Inaudible Voice Commands: The Long-Range Attack and Defense


Nirupam Roy


Sheng Shen

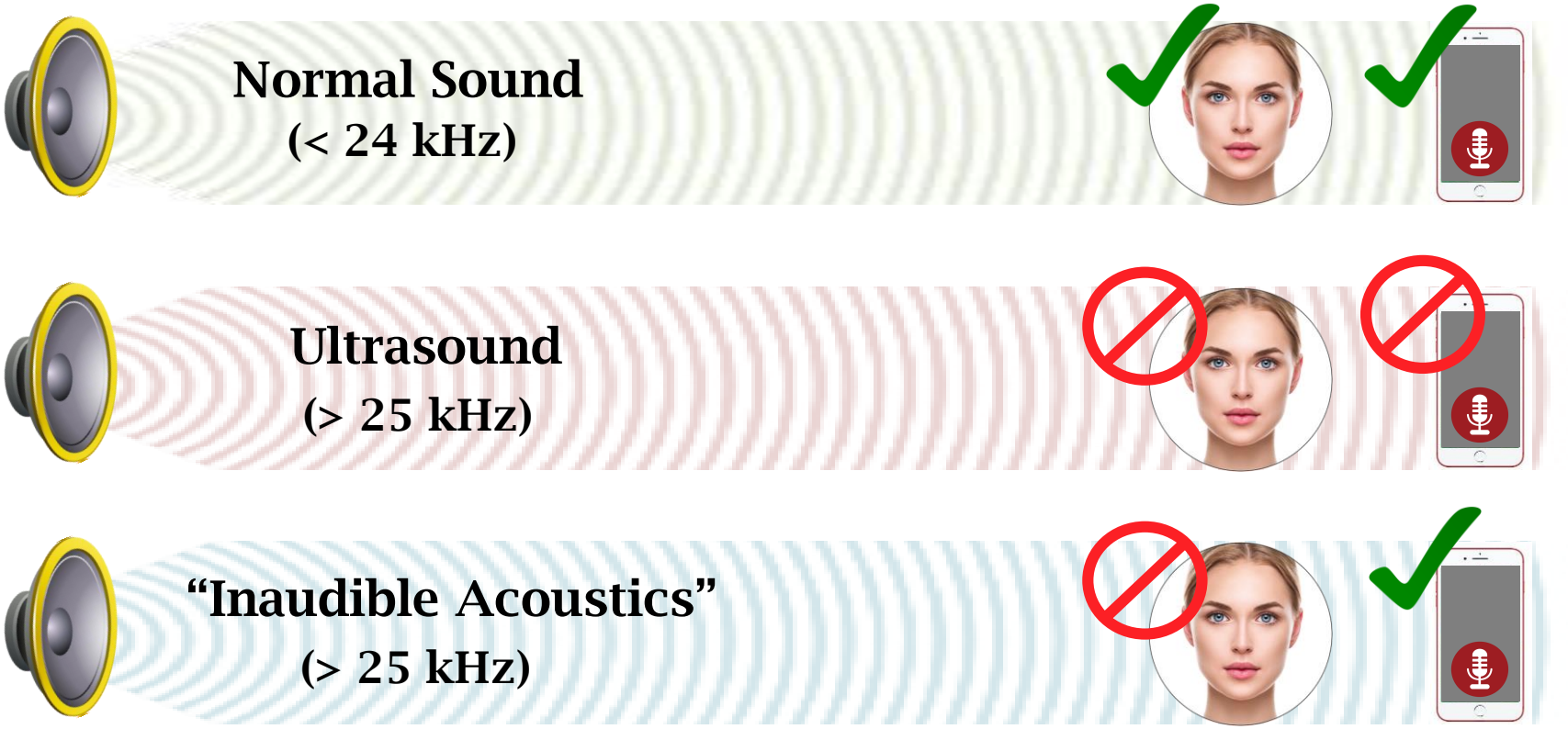

Haitham Hassanieh


Romit Roy Choudhury

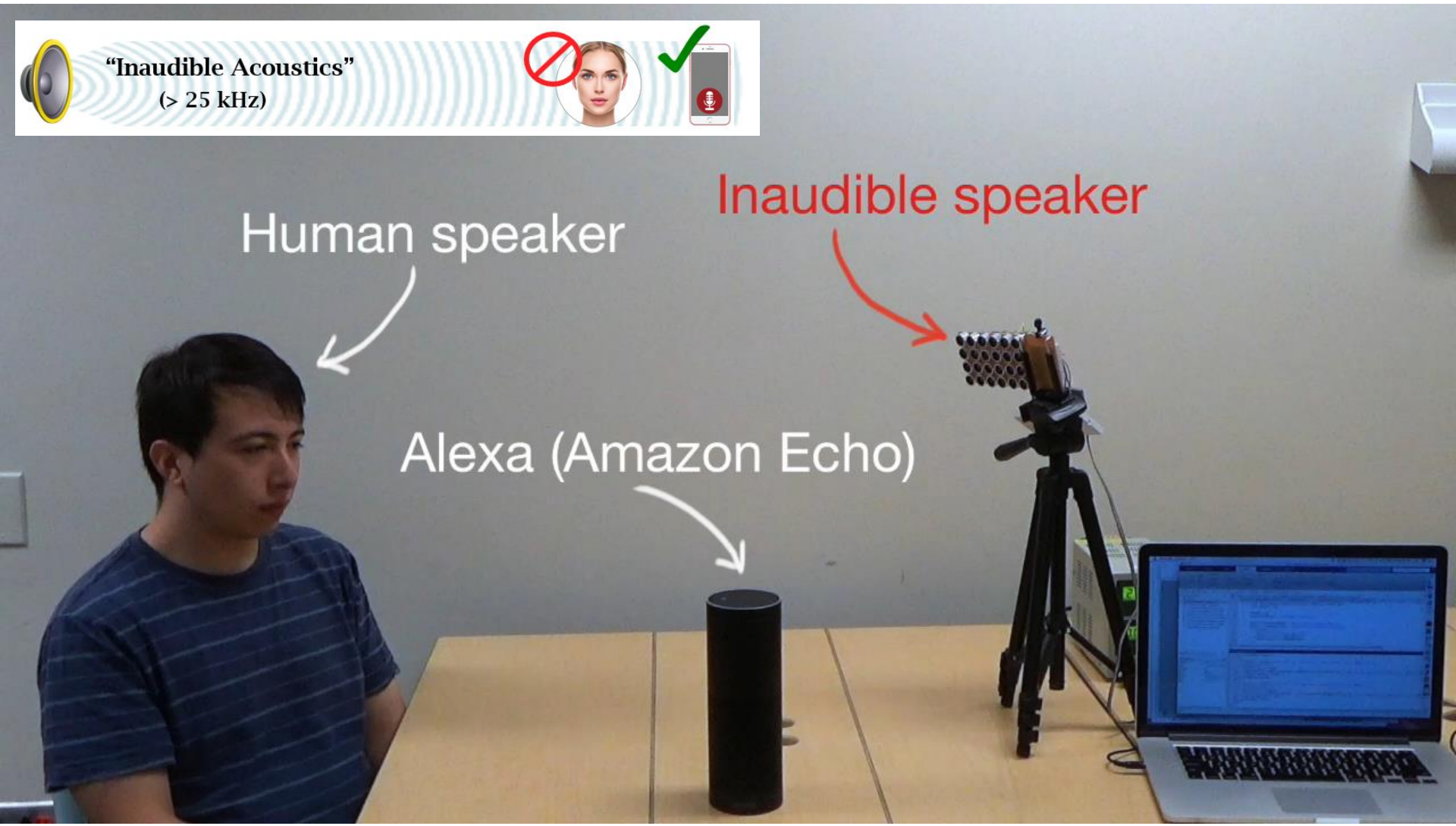
University of Illinois at Urbana-Champaign

50 million voice assistants are sold in US

Inaudible Acoustics



Inaudible Acoustics



Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

MobiSys'17
(Best Paper)

CCS'17

arXiv

Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

Today's Talk:

1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?

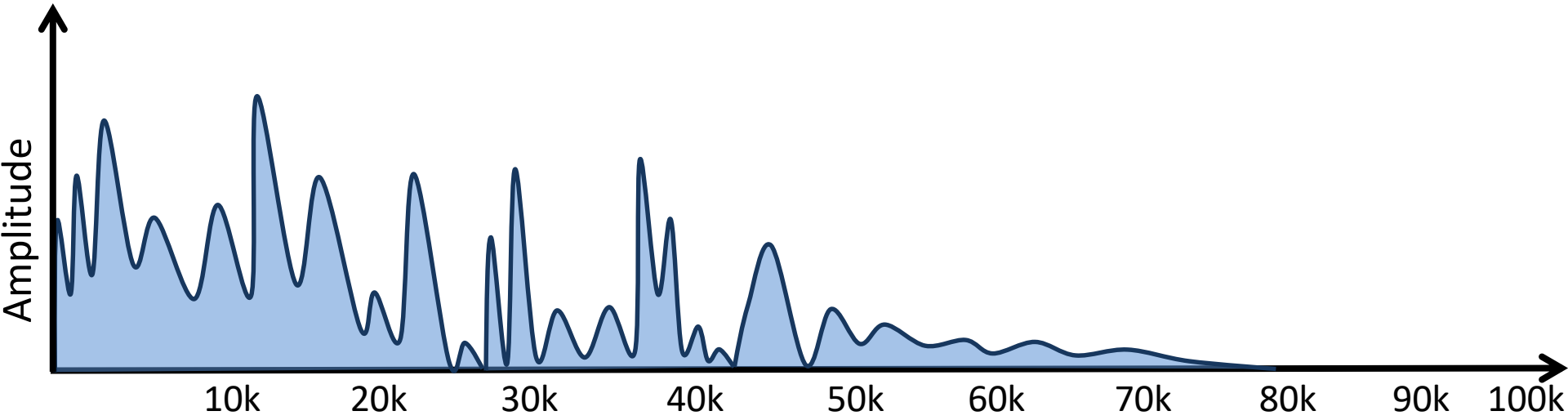
Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

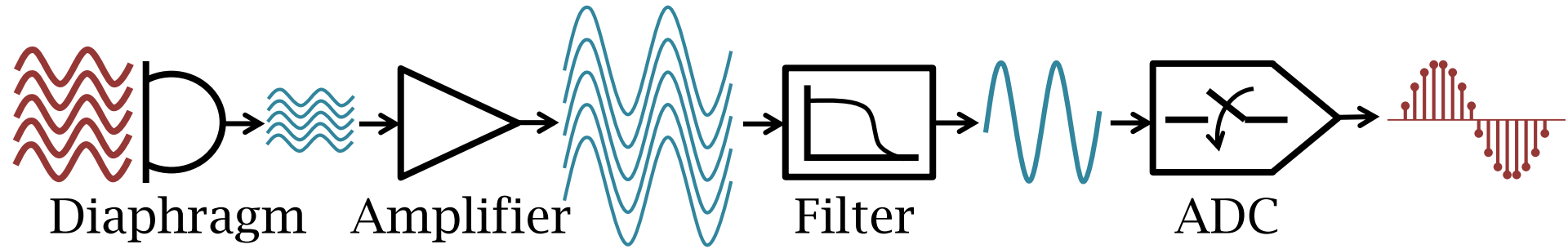
Today's Talk:

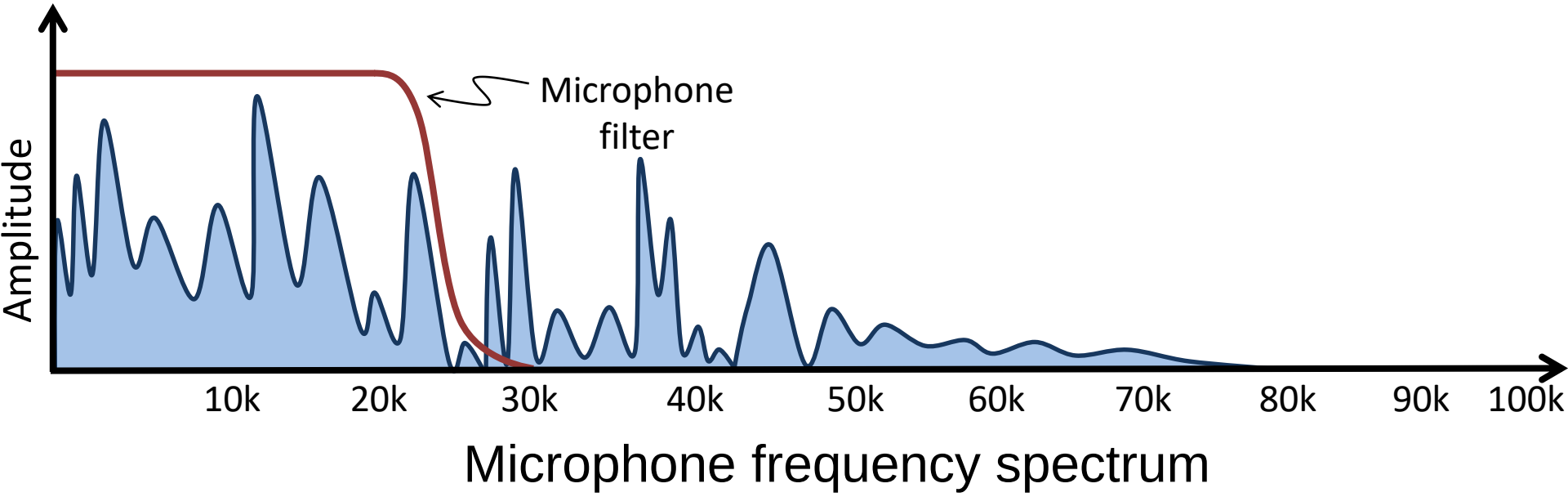
1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?

