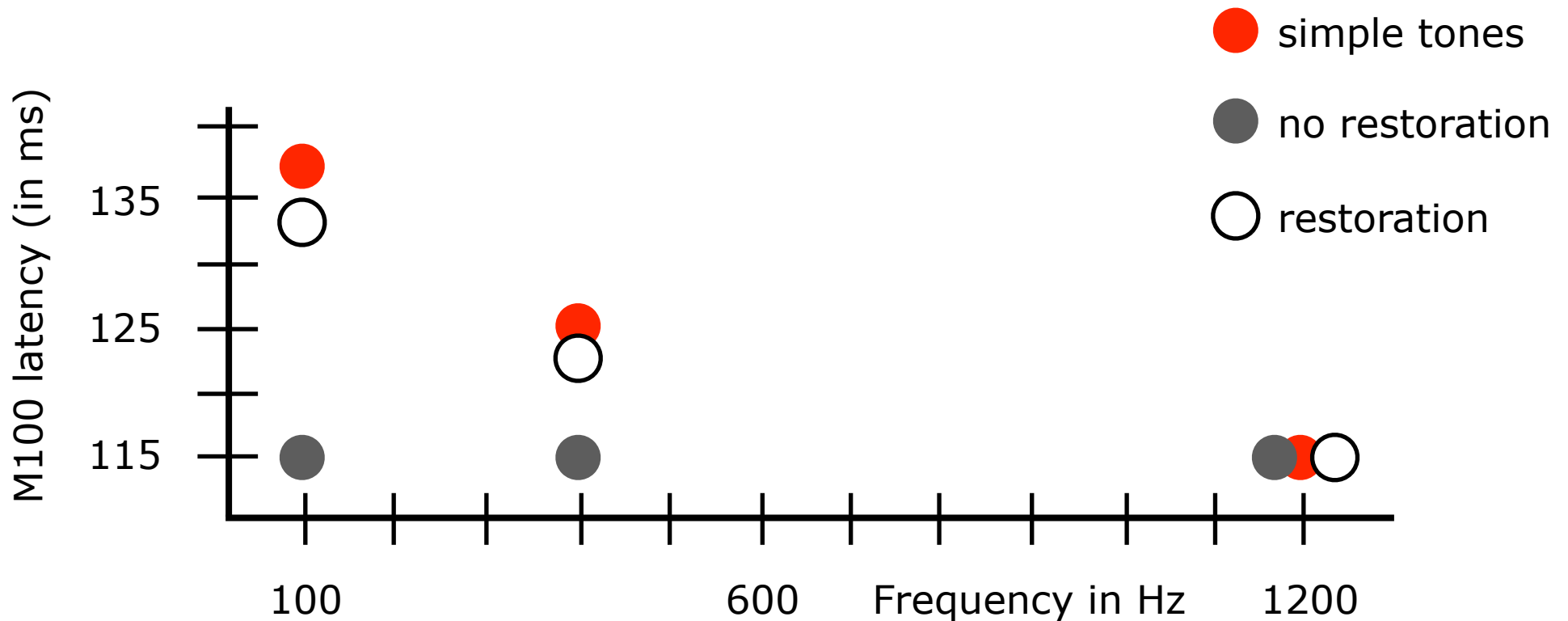
The lowest frequency in each complex tone was 1200Hz. So if the M100 simply responds to the lowest frequency in a complex tone, then the latency should never change.

Restoration hypothesis:

If the M100 responds to the restored fundamental, then each tone will have a different latency. Furthermore, the latency will be similar to the latency for simple tones at that the restored frequency.