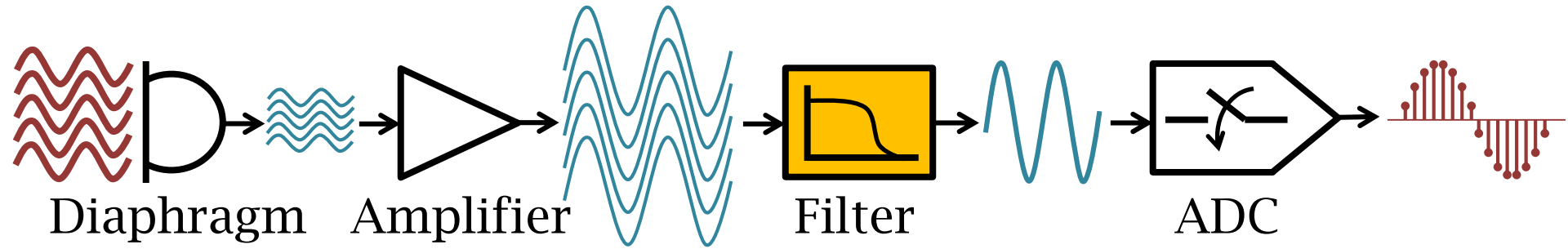


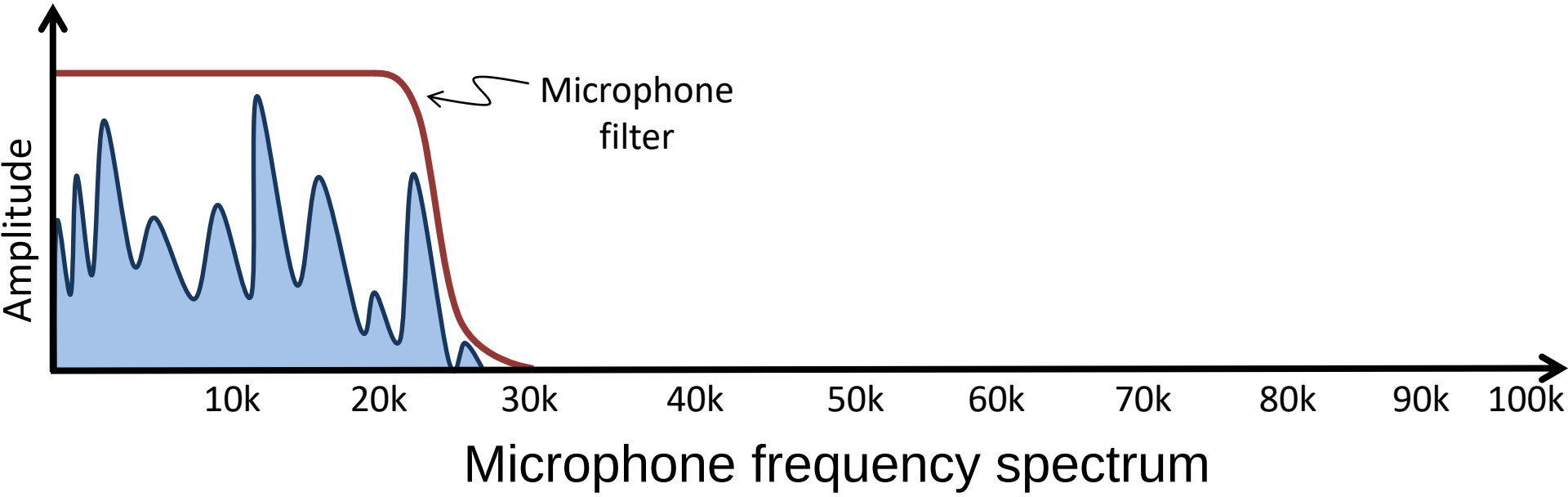
Microphone frequency spectrum

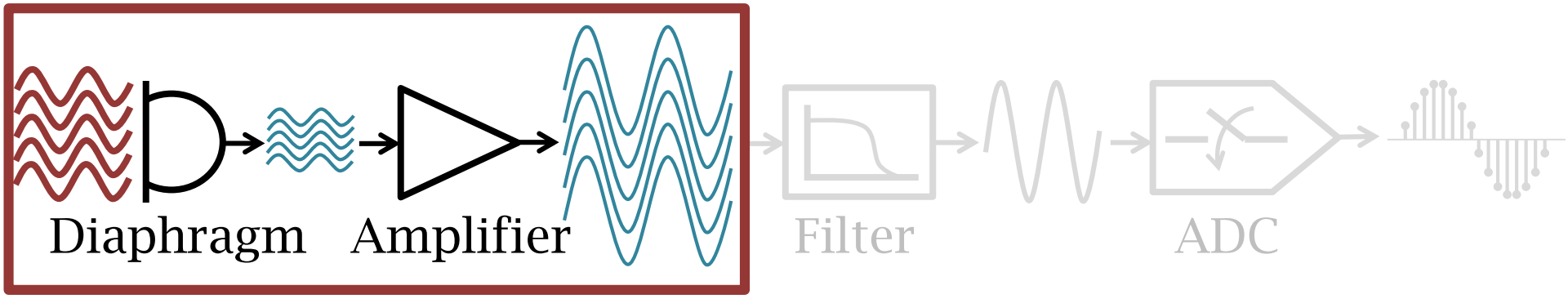




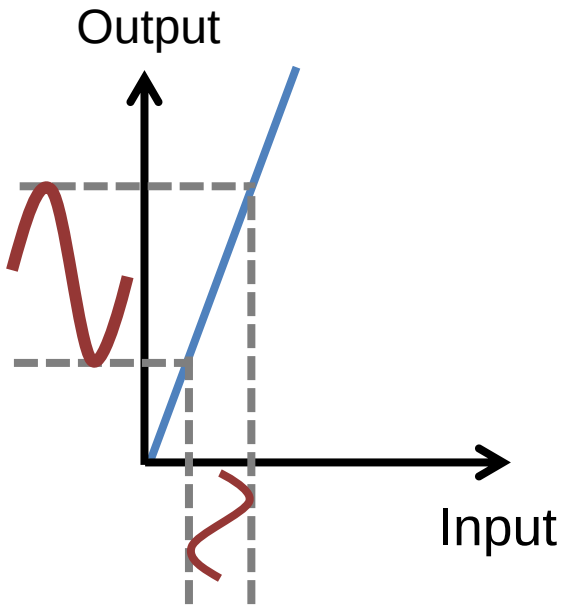
Microphone frequency spectrum



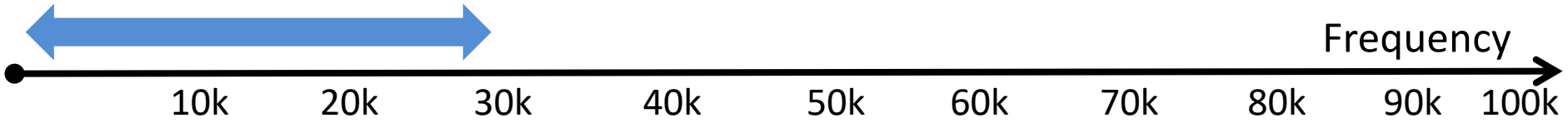


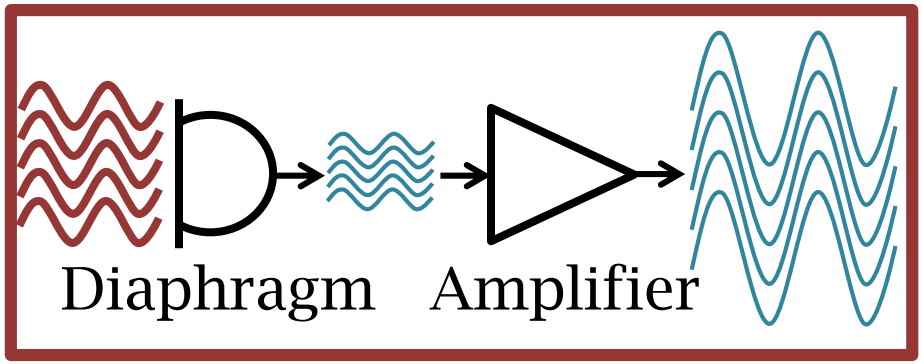


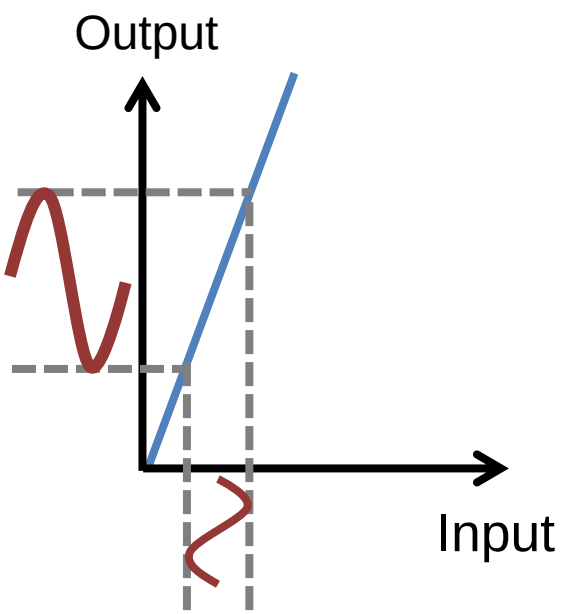
Air Vibration → Electric Voltage

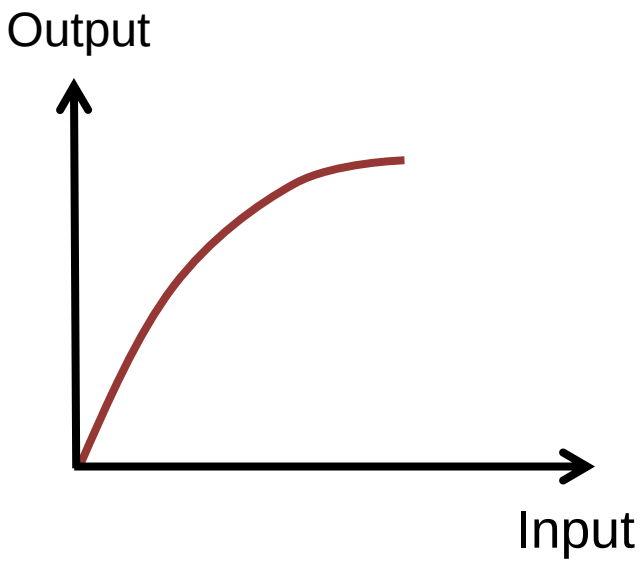

Output
Input

$V_{out} = a_1 V_{in}$


Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k


Diaphragm
Amplifier

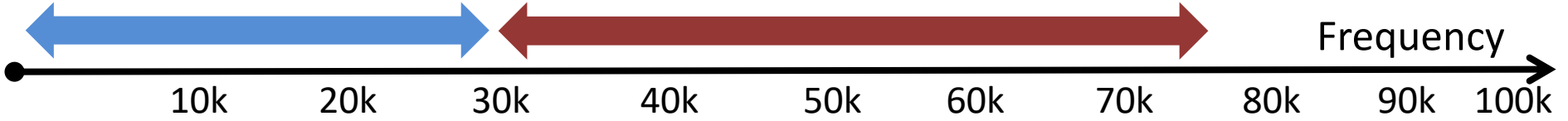


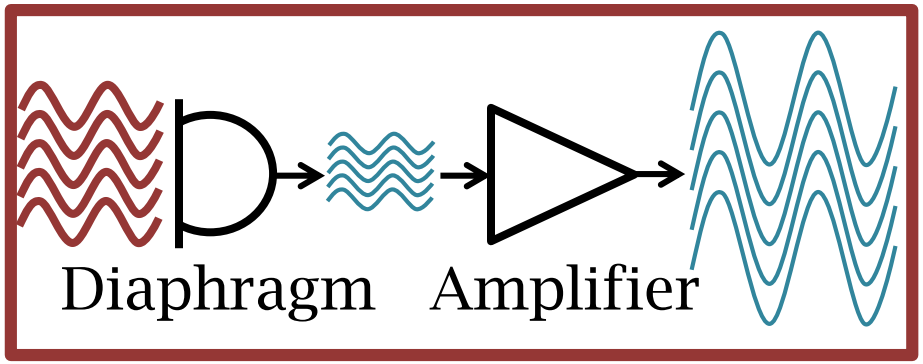


Nonlinear

$V_{out} = a_1 V_{in}$

$V_{out} = a_1 V_{in} + a_2 V_{in}^2 + \ldots$





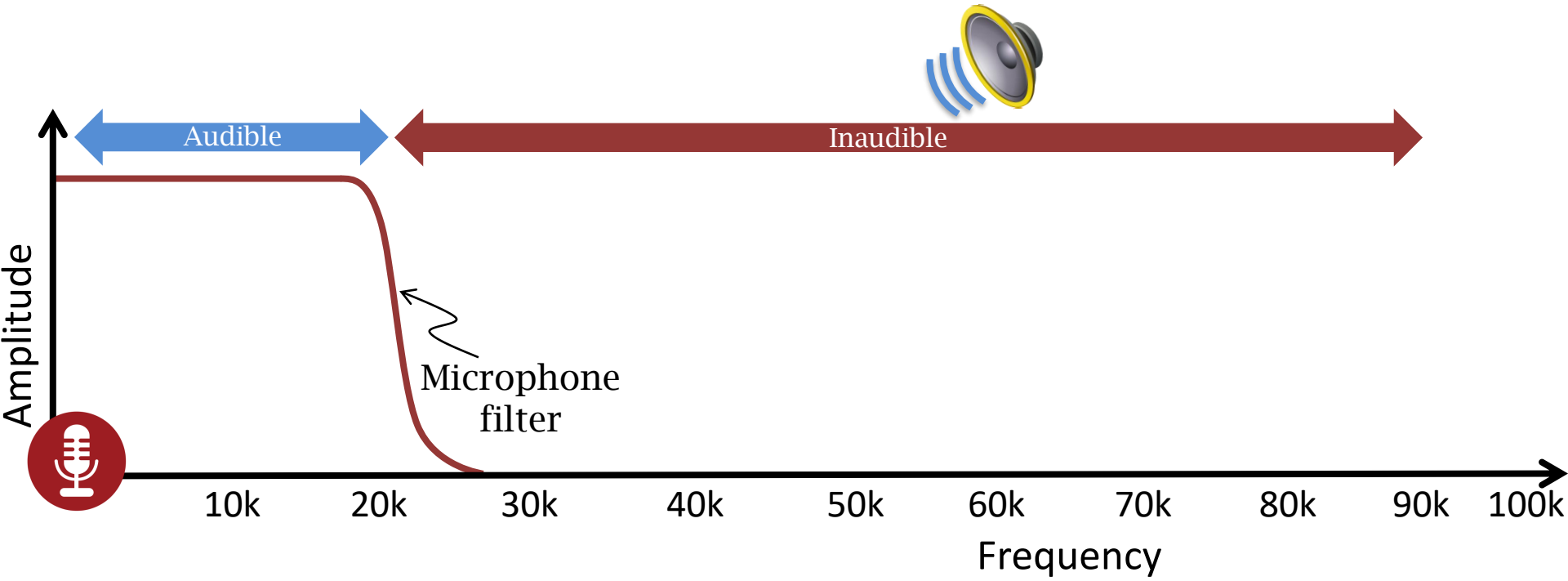
Nonlinear

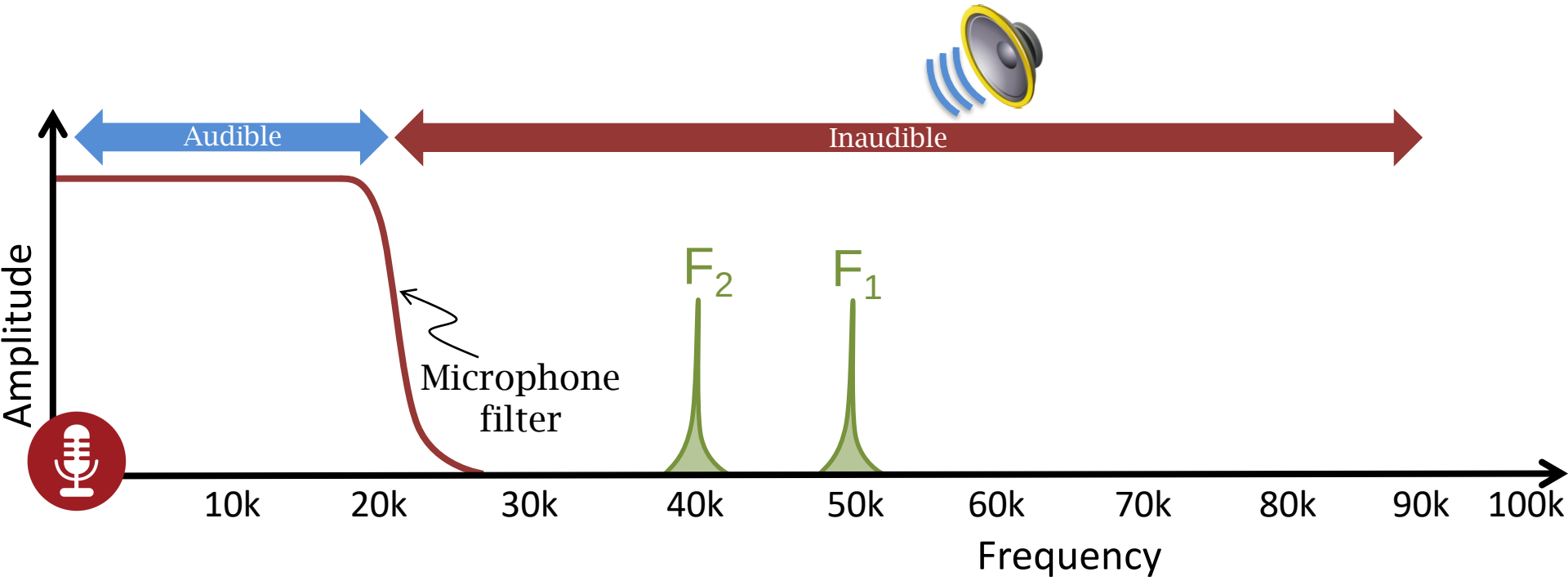
$V_{out} = a_1 V_{in}$

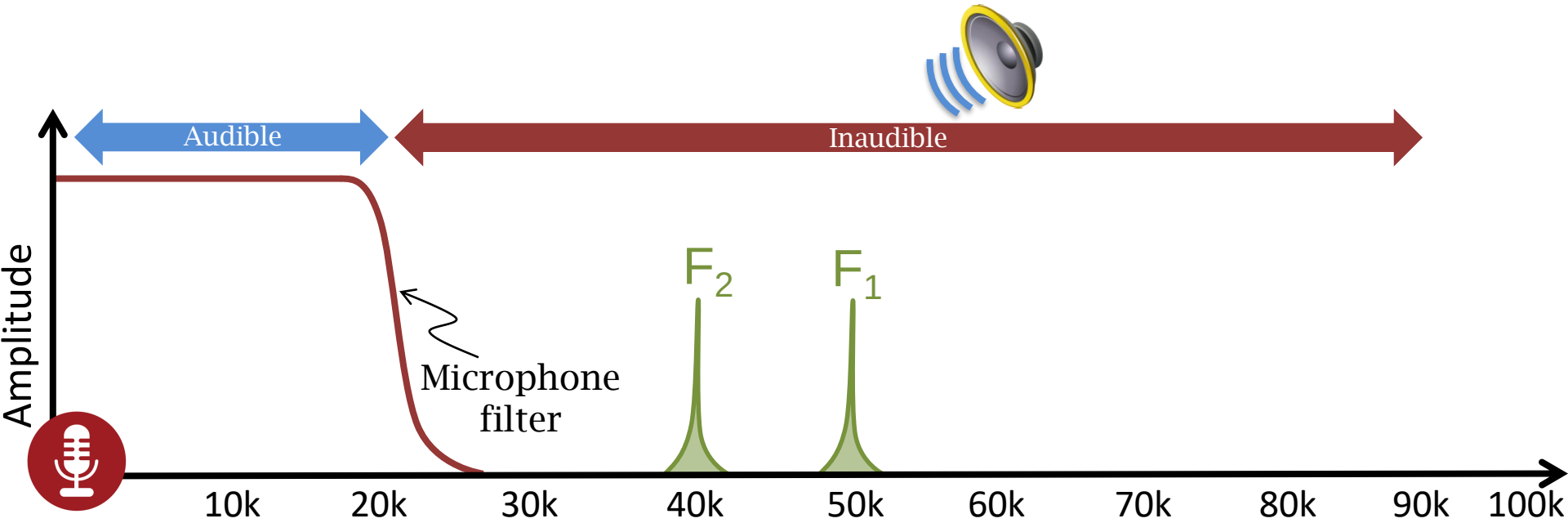
$V_{out} = a_1 V_{in} + a_2 V_{in}^2 + \ldots$

Frequency

10k 20k 30k 40k 50k 60k 70k 80k 90k 100k

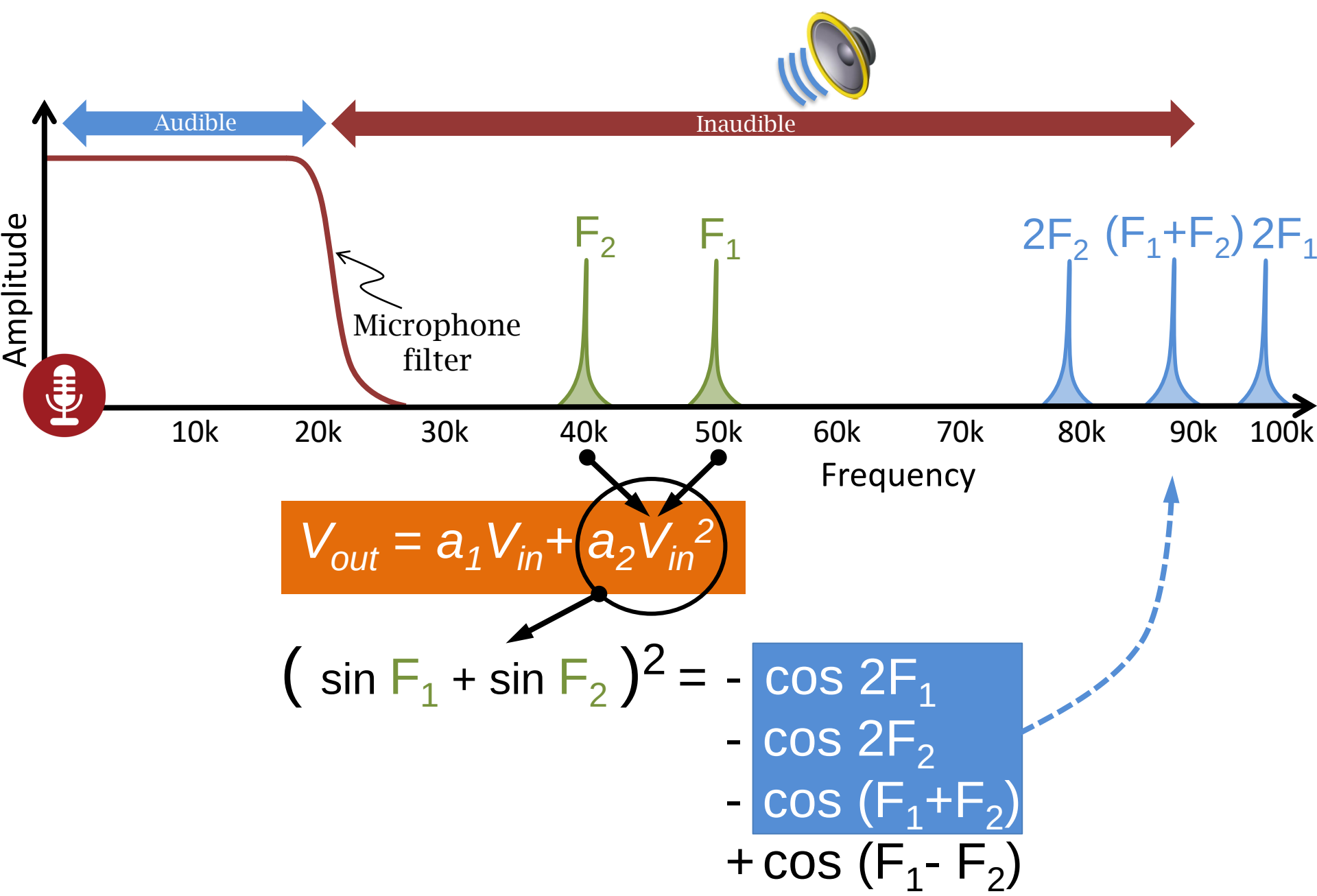

Audible
Inaudible
Amplitude
Microphone
filter
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Frequency


Audible
Inaudible
Amplitude
Microphone
filter
F2
F1
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Frequency



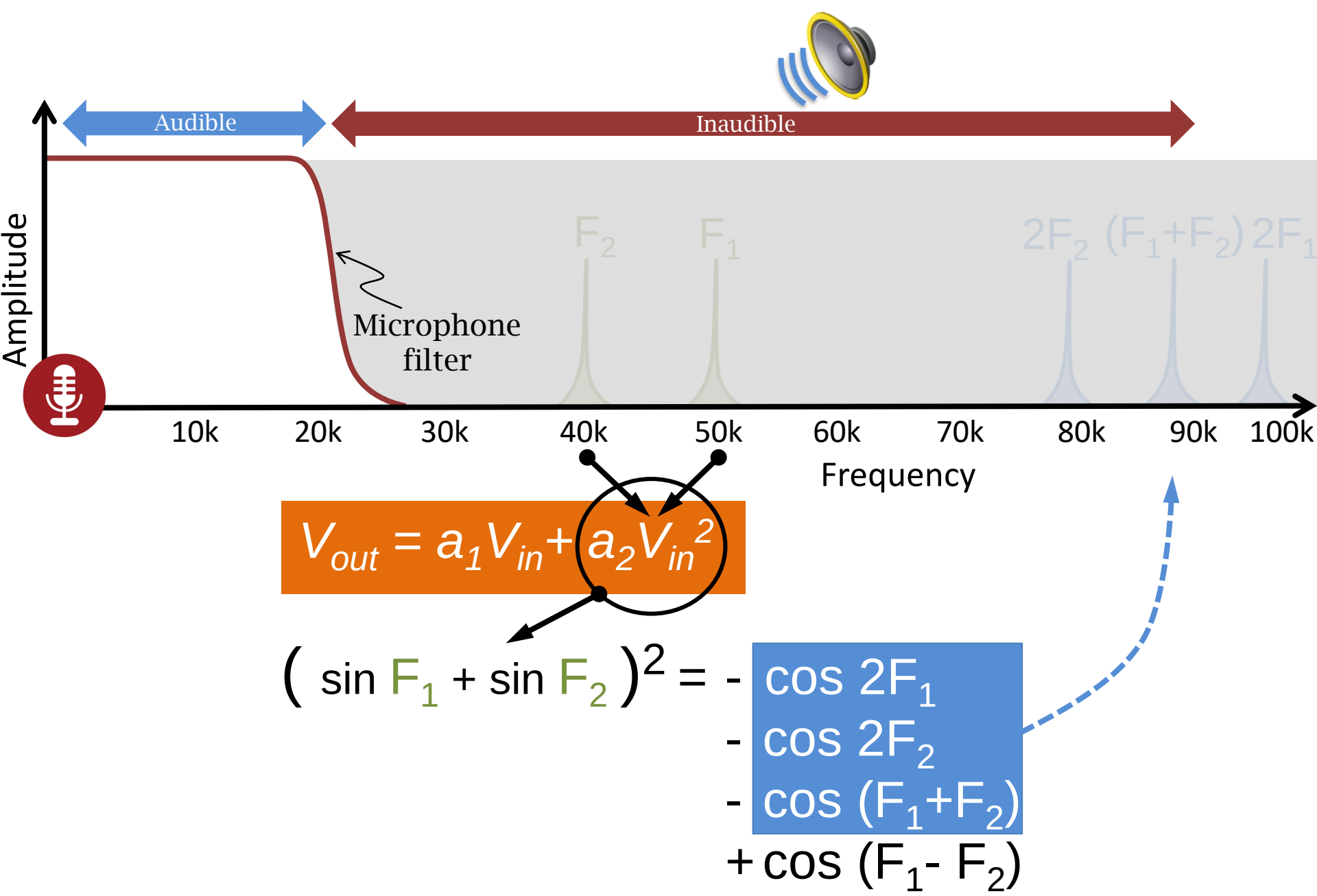
$$V_{out} = a_1 V_{in} + a_2 V_{in}^2$$

$$(\sin F_1 + \sin F_2)^2 = -\cos 2F_1 - \cos 2F_2 - \cos (F_1+F_2) + \cos (F_1 - F_2)$$



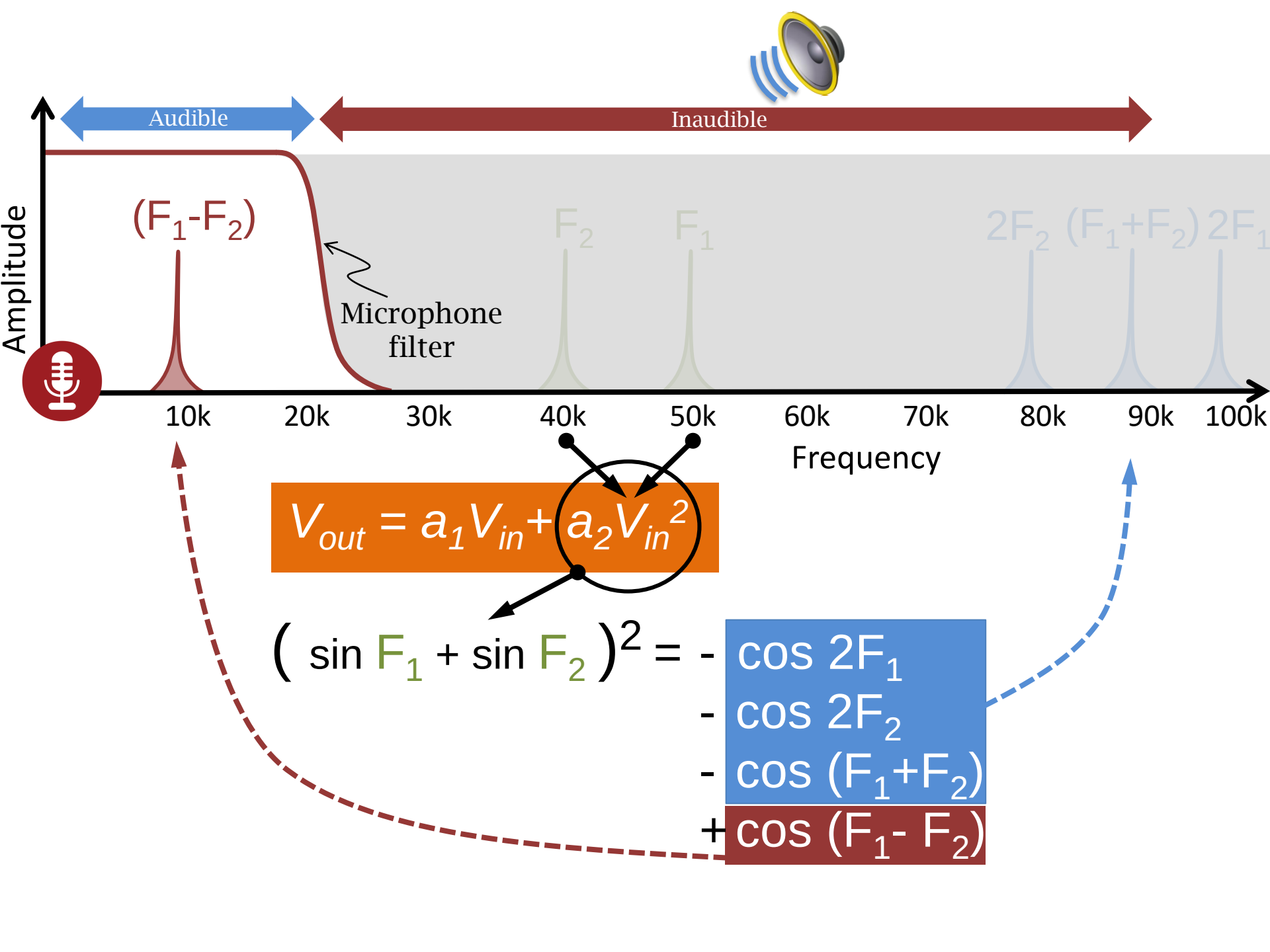
$$V_{out} = a_1 V_{in} + a_2 V_{in}^2$$

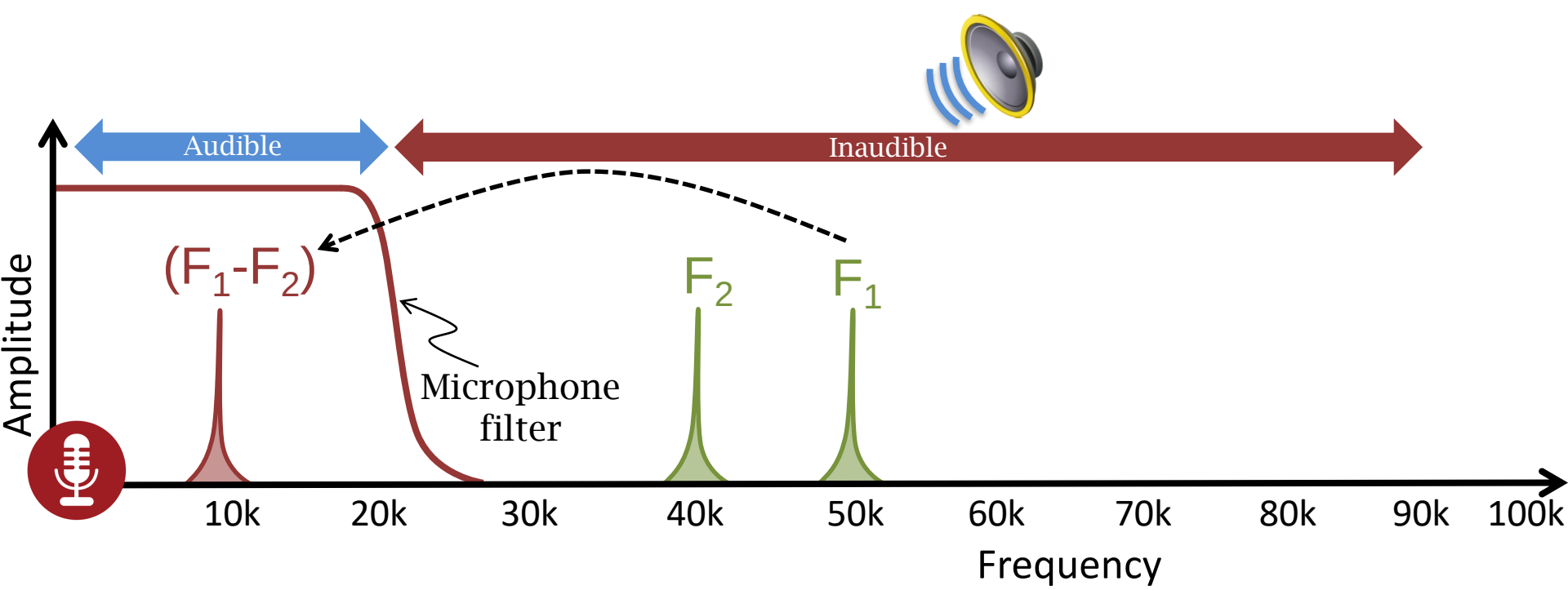
$$(\sin F_1 + \sin F_2)^2 = -\cos 2F_1 - \cos 2F_2 - \cos (F_1+F_2) + \cos (F_1 - F_2)$$

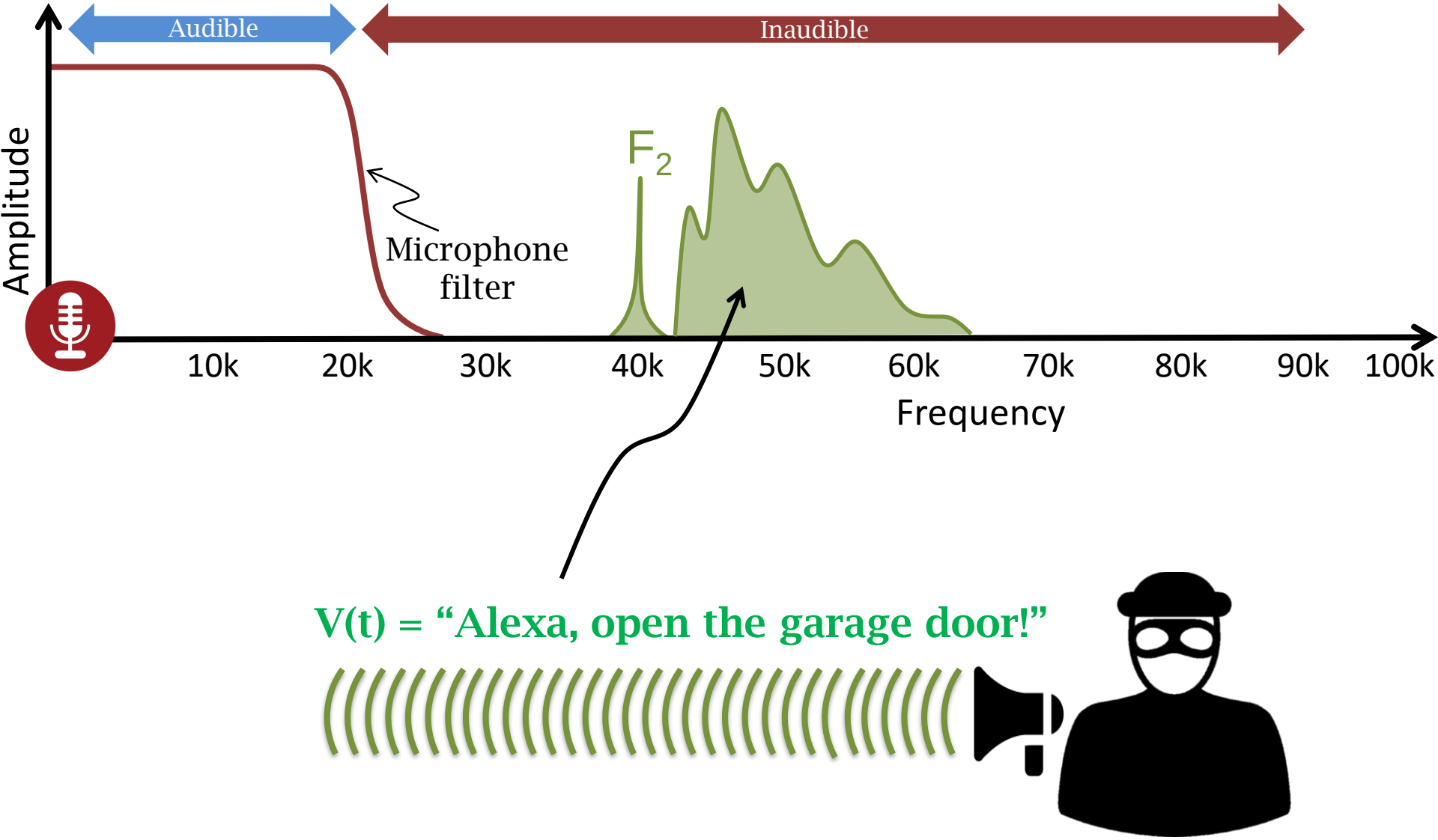


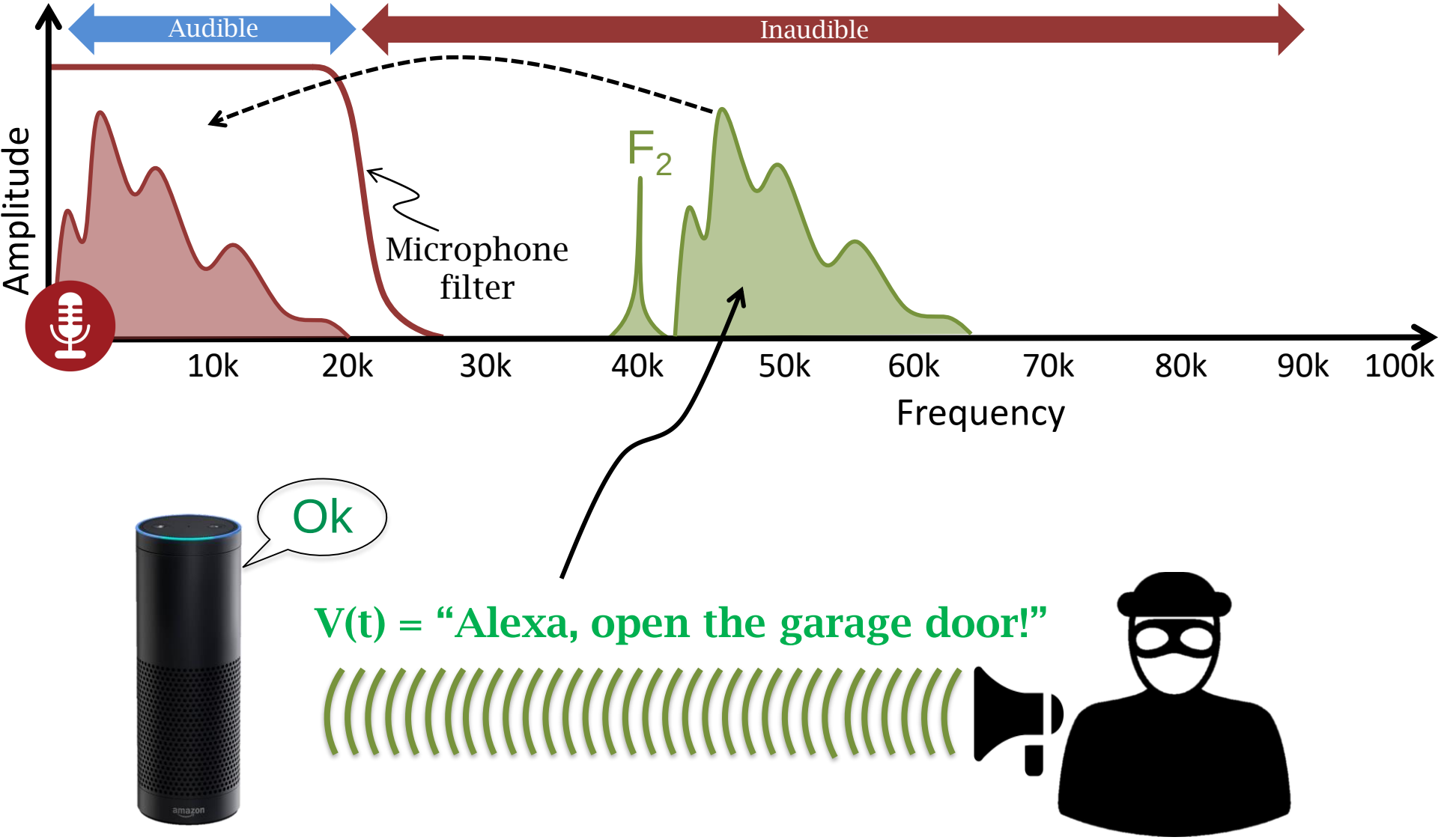
$$V_{out} = a_1 V_{in} + a_2 V_{in}^2$$

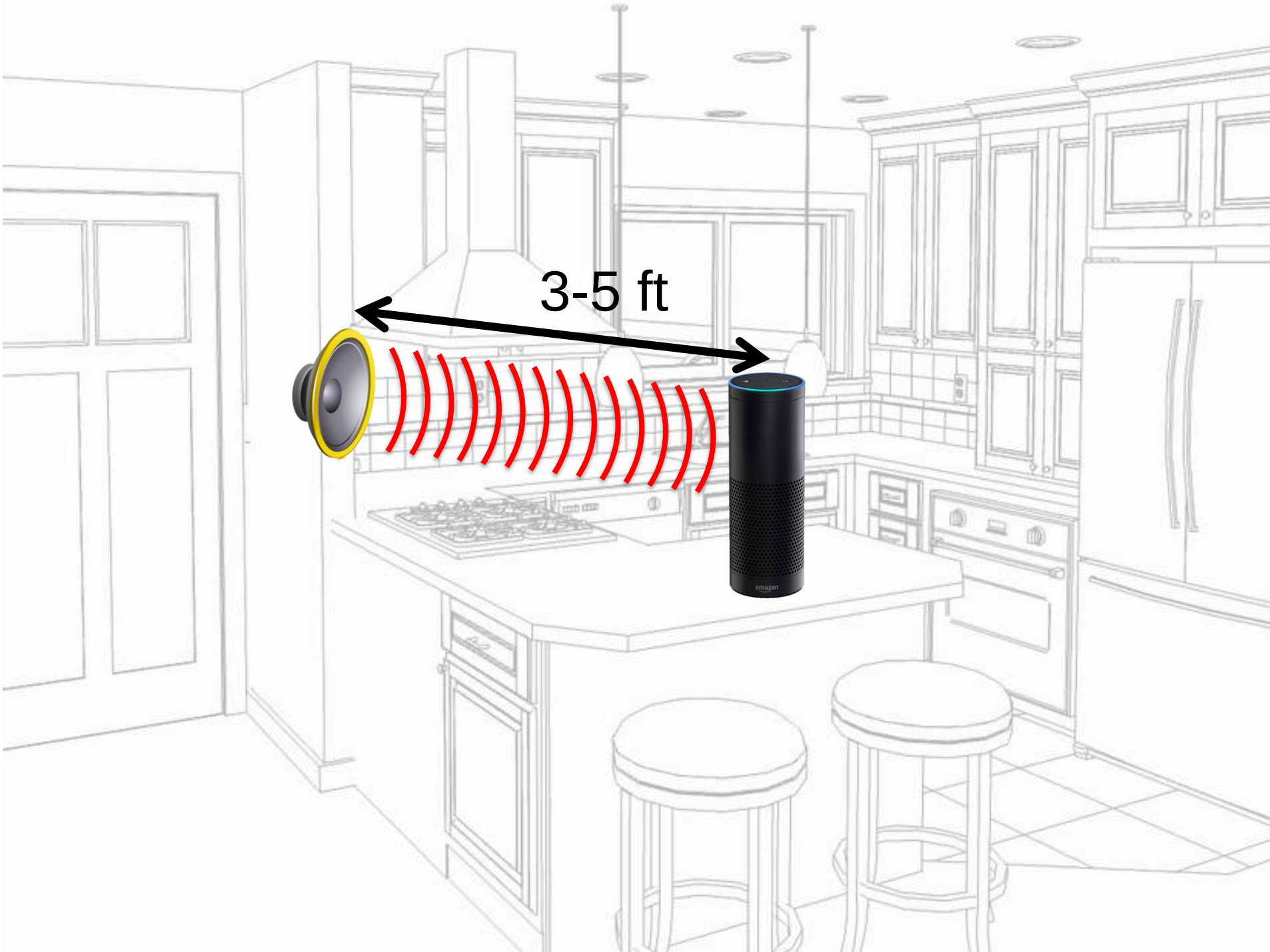
$$(\sin F_1 + \sin F_2)^2 = -\cos 2F_1 - \cos 2F_2 - \cos (F_1+F_2) + \cos (F_1 - F_2)$$


Audible
Inaudible
Amplitude
(F1-F2)
F2
F1
2F2
(F1+F2)
2F1
Microphone filter
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Frequency
Vout = a1Vin + a2Vin2
(sin F1 + sin F2)2 = - cos 2F1
- cos 2F2
- cos (F1+F2)
+ cos (F1- F2)

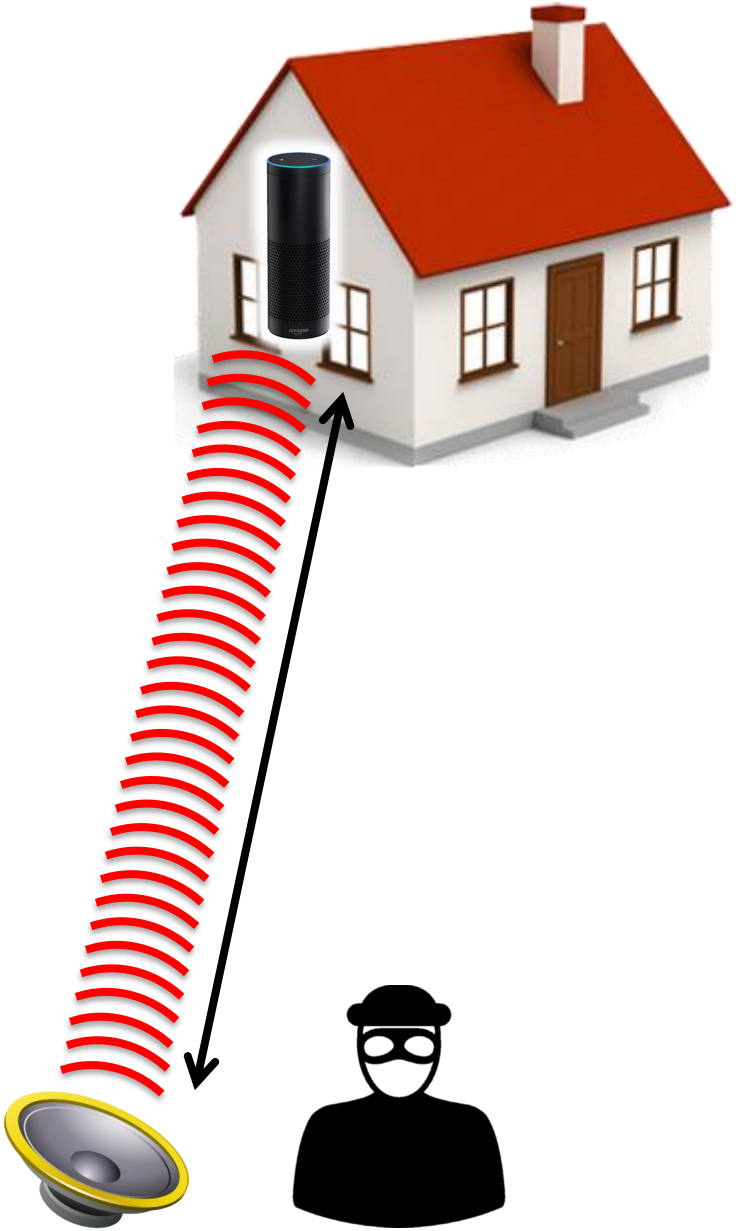

Audible
Inaudible
Amplitude
(F_1-F_2)
F_2
F_1
Microphone filter
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Frequency


Audible
Inaudible
Amplitude
Microphone
filter
F_2
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Frequency
V(t) = “Alexa, open the garage door!”