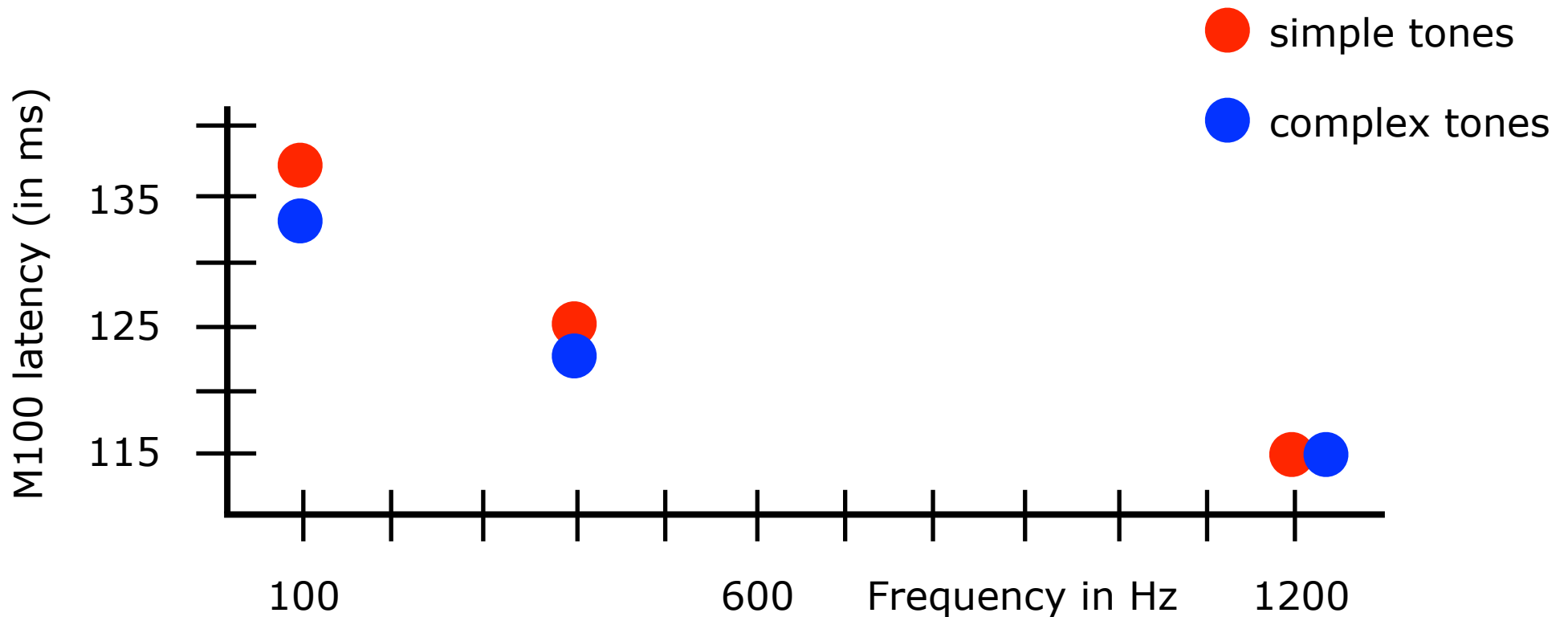
Brain evidence for F0 restoration

We can schematize the three predictions like this:



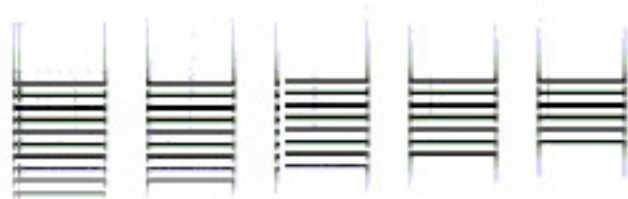
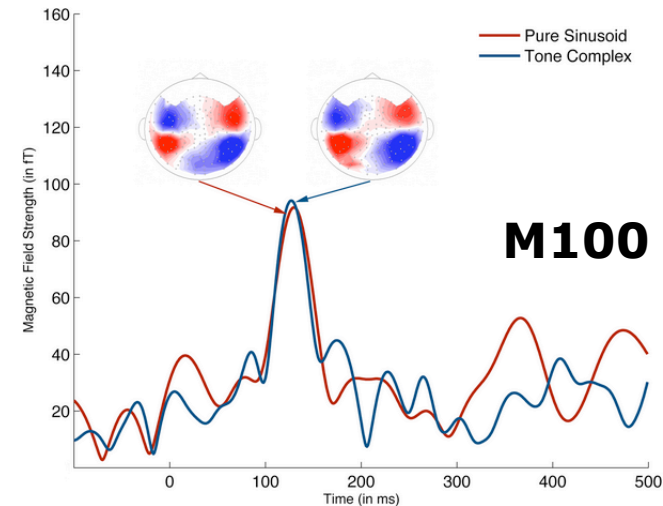
Brain evidence for F0 restoration

And the winner is..... the [restoration hypothesis](#)



What does this mean?

What we just saw is a **neuromagnetic response** to auditory processing (localized to the auditory cortex)...



... that appears to be sensitive to **restored fundamentals**.

In other words, a brain response to a sound **that isn't there!**

This suggests that the auditory cortex can make use of **information that is not present in the signal** based on simple calculations such as **factorization** (or **division**)

Some trivia for you...

Telephones only transmit a narrow band of frequencies 300Hz-3400Hz:

Telephone companies originally did this to save money on wires... but now they do this to save money on data transfer rates.

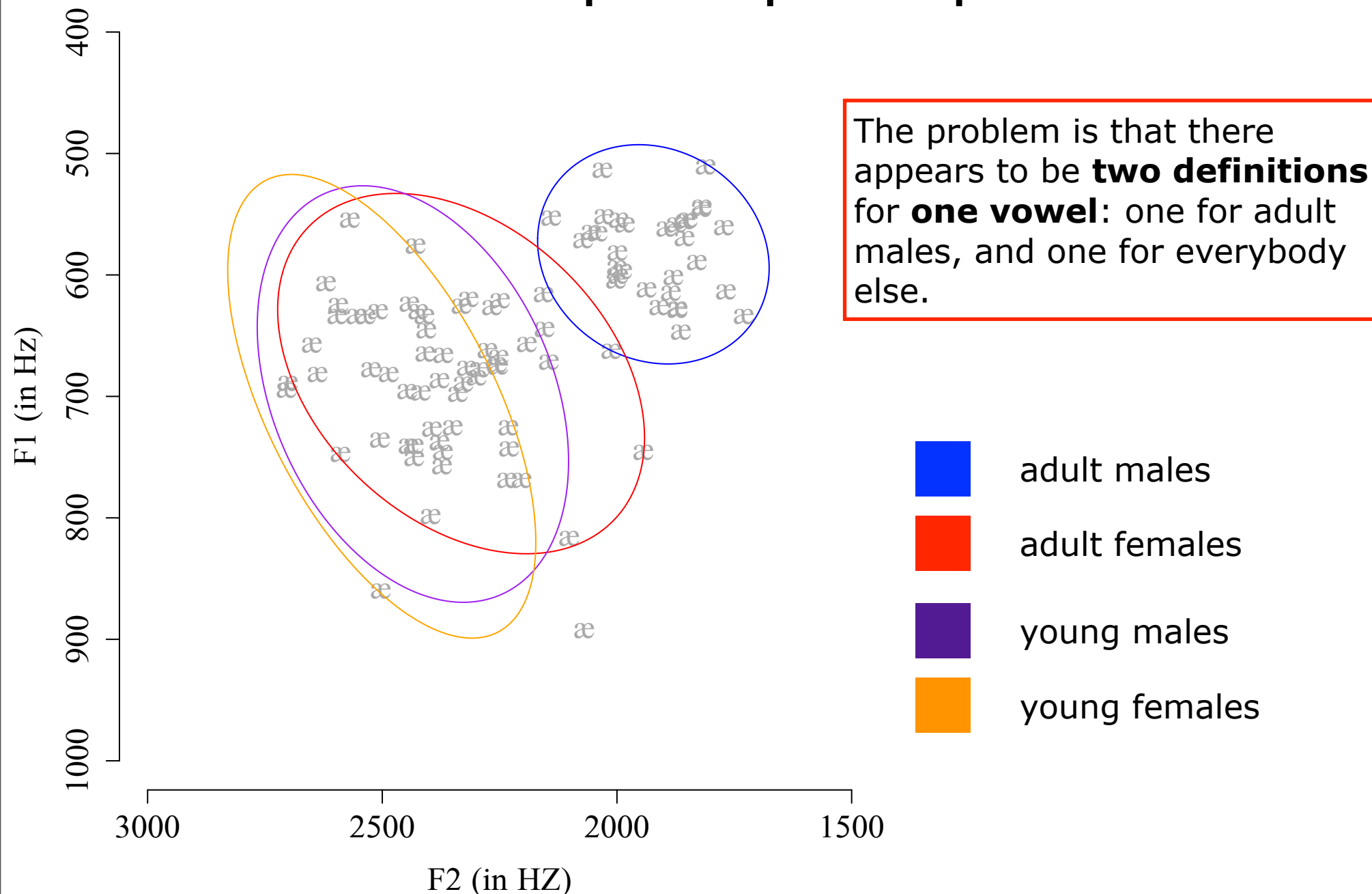


Recall what I said about the F0 human voice: the male average is 130Hz, the female average is 220Hz.

This means that telephones **do not transmit the F0 of our voices!**

The reason that we can discriminate the gender of the people we are talking to is because our brains can restore the fundamental from higher harmonics!

Let's use this new information to tackle our vowel perception problem



How about we divide by F3?