Audible
Inaudible
Amplitude
F_2
Microphone filter
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Frequency
Ok
V(t) = "Alexa, open the garage door!"


3-5 ft

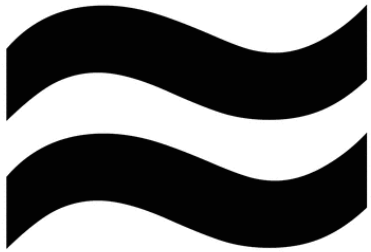
Can someone attack from a longer range?

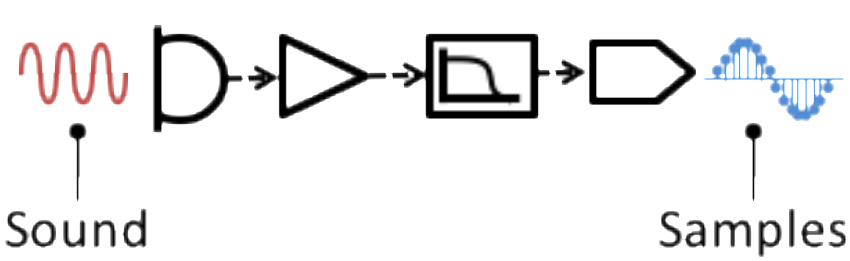


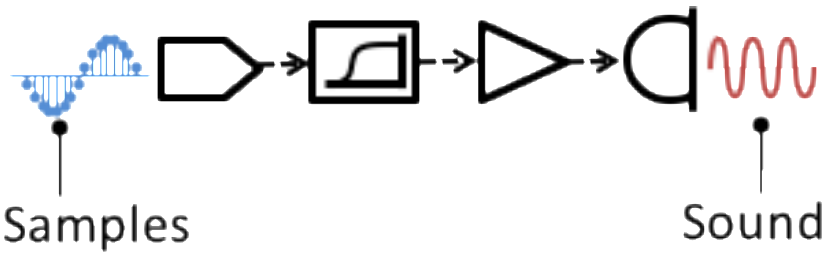
Can someone attack from a longer range?

Can someone attack from a longer range?

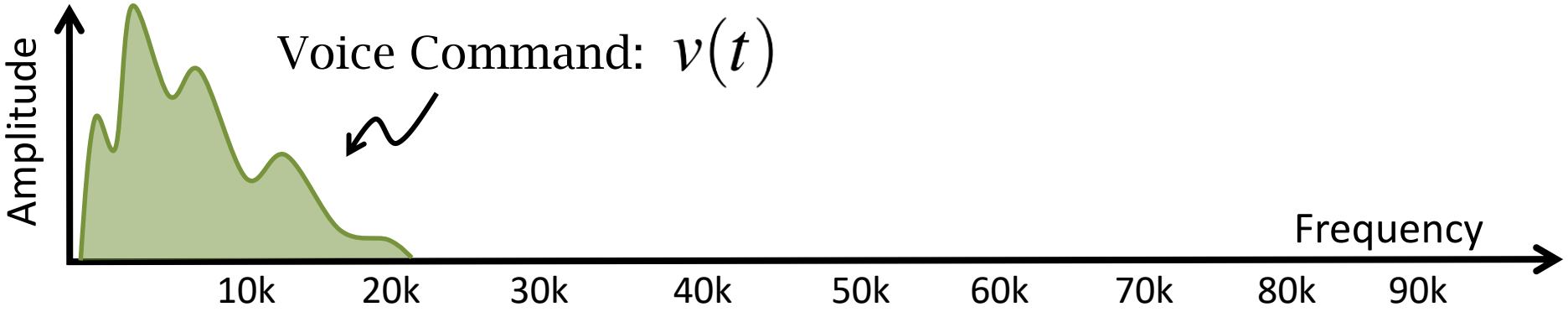
High power makes ultrasonic speakers audible

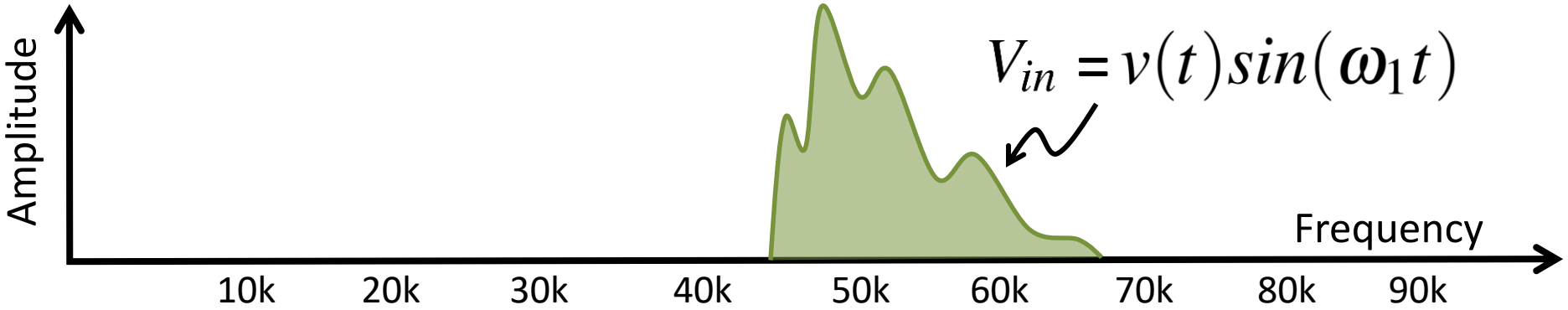


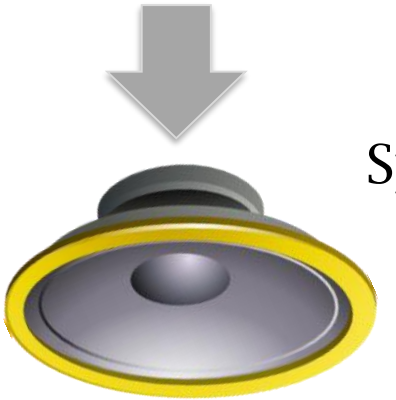




Speakers have nonlinearity too!

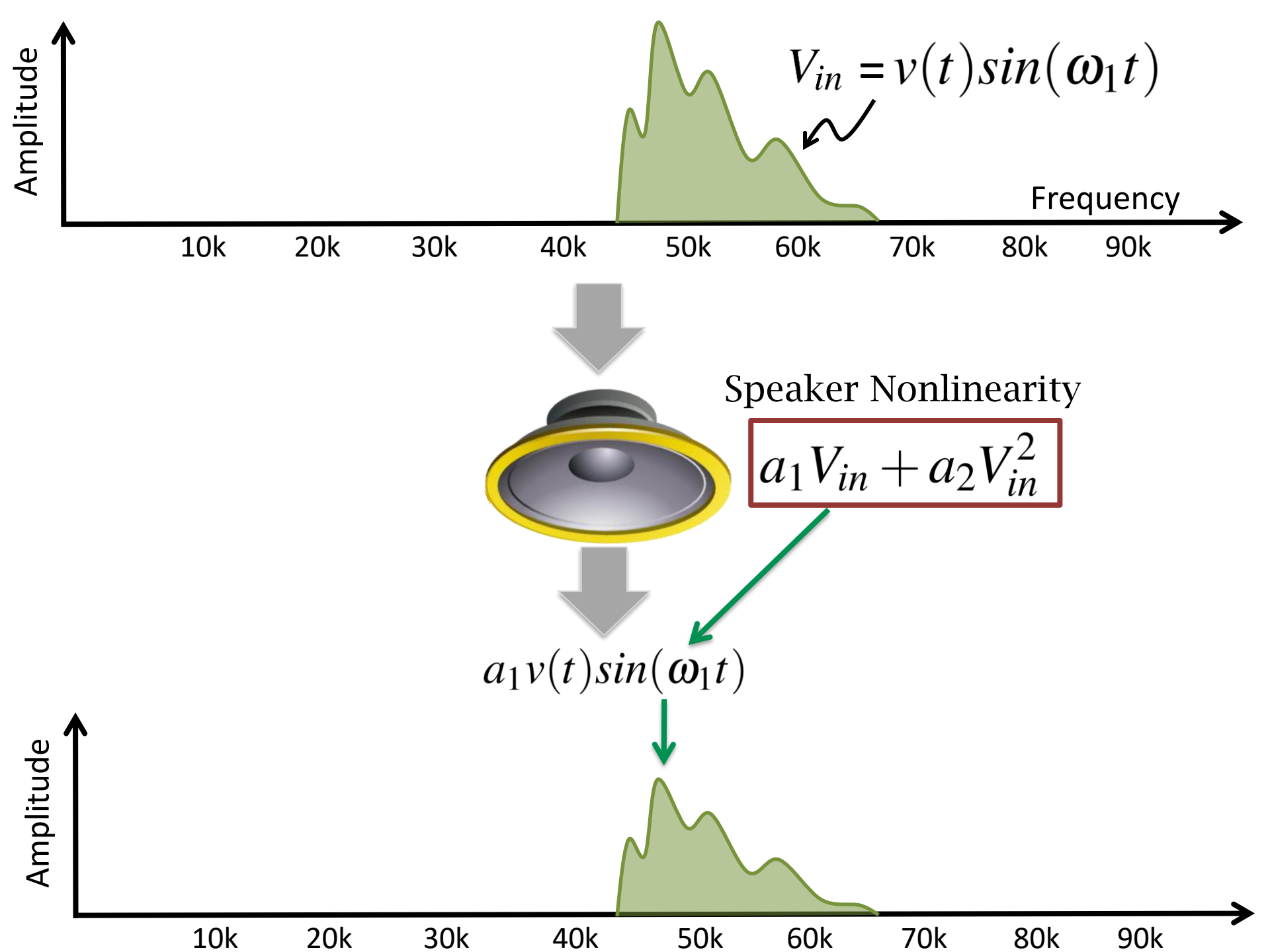

Amplitude
Voice Command: $v(t)$
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k

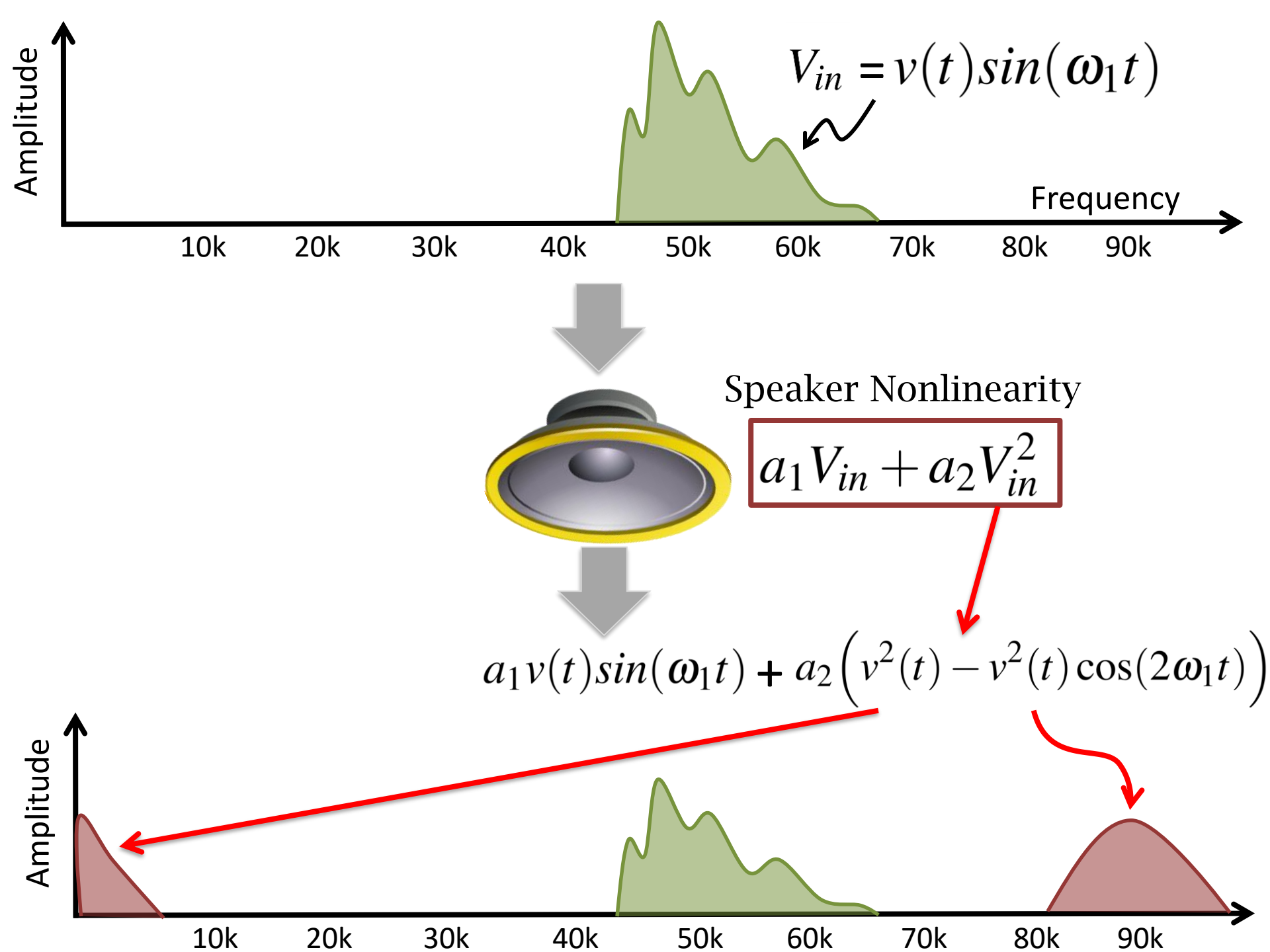


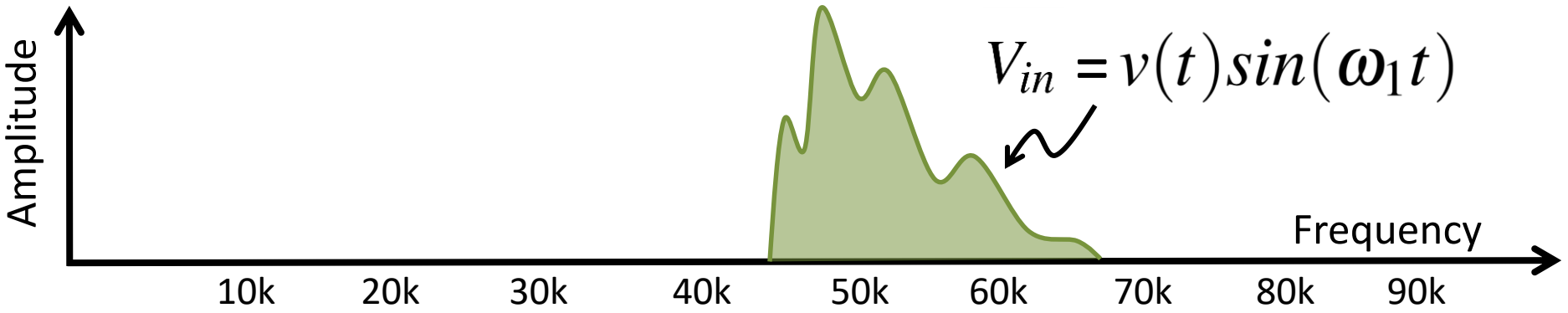


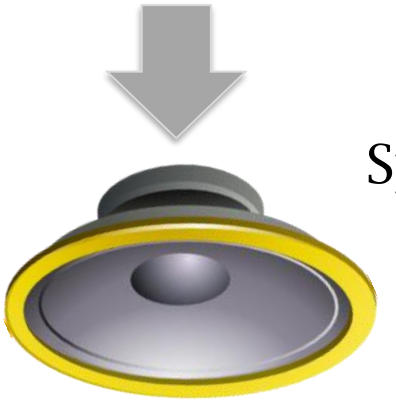
Speaker Nonlinearity

$$a_1 V_{in} + a_2 V_{in}^2$$


Amplitude
$V_{in} = v(t)sin(\omega_1 t)$
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
Speaker Nonlinearity
$a_1 V_{in} + a_2 V_{in}^2$
$a_1 v(t)sin(\omega_1 t)$
Amplitude
10k
20k
30k
40k
50k
60k
70k
80k
90k


Amplitude
$V_{in} = v(t)sin(\omega_1 t)$
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
Speaker Nonlinearity
$a_1 V_{in} + a_2 V_{in}^2$
$a_1 v(t)sin(\omega_1 t) + a_2\left(v^2(t) - v^2(t)\cos(2\omega_1 t)\right)$
Amplitude
10k
20k
30k
40k
50k
60k
70k
80k
90k