The **third formant** (F3) has some interesting properties:

1. It doesn't change from vowel to vowel (it is always the same)
2. It does change from person to person (gender, age, vocal tract length)
3. It doesn't change that much with the F0 of your voice
4. It is present when you whisper

This means you can use F3 as a unique speaker identifier, and gauge how much variation is due to the person, and how much is due to the vowels.

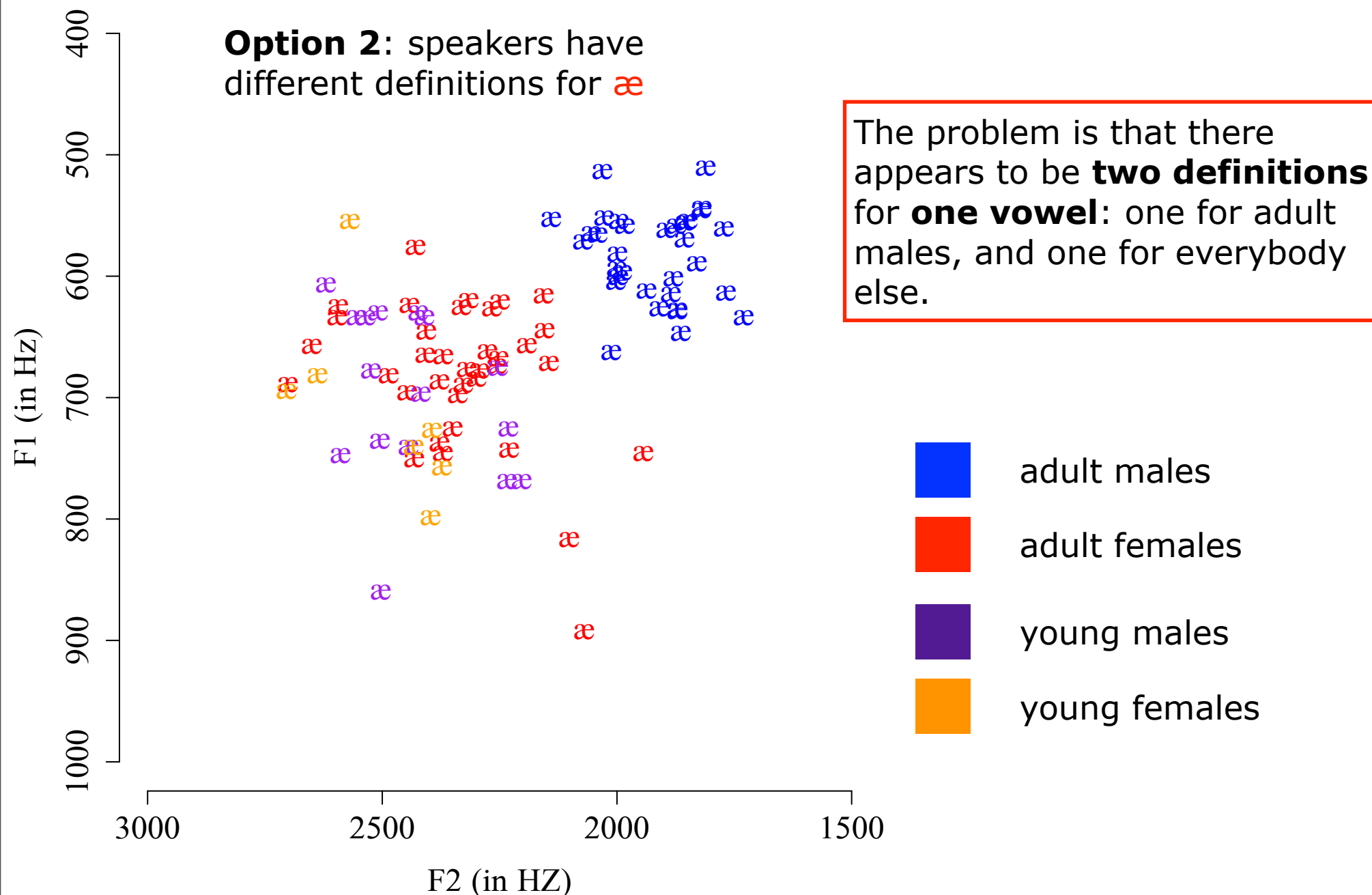
This led Monahan and Idsardi 2010 to propose **standardizing** vowels using F3. In mathematical terms, this simply means dividing F1 by F3 and dividing F2 by F3:

$F1/F3$

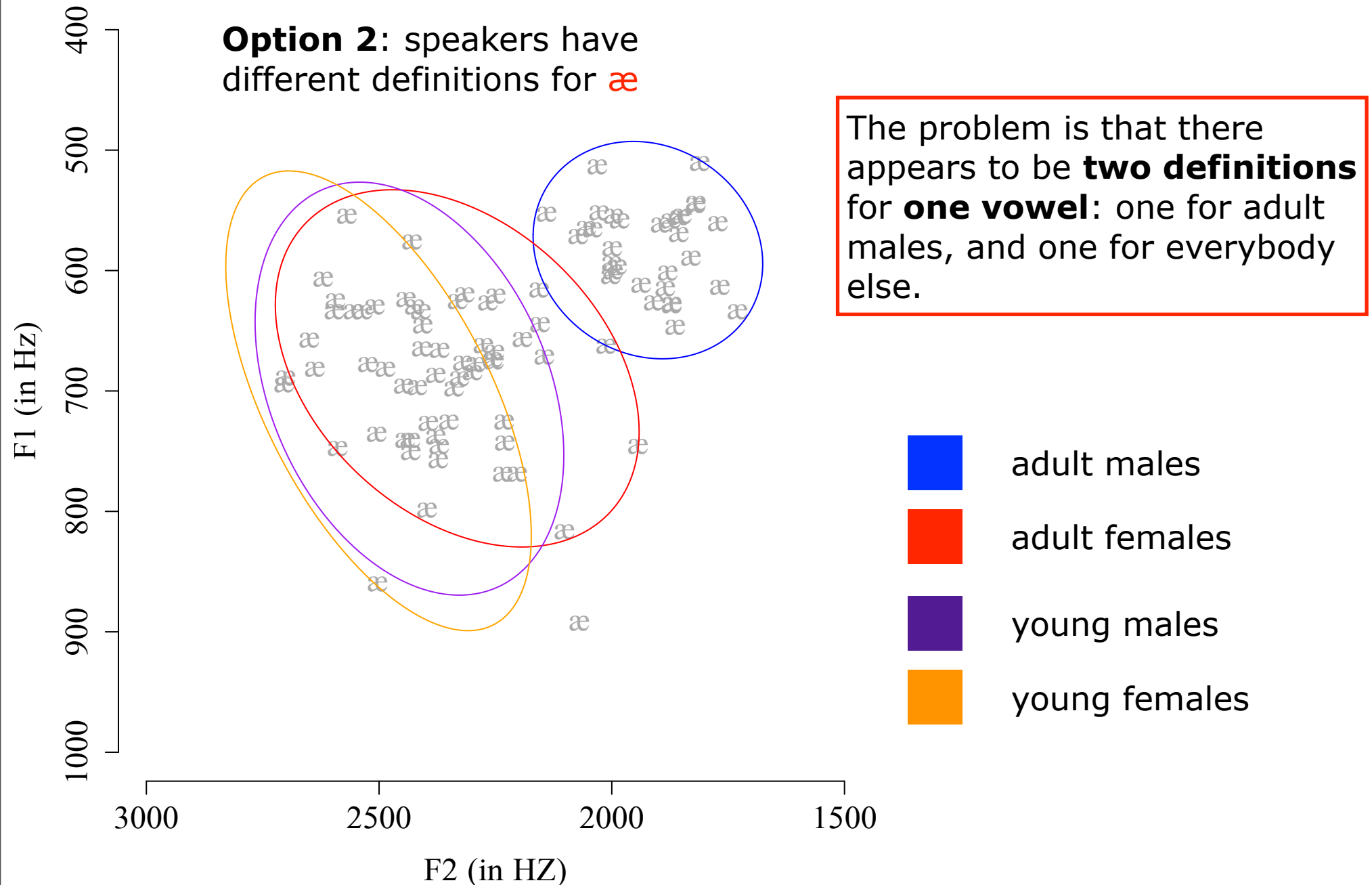
$F2/F3$

The idea is that this will eliminate a large portion of the difference in formants between people with different F0s, but without relying on F0 (which varies more than F3 when you speak)