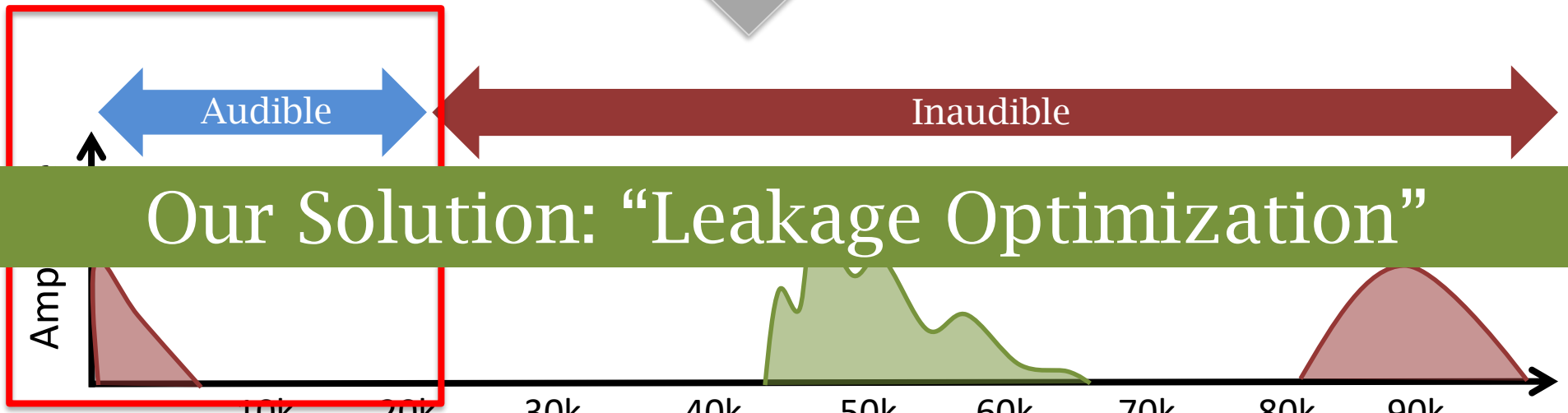




Speaker Nonlinearity

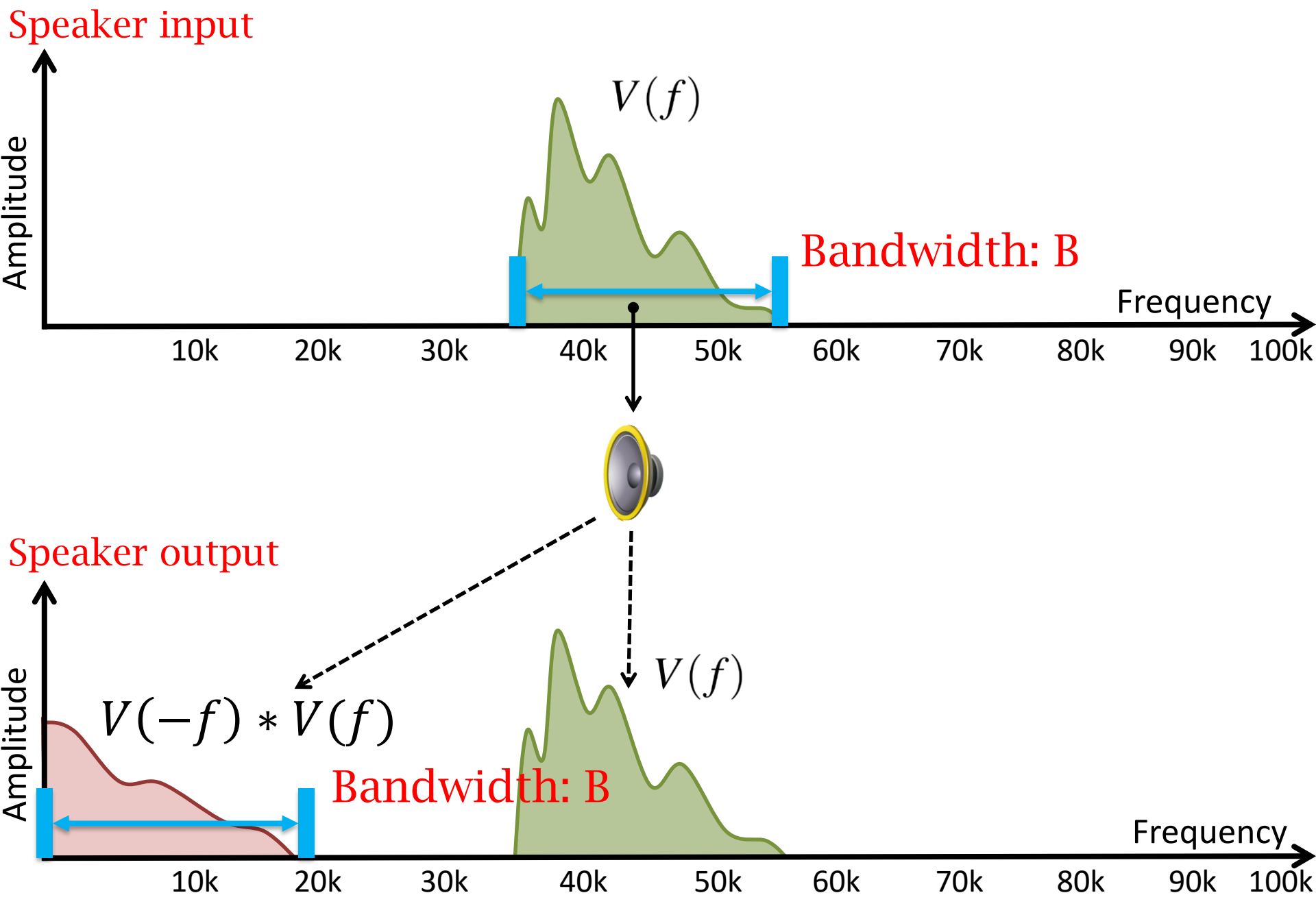
$$a_1 V_{in} + a_2 V_{in}^2$$

General to all speakers!

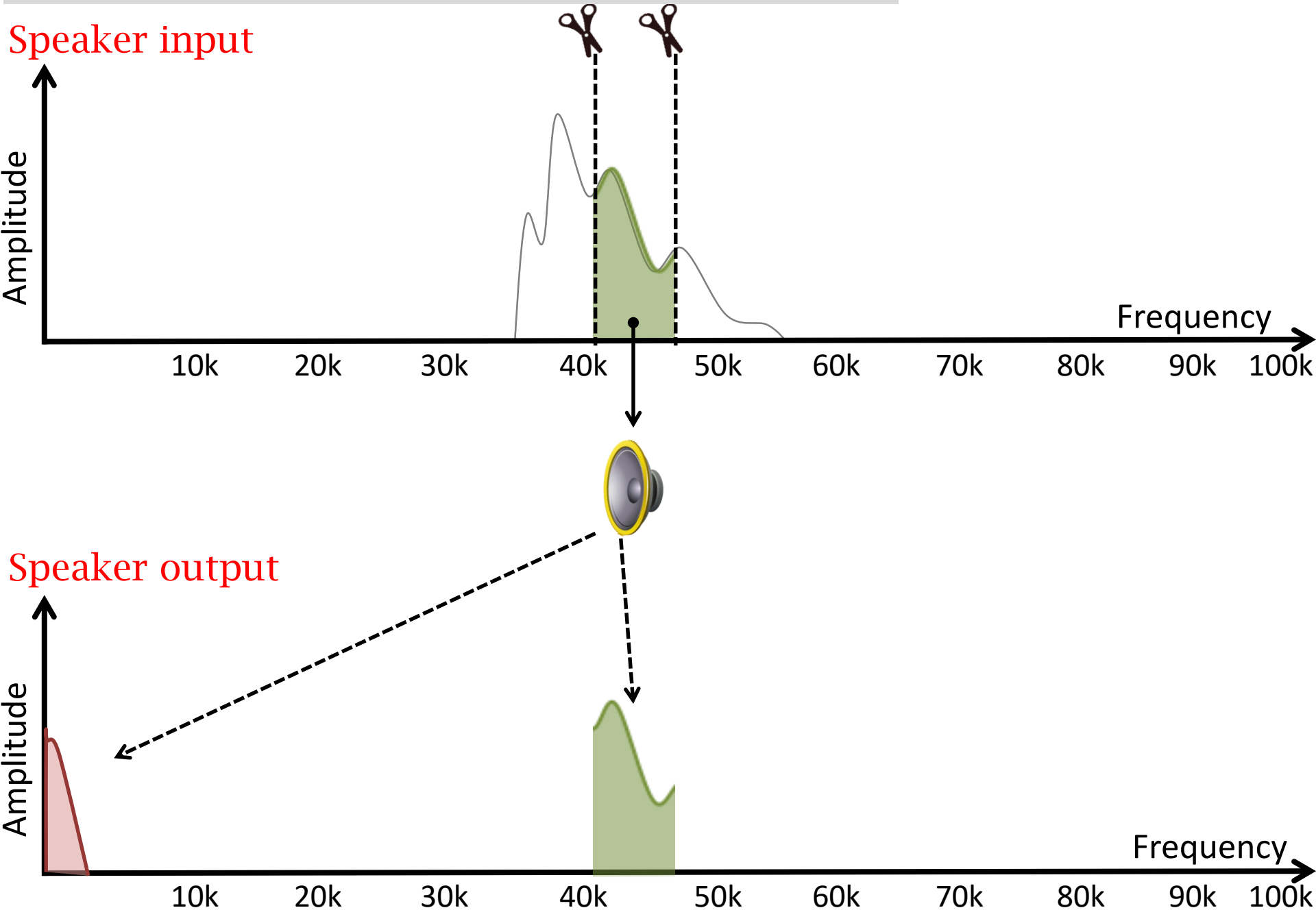


Our Solution: "Leakage Optimization"