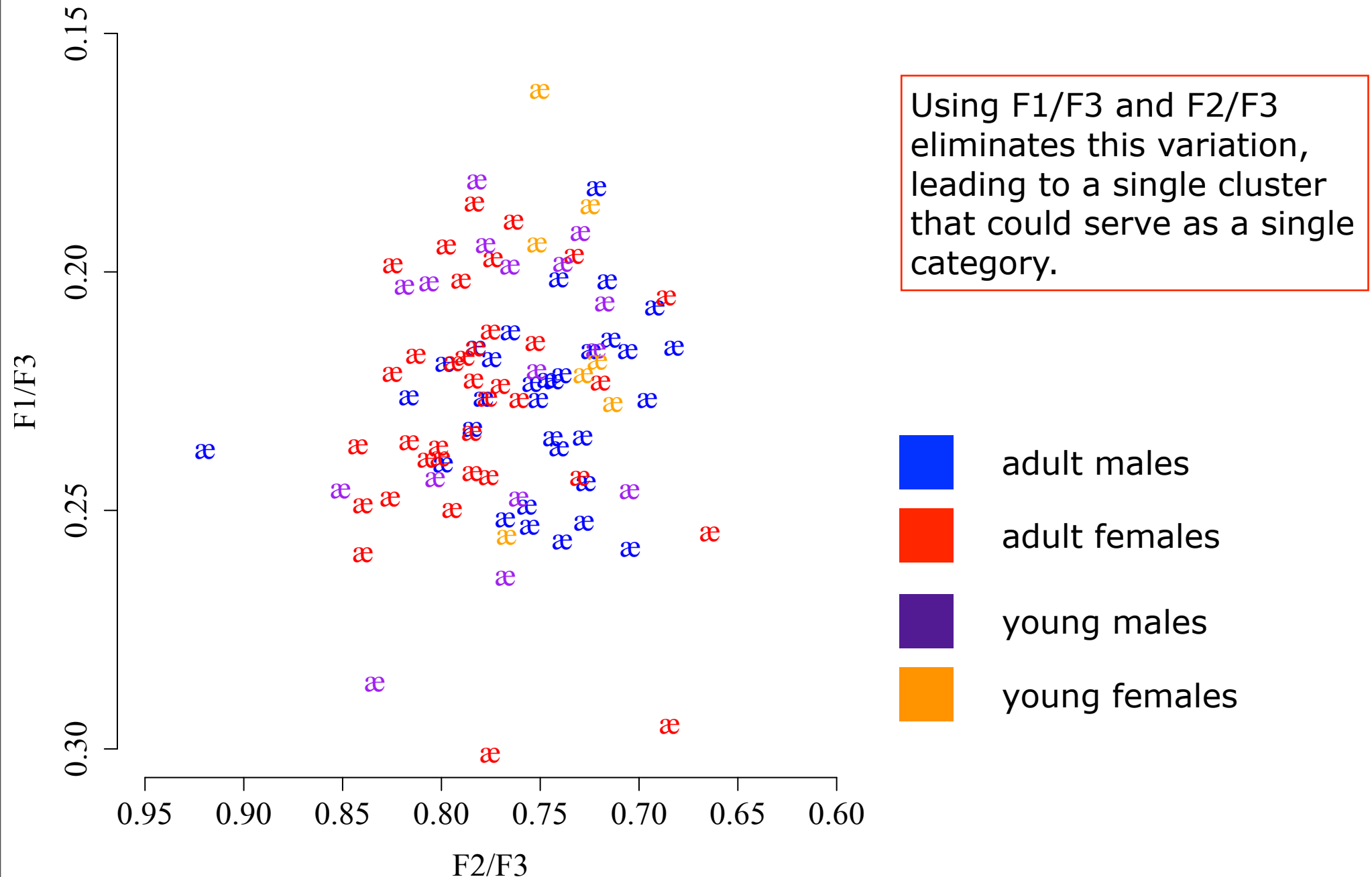
Here are the raw formant values for æ



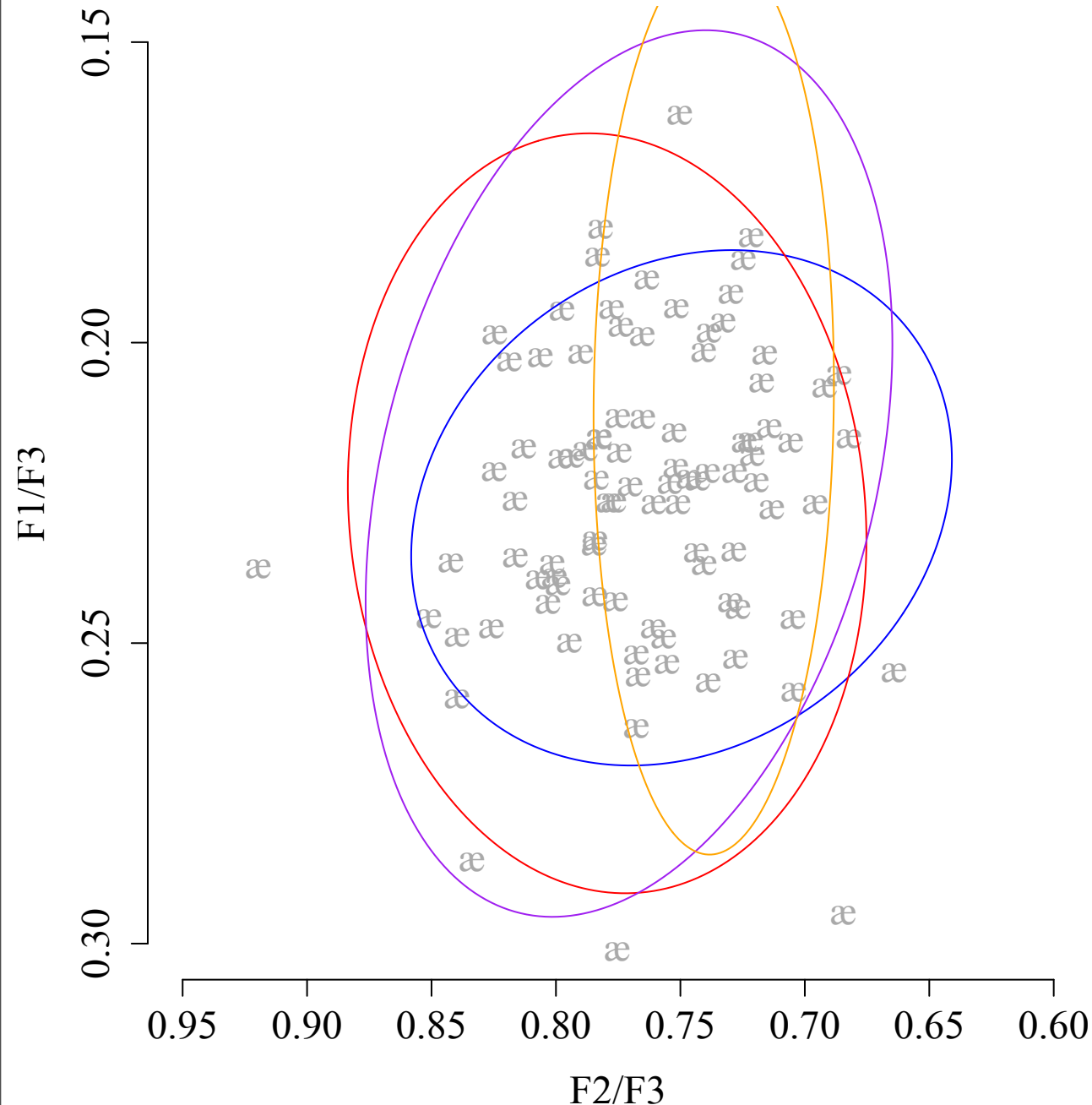
Here are the raw formant values for æ



Here are the F1/F3 and F2/F3 values for æ



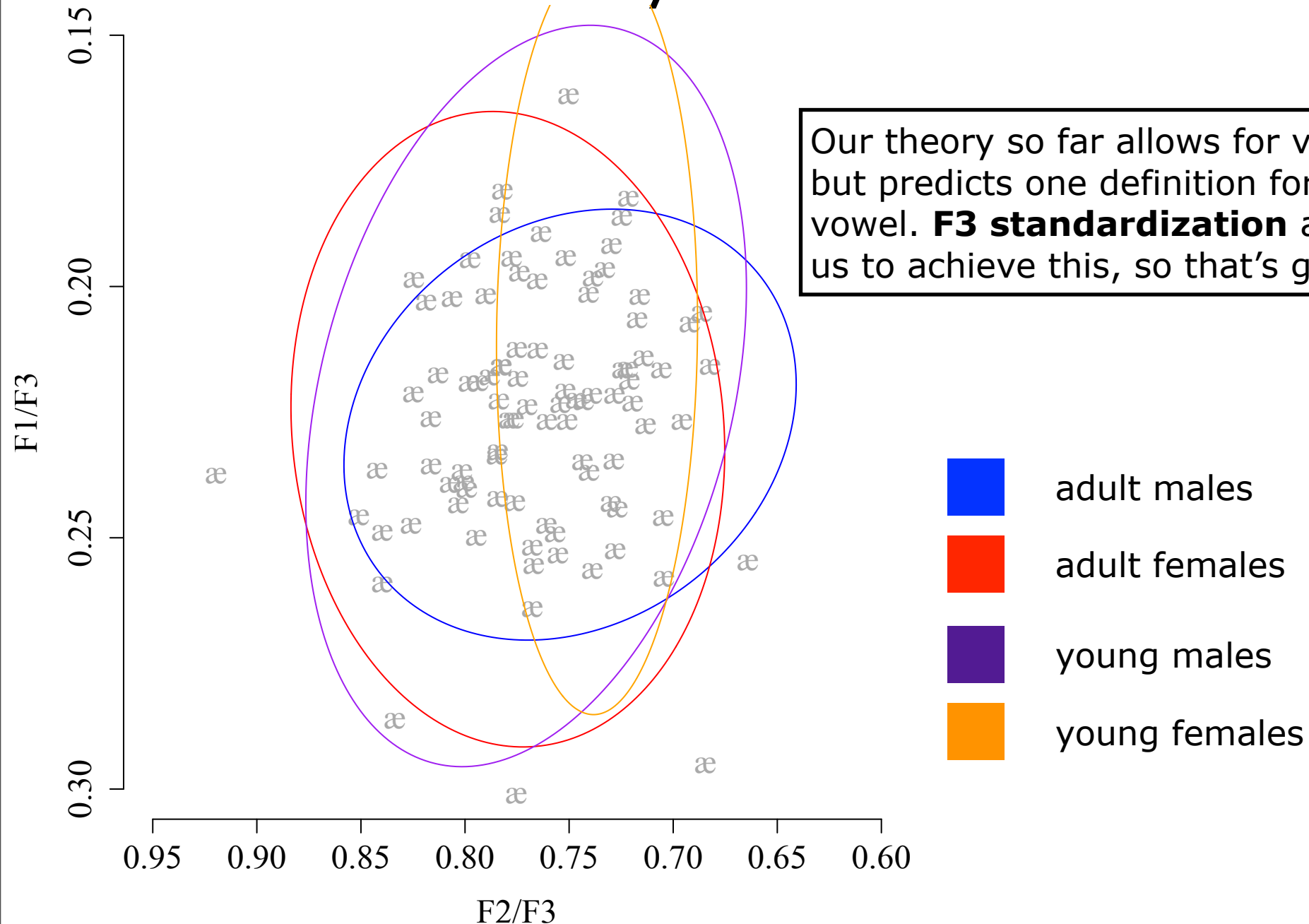
Here are the F1/F3 and F2/F3 values for æ



Using F1/F3 and F2/F3 eliminates this variation, leading to a single cluster that could serve as a single category.

- adult males
- adult females
- young males
- young females

There is still some variation, but only one cluster!



Recap for the day

starting
representation

**sensory
representation**