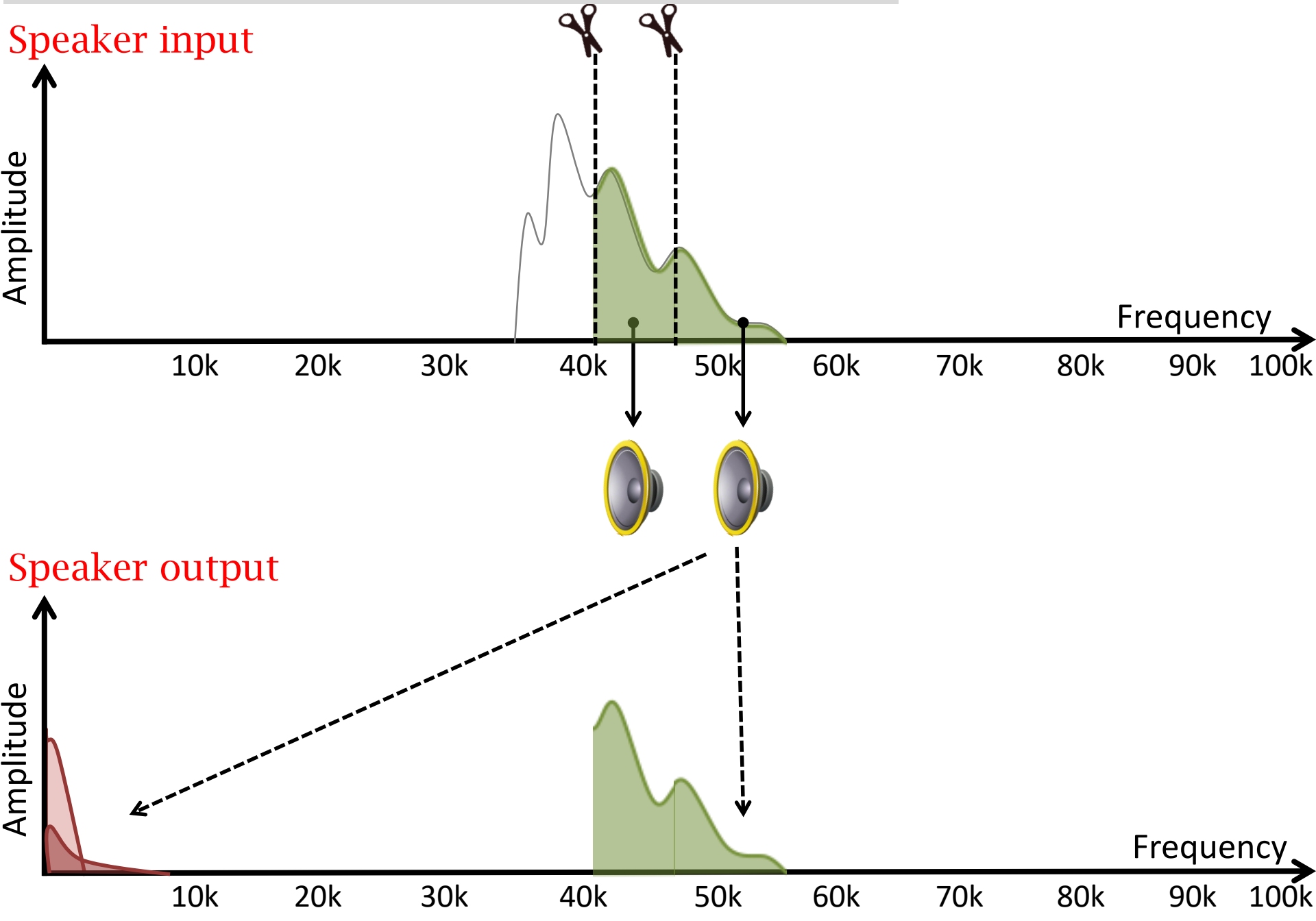
Speaker Nonlinearity → Audible Leakage



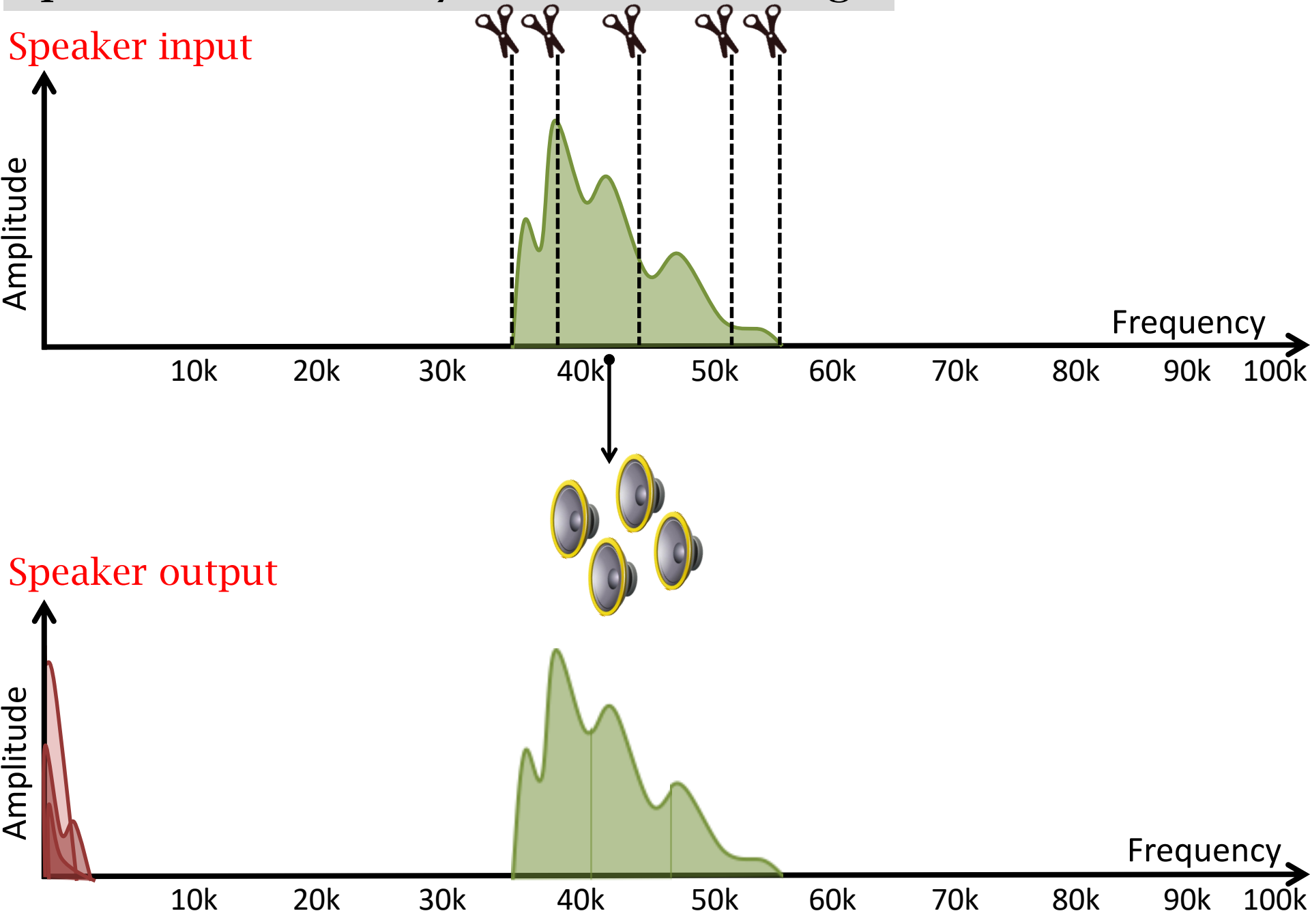
Speaker Nonlinearity → Audible Leakage



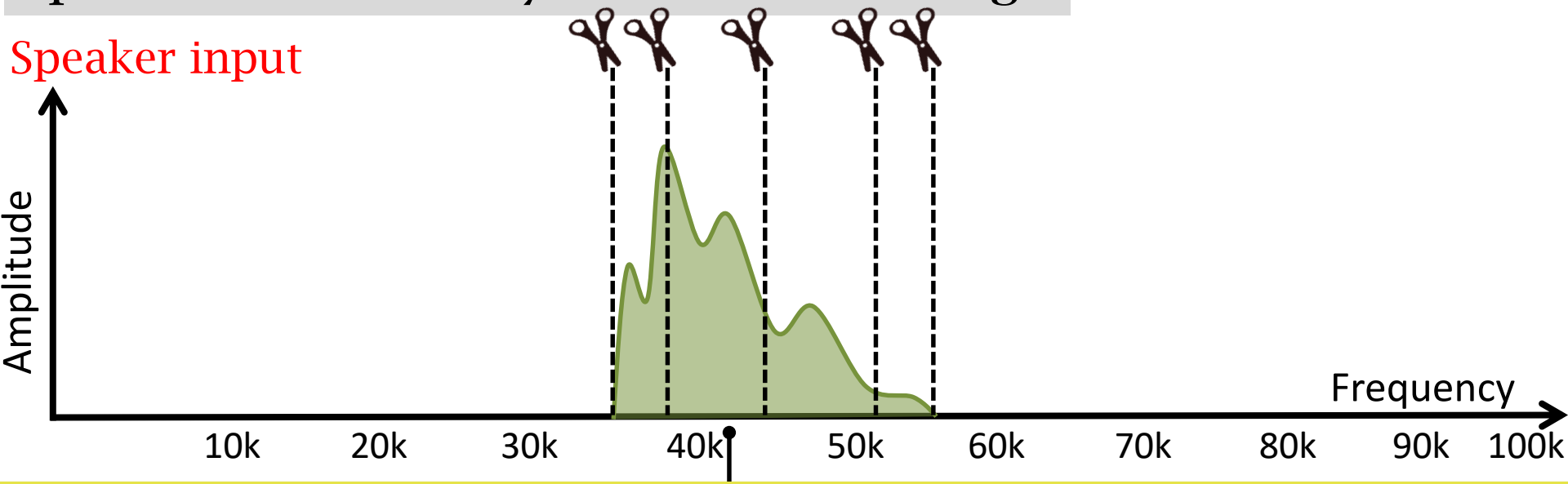
Speaker Nonlinearity → Audible Leakage



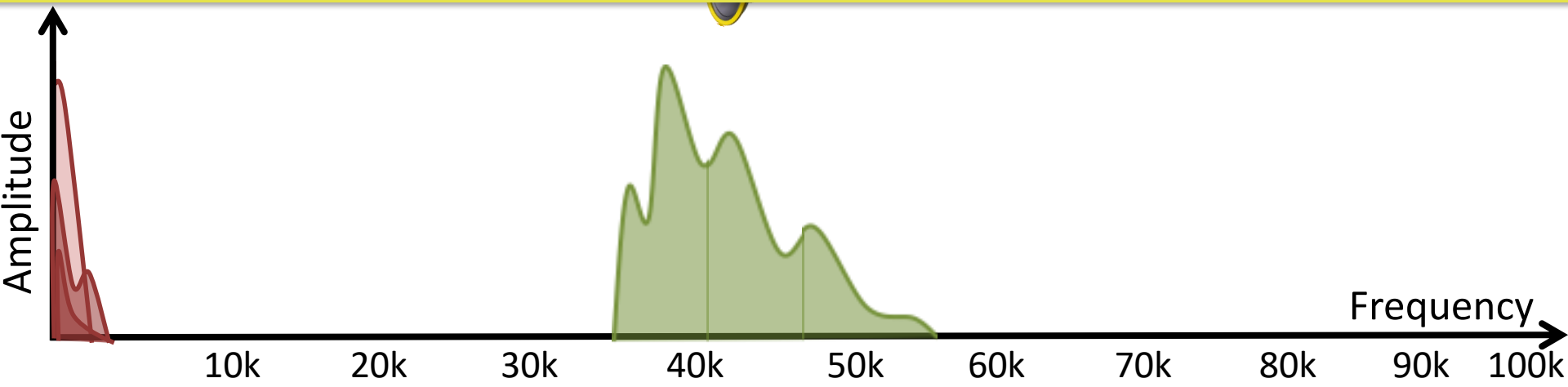
Speaker Nonlinearity → Audible Leakage

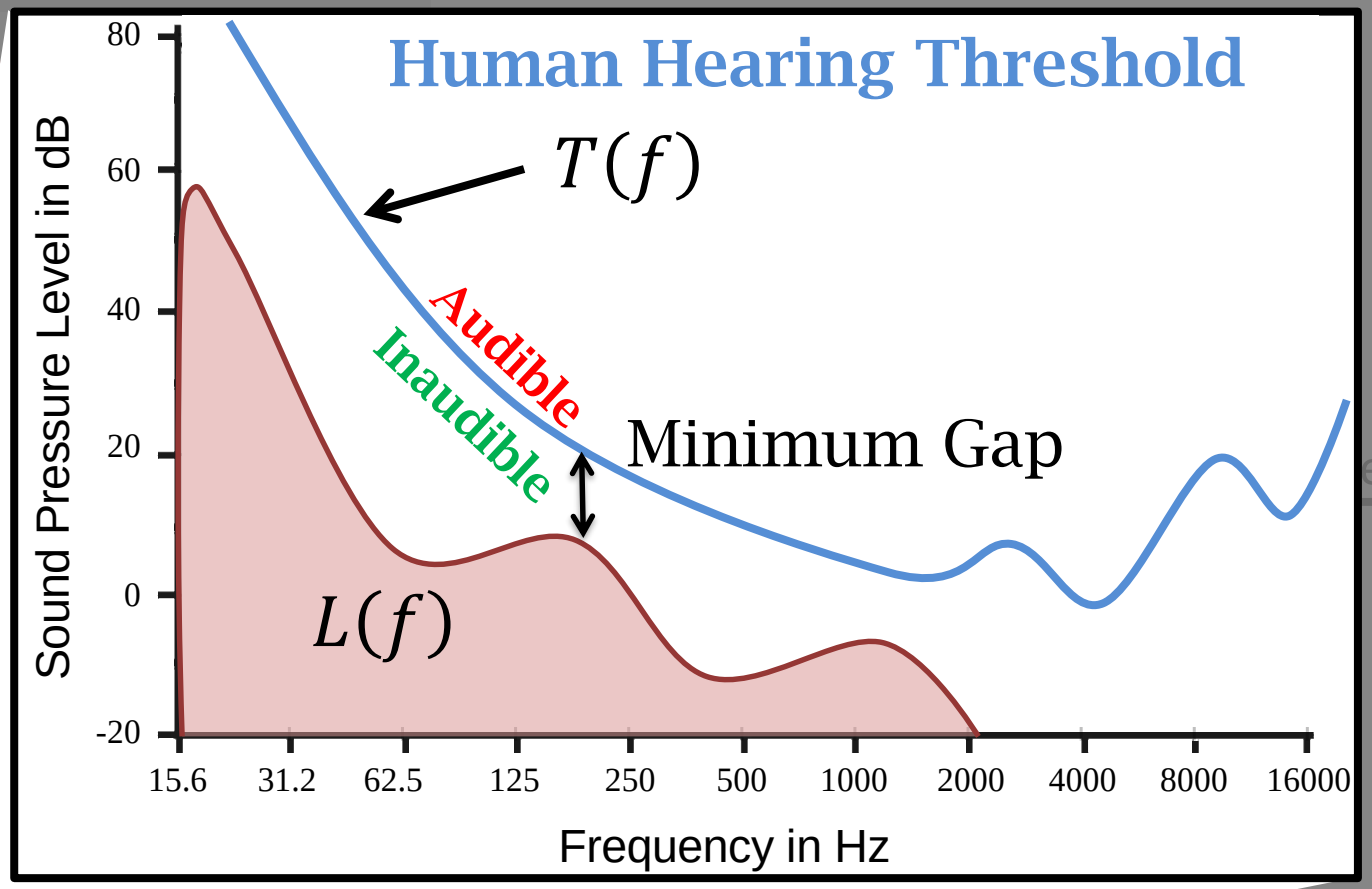


Speaker Nonlinearity → Audible Leakage



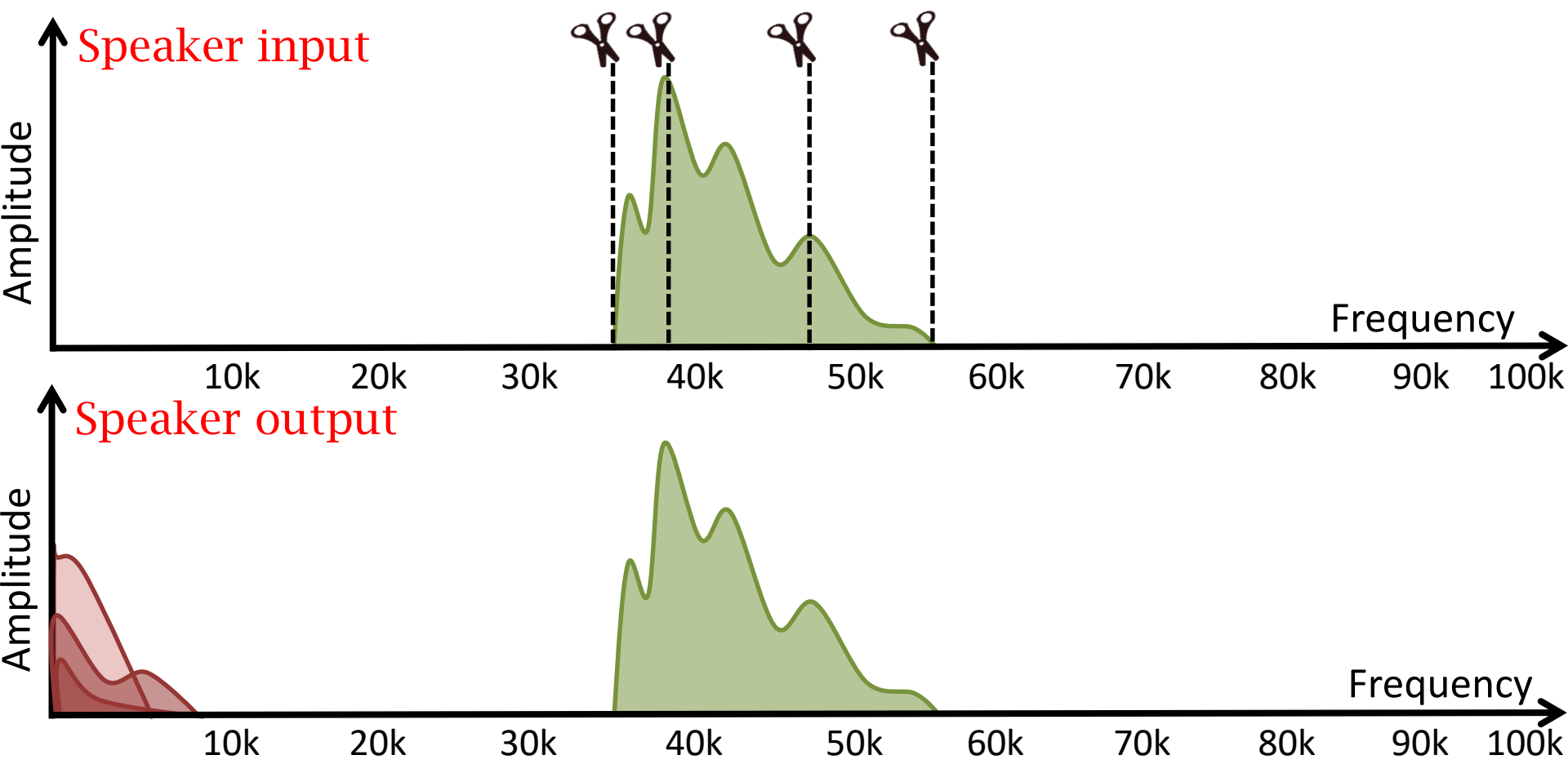
Chopping compresses the leakage band

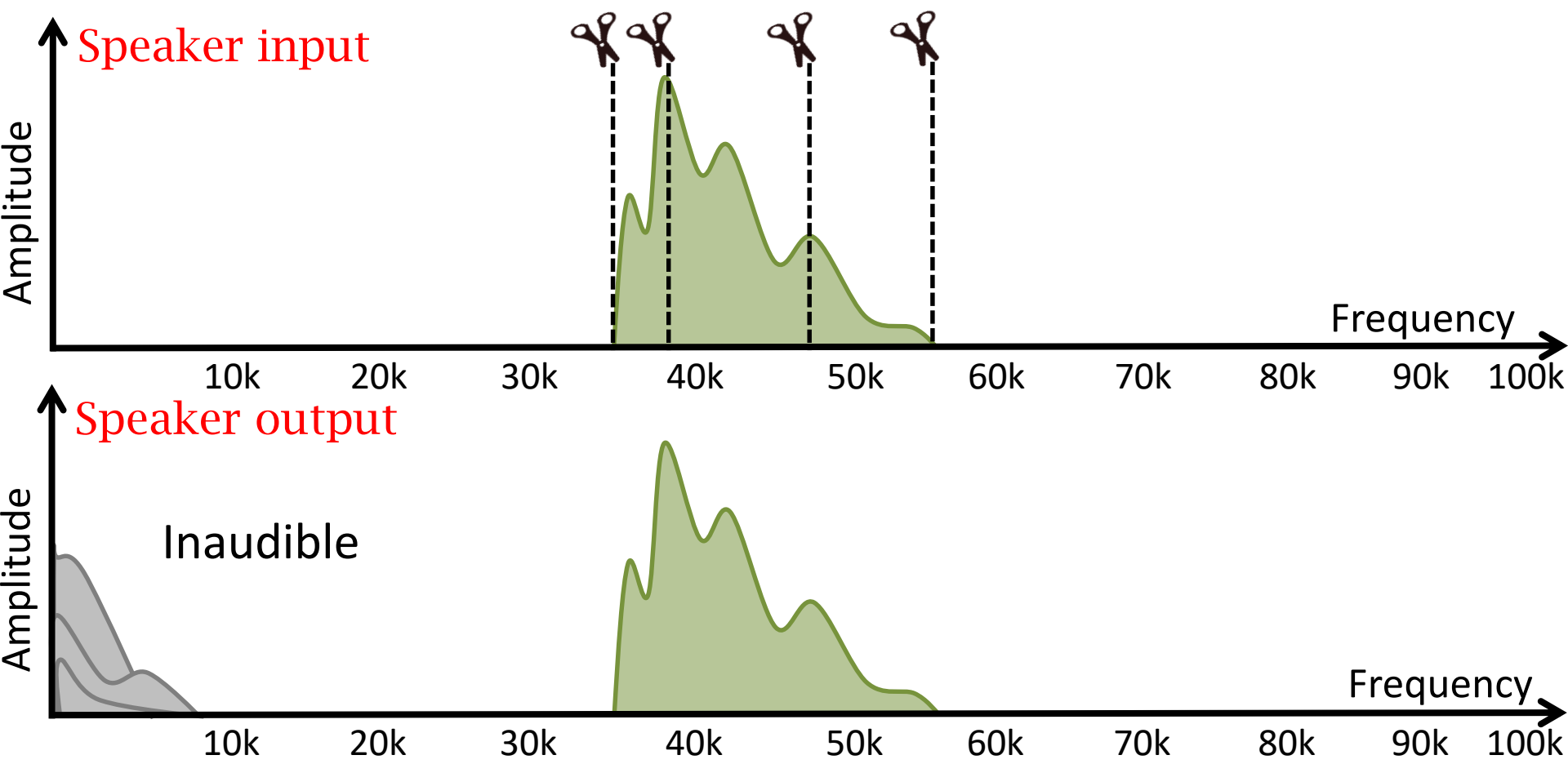


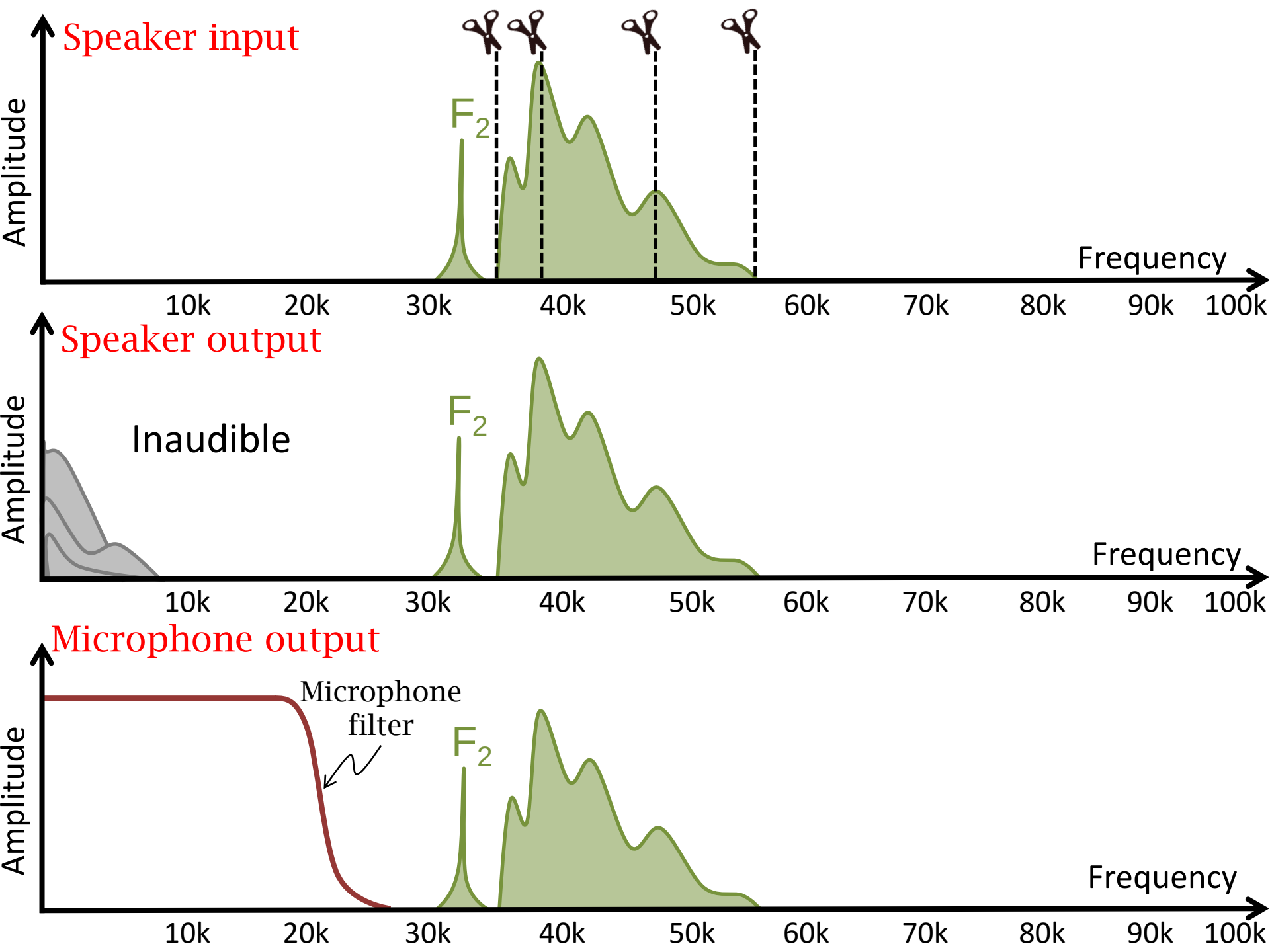


$$Maximize \quad \min_{f}[T(f) - L(f)]$$

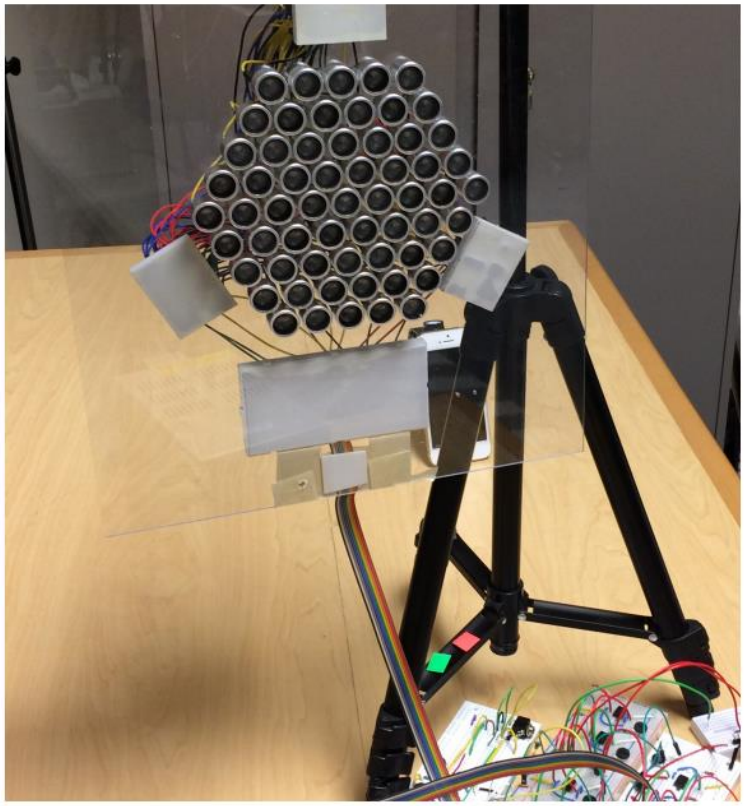
$$subject\ to \quad f_0 \leq f_1 \leq f_2 \leq \ \ldots \leq f_N$$

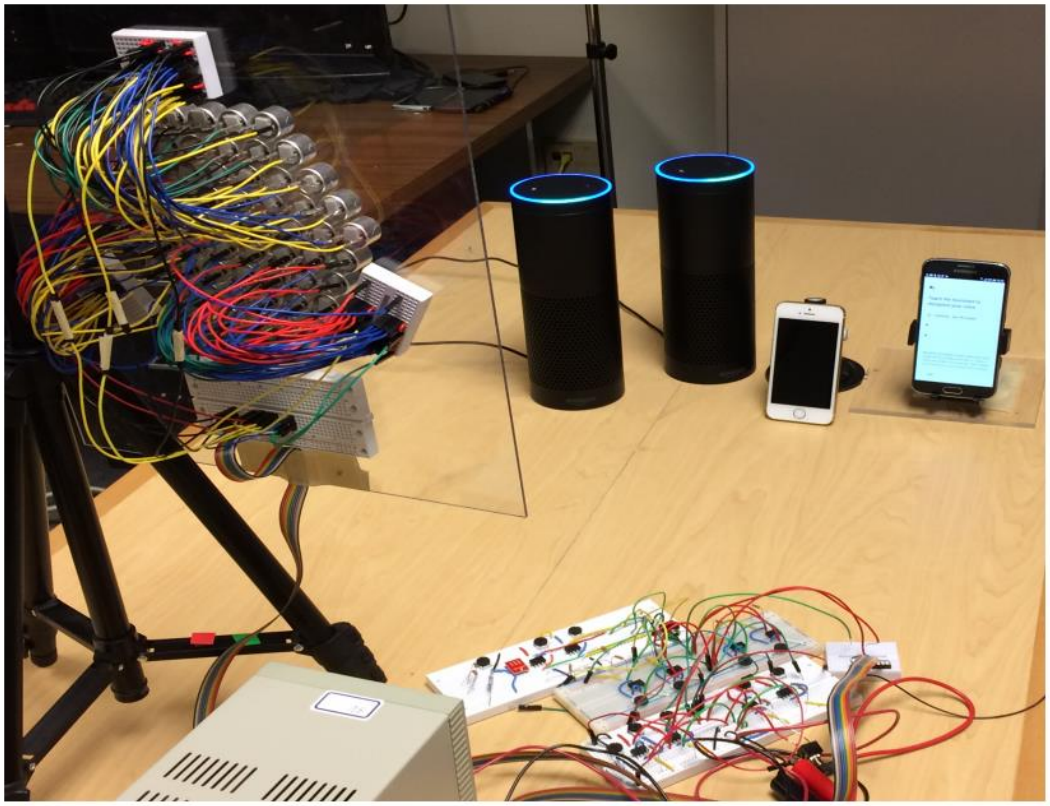

Speaker input
Amplitude
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Speaker output
Amplitude
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k


Speaker input
Amplitude
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Speaker output
Inaudible
Amplitude
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k

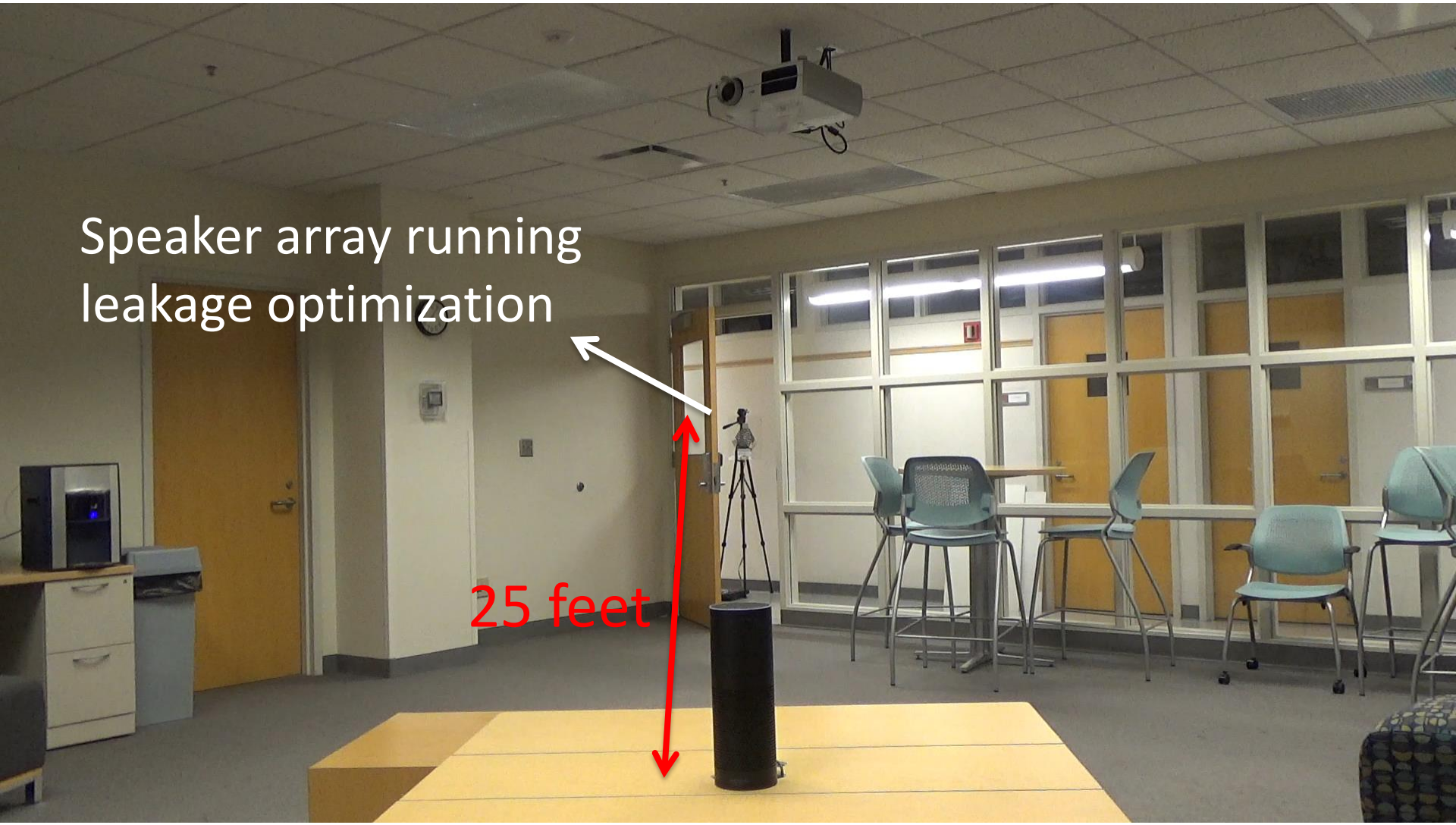

Speaker input
Amplitude
F2
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Speaker output
Amplitude
Inaudible
F2
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Microphone output
Amplitude
Microphone
filter
F2
Frequency
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k