Every property of sound:
-amplitude
-frequency (F0, F1, F2)



IPA CHART
IPA SYMBOLS

VOWELS

ɪ:	ɪ	ʊ	u:
SLEEP	SIT	BOOK	FOOT
e	ə	ɜ:	ɔ:
BED	AFTER	BIRD	PORT
æ	ʌ	ɑ:	ɒ
MAN	CUP	CAR	POT

DIPHTHONGS

ɪə	eɪ	
EAR	FLAY	
ʊə	ɔɪ	əʊ
FEWER	BOY	NO
eə	aɪ	aʊ
HAR	BUY	COW

CONSONANTS

p	b	t	d	tʃ	dʒ	k	g
PEN	BIG	TEN	DOG	CHURCH	JUDGE	CAKE	GUEST
f	v	θ	ð	s	z	ʃ	ʒ
FAIR	VERY	THREE	THIS	BUS	BUZZ	CASH	LEISURE
m	n	ŋ	h	l	r	w	j
MORE	NOW	SONG	HAPPY	LIP	ROSE	WAT	YES

Big Question:

Can we find a set of acoustic properties that **uniquely define** each sound in our language?

Recap for the day

Big Question:

Can we find a set of acoustic properties that **uniquely define** each sound in our language?

Possible Answer:

Each speech sound can be **uniquely defined** by the first two formants (**F1** and **F2**).

Prediction 1:

Only F1 and F2 should be necessary for perception.

In fact, F3 is also necessary.

Prediction 2:

We should be able to observe a single definition for each vowel despite variation in the F0 of speakers (male/female, adult/child).

We achieved this using F3 normalization (dividing F1 and F2 by F3), based on insights into auditory processing provided by the restoration of missing fundamentals!