Evaluation



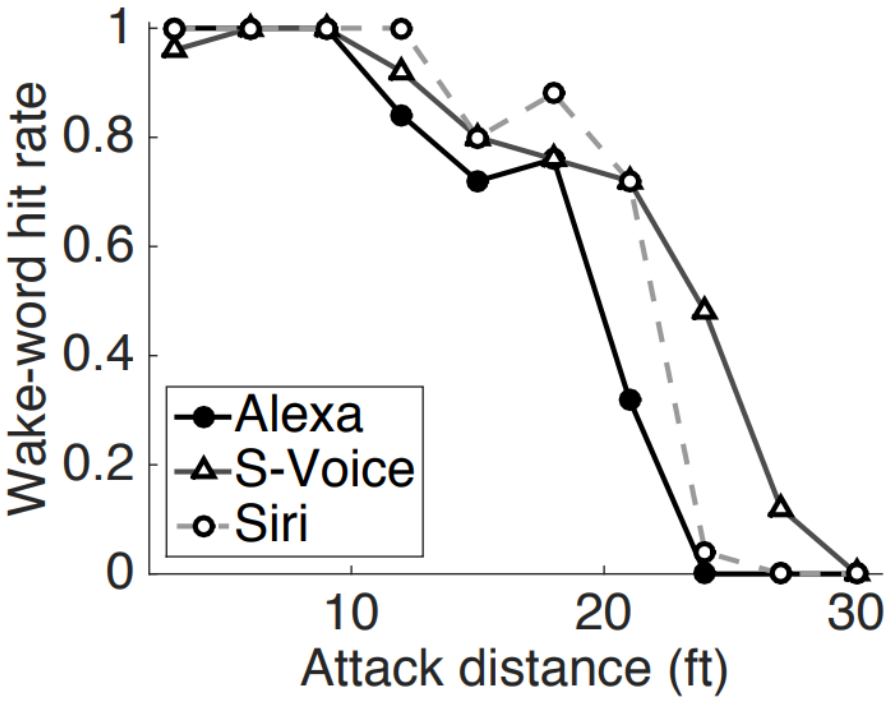


Inaudible voice commands: Long range



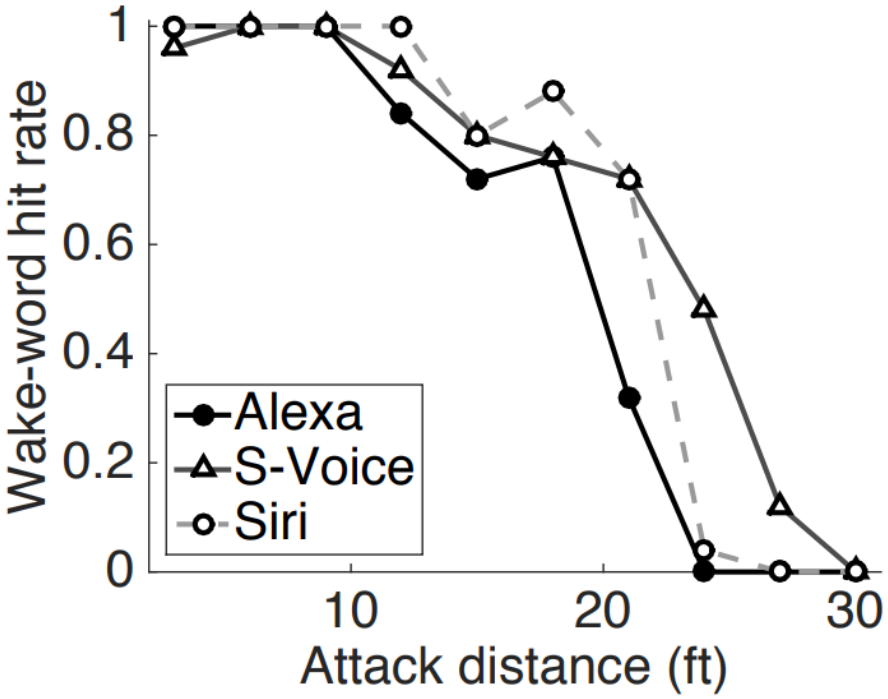
Evaluation

Wake-word hit rate

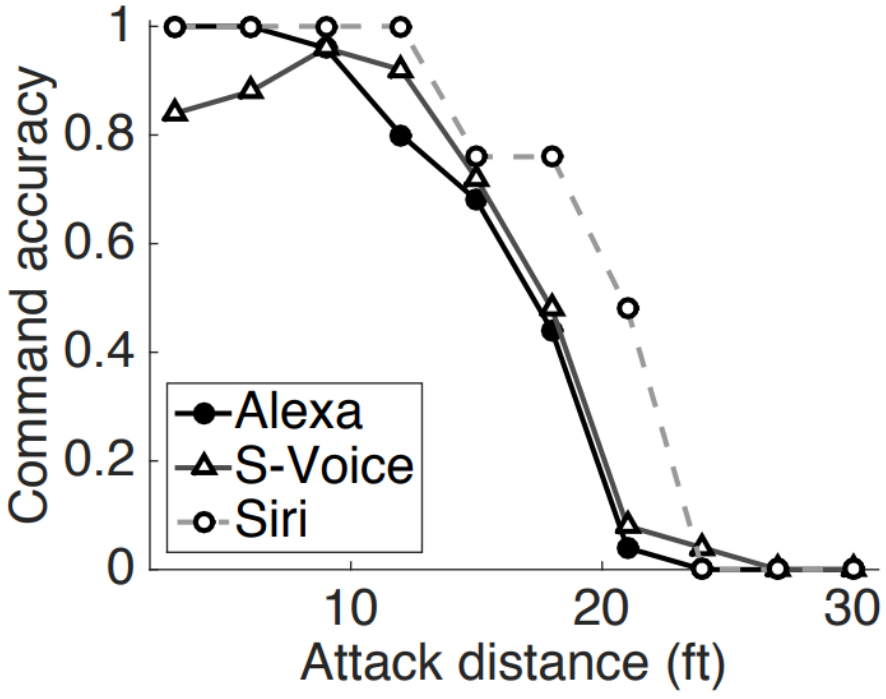


Evaluation

Wake-word hit rate

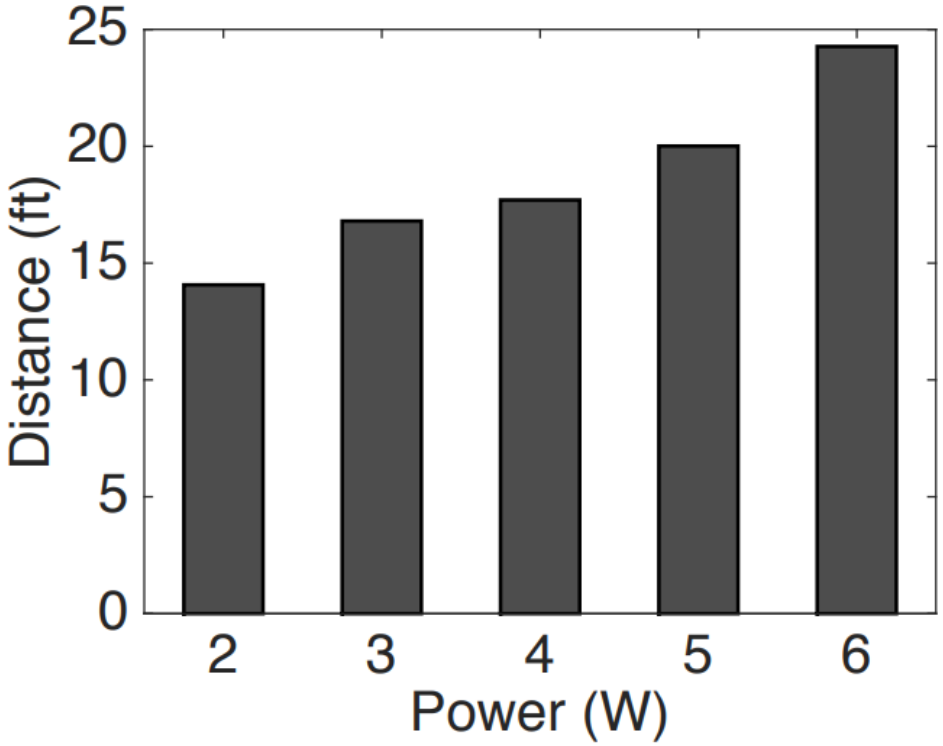


Command detection accuracy



Evaluation

Maximum activation distance for different input power



Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

Today's Talk:

1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?

Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

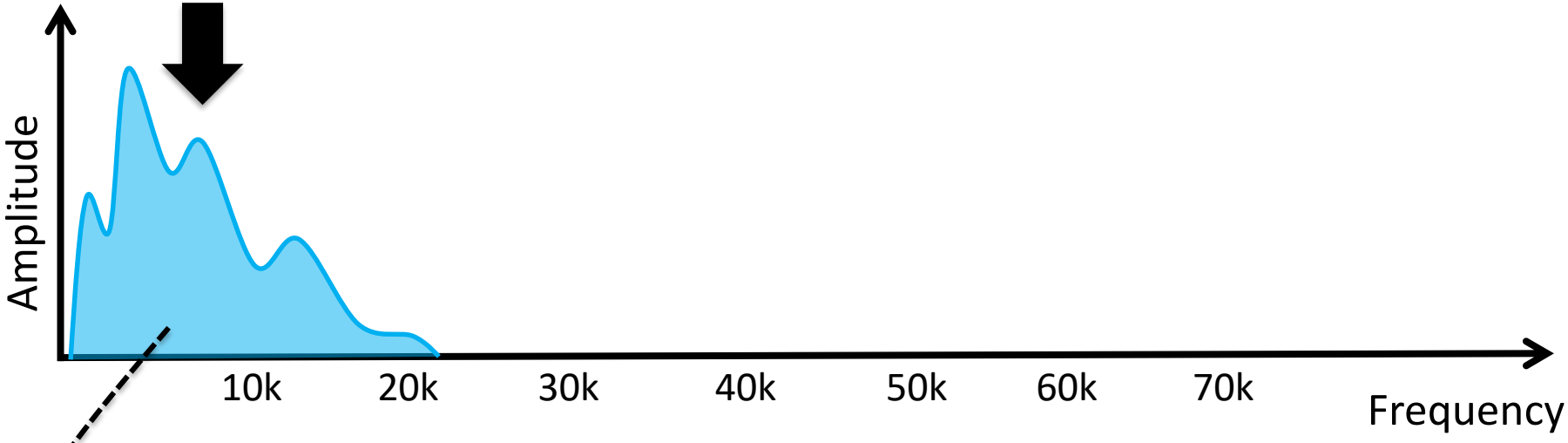
Today's Talk:

1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?

Core Question:

Is this a “non-linear signal” or normally recorded signal



?

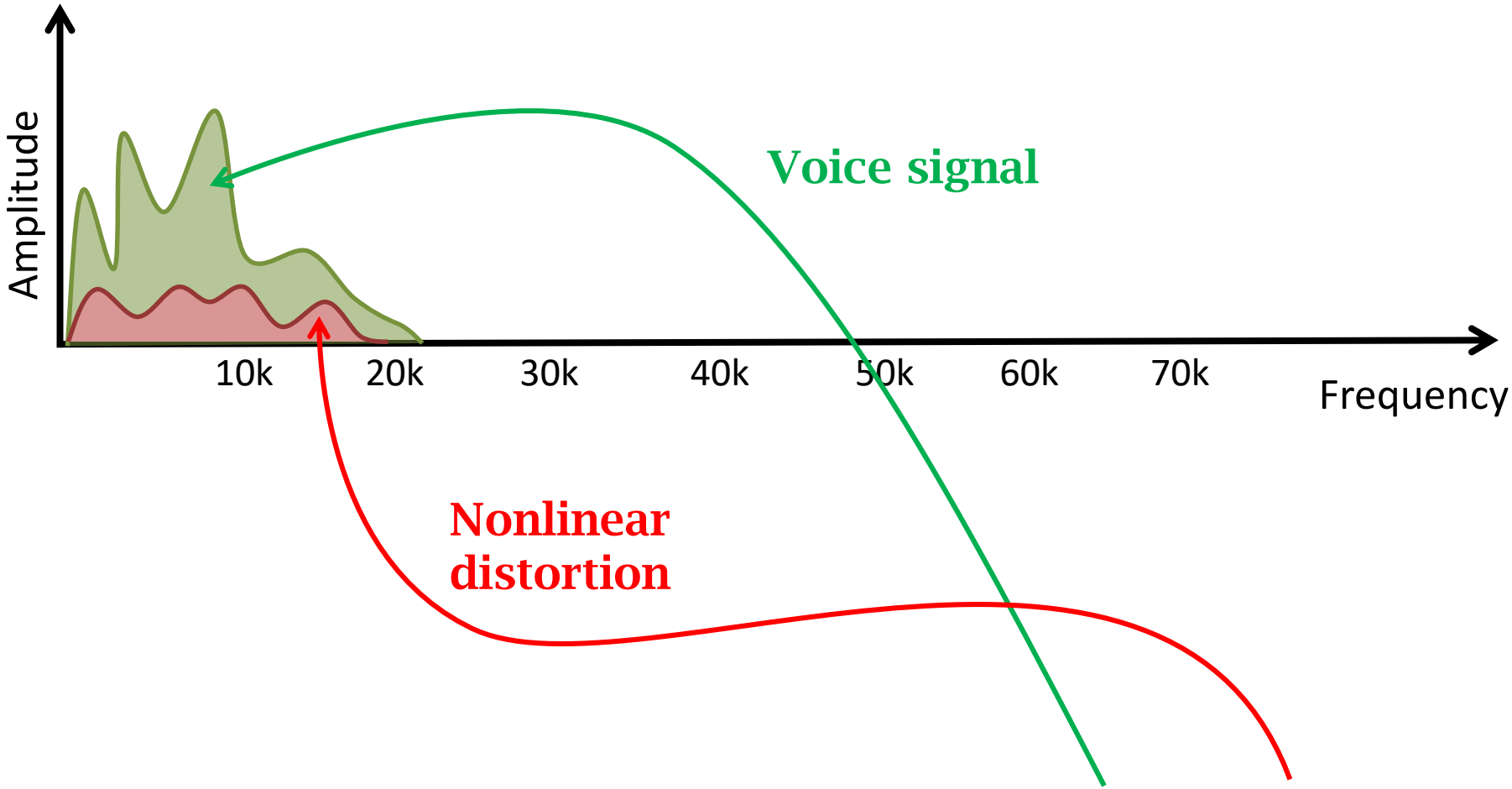
Voice signal $v(t)$

Inaudible Voice Attack

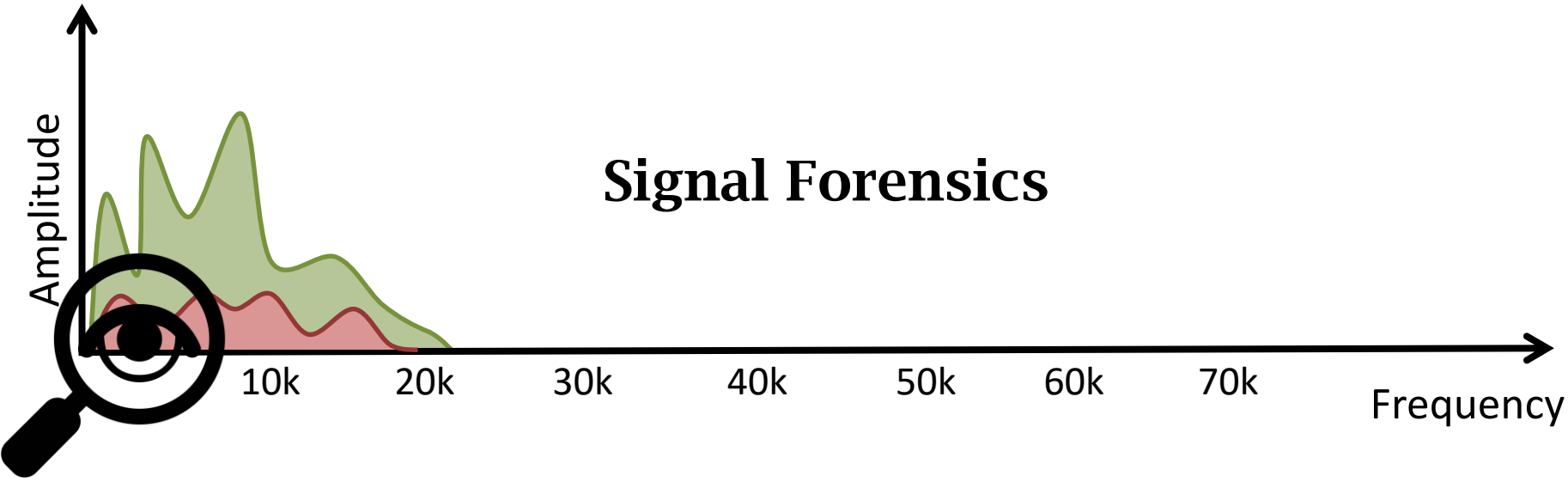
$$[v(t) \cdot \sin(\omega_1 t) + c \cdot \sin(\omega_1 t)]^2 = v(t) + c' v^2(t) + \cdots$$

Core Question:

Is this a “non-linear signal” or normally recorded signal



$$[v(t) \cdot \sin(\omega_1 t) + c \cdot \sin(\omega_1 t)]^2 = v(t) + c'v^2(t) + \cdots$$

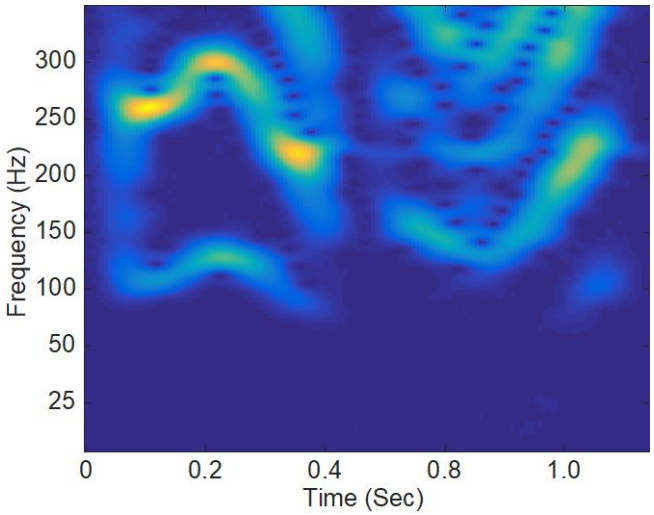


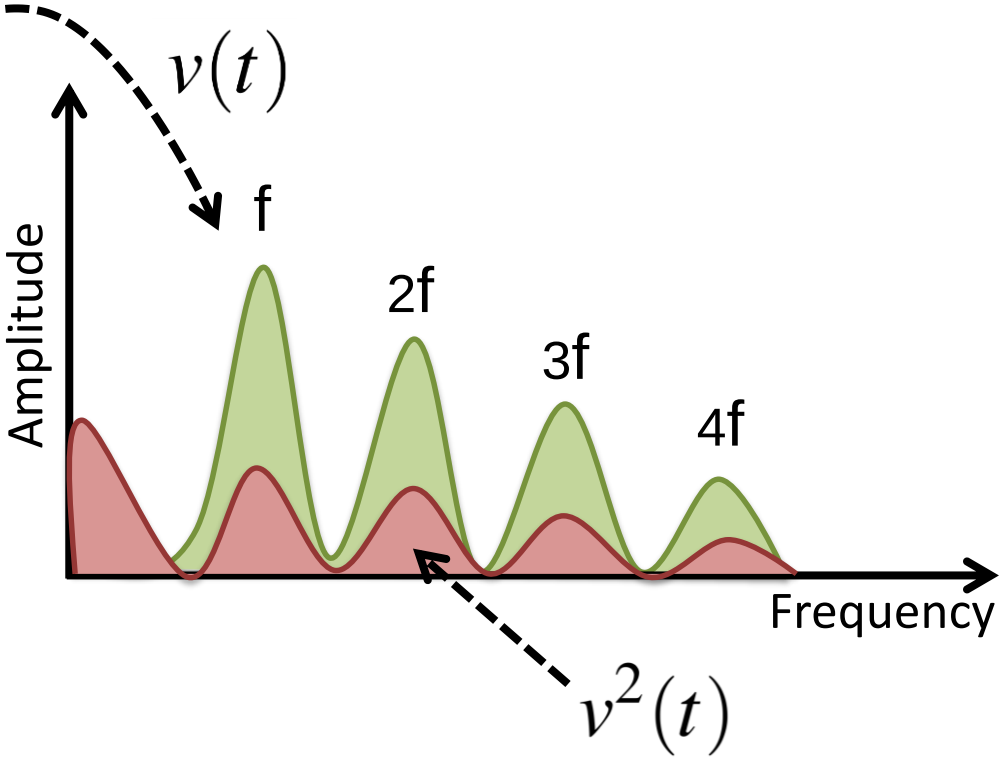
Difficult to decouple “voice signal” and “non-linear signal”

Human voice signals present opportunities …

Opportunity #1: Voice > 50 Hz

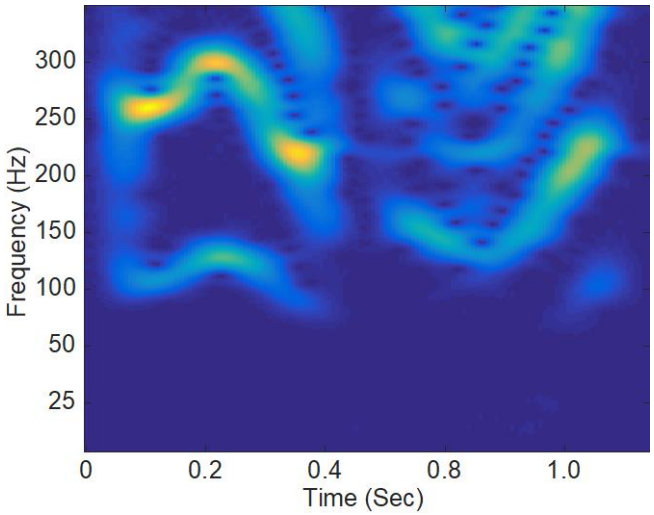
Human Voice

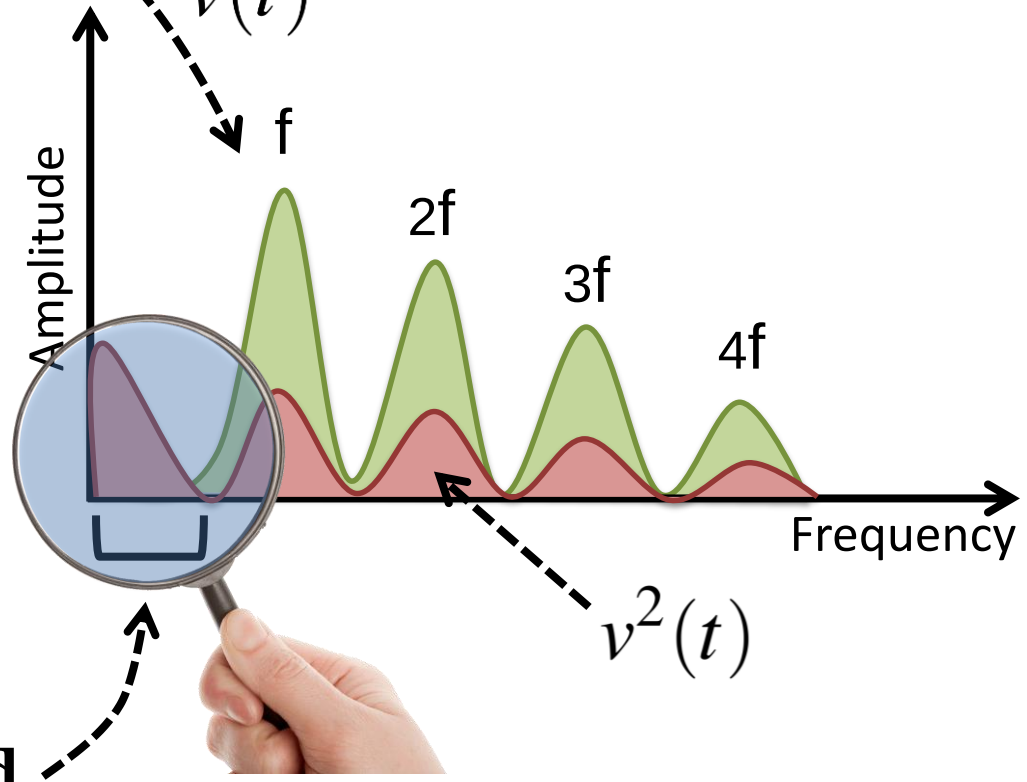




Opportunity #1: Voice > 50 Hz

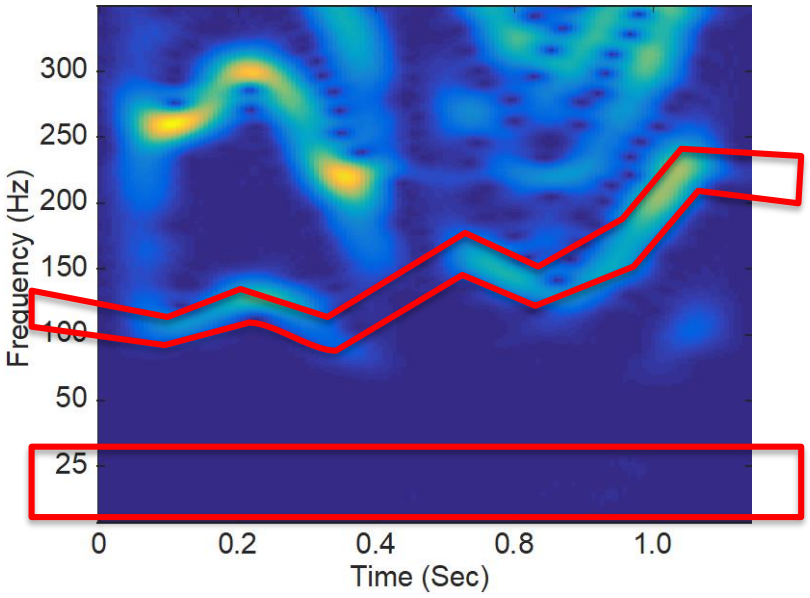
Human Voice

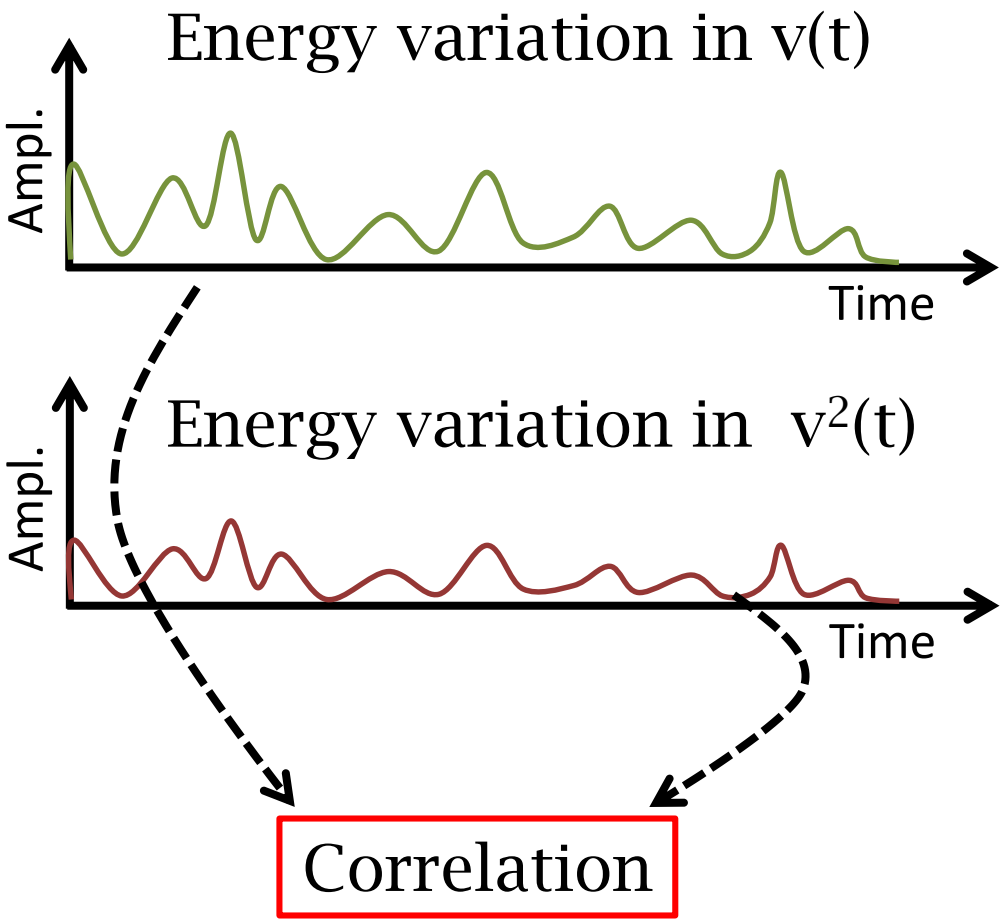




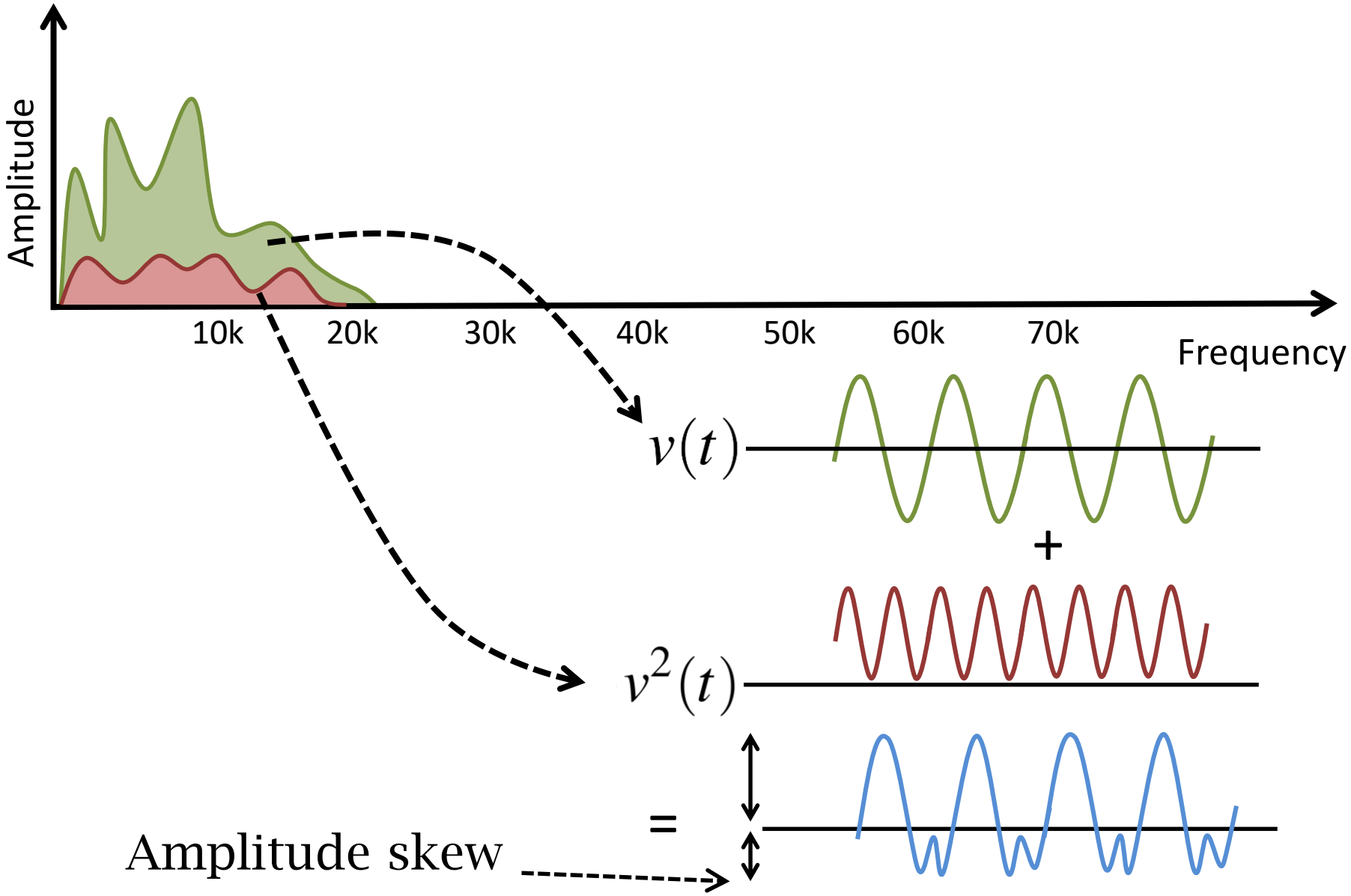
Energy at sub-50Hz band

Opportunity #2: Correlation

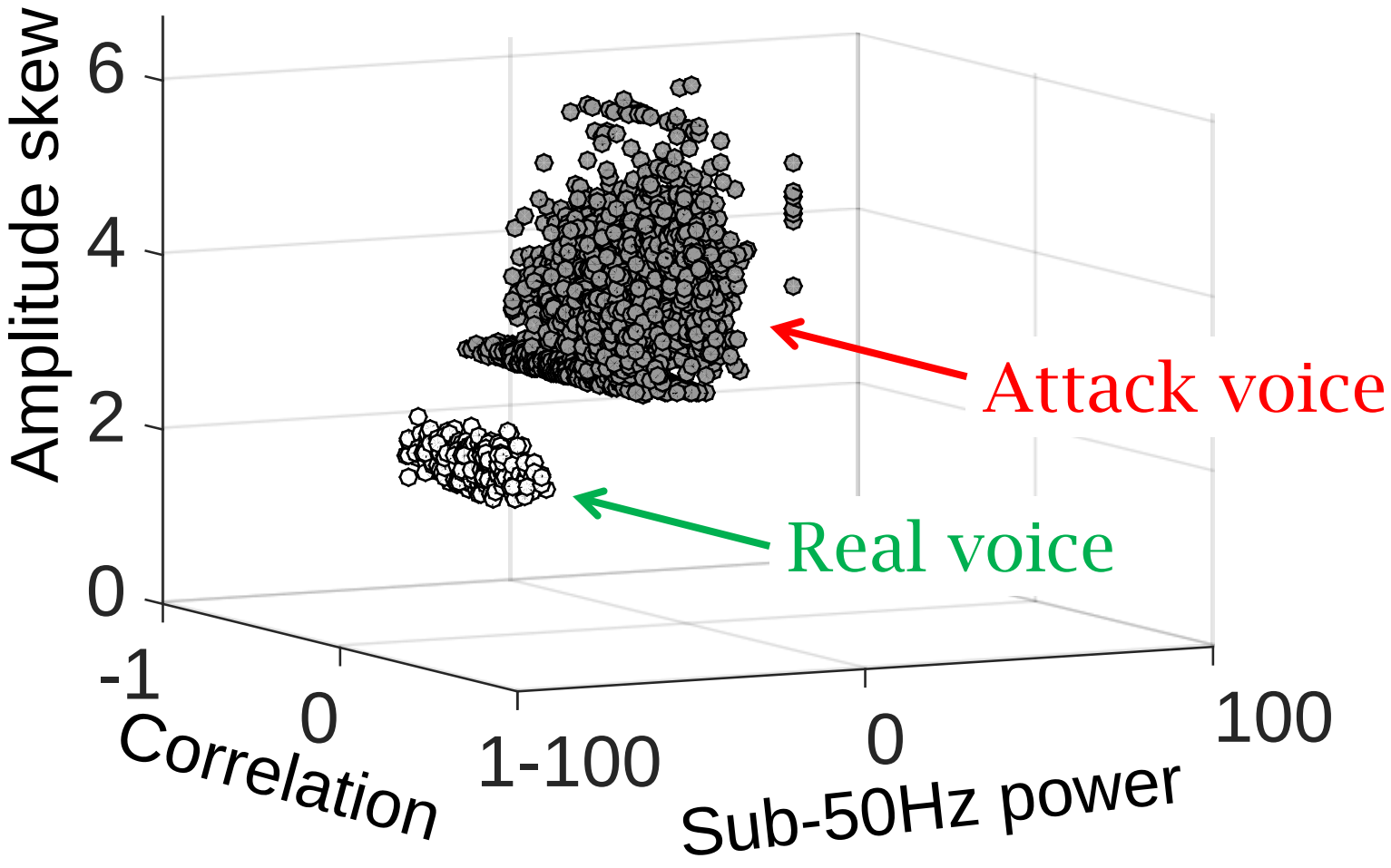




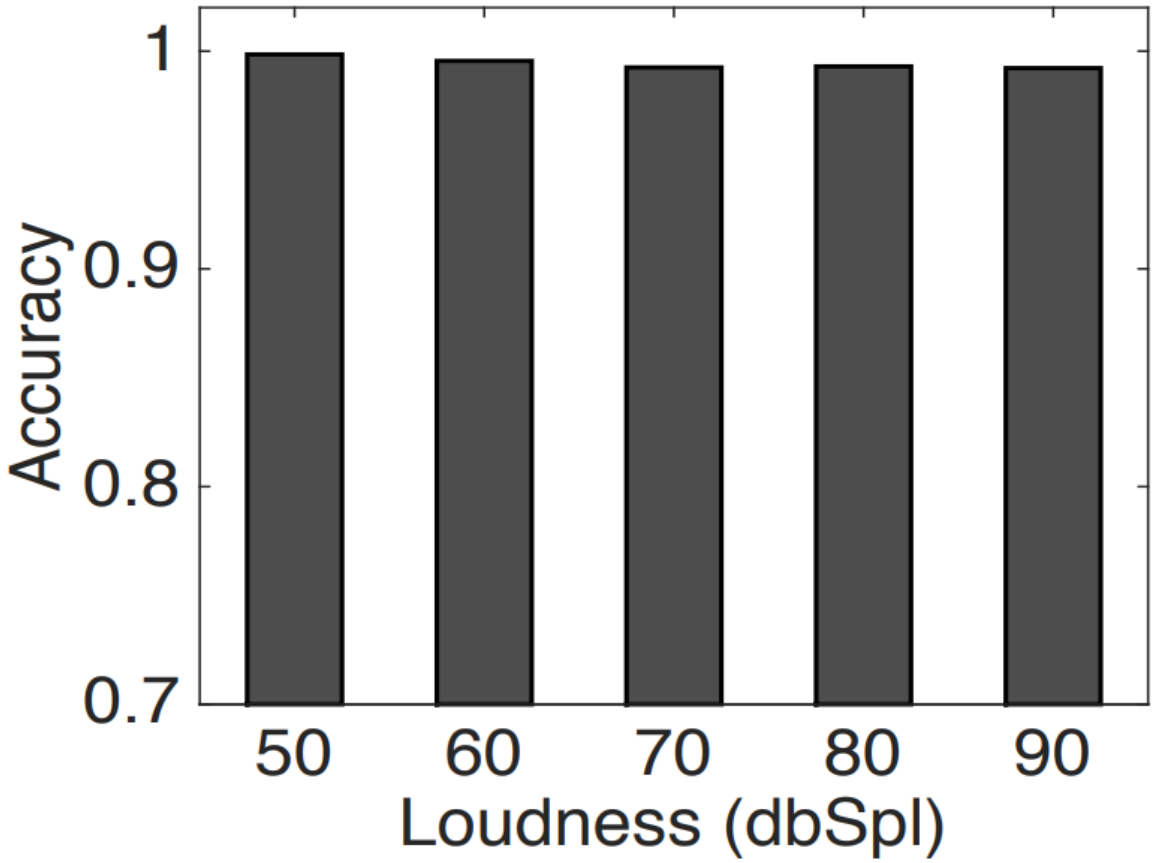
Opportunity #3: Amplitude Skewness



5000 Test Cases



Overall Detection Accuracy



To summarize…

Inaudible Acoustics (> 25 kHz): **"Alexa, open the garage door!"**