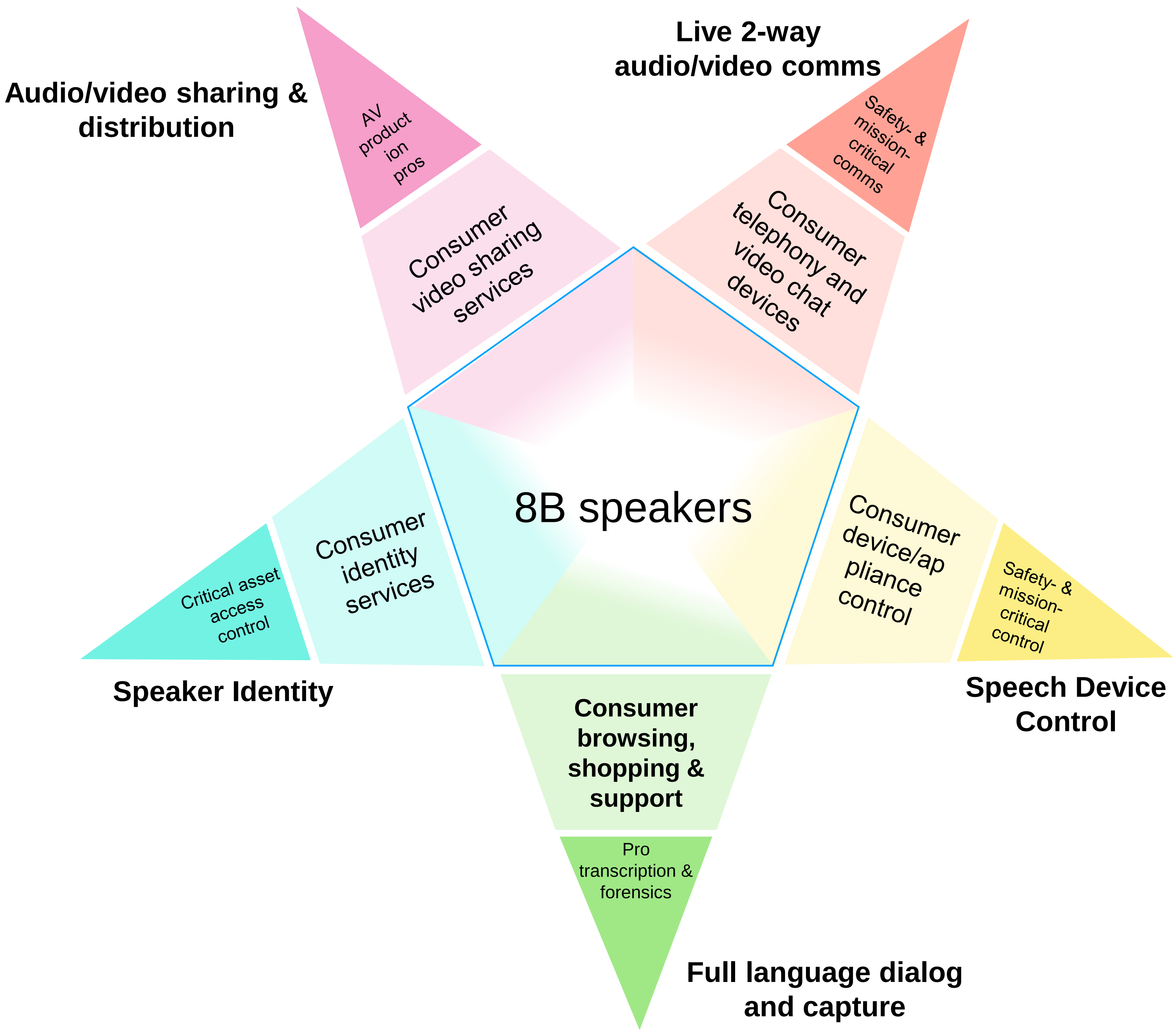
Samer Hijazi

CTO

BabbleLabs Inc.

The New World of Speech Technology



babblelabs

The Opportunity

- 22B microphones by 2020
- 7B phones + radios + TVs delivering voice
- YouTube uploads: 13B minutes per year
- 200T minutes per year of device interaction
- 1Q words per year in voice calls

Speech Market Growth: 38.3%

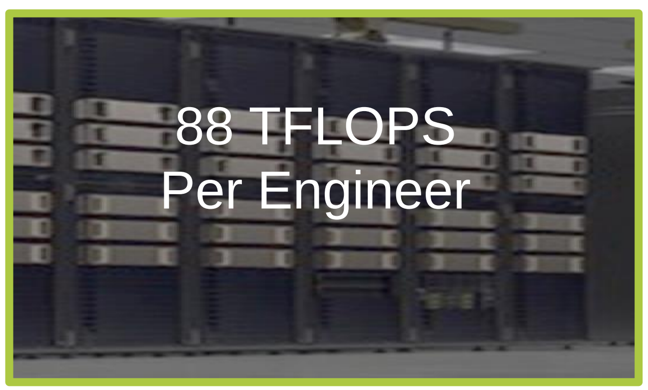
-Statista 2018: speech recognition technology market 2016-2014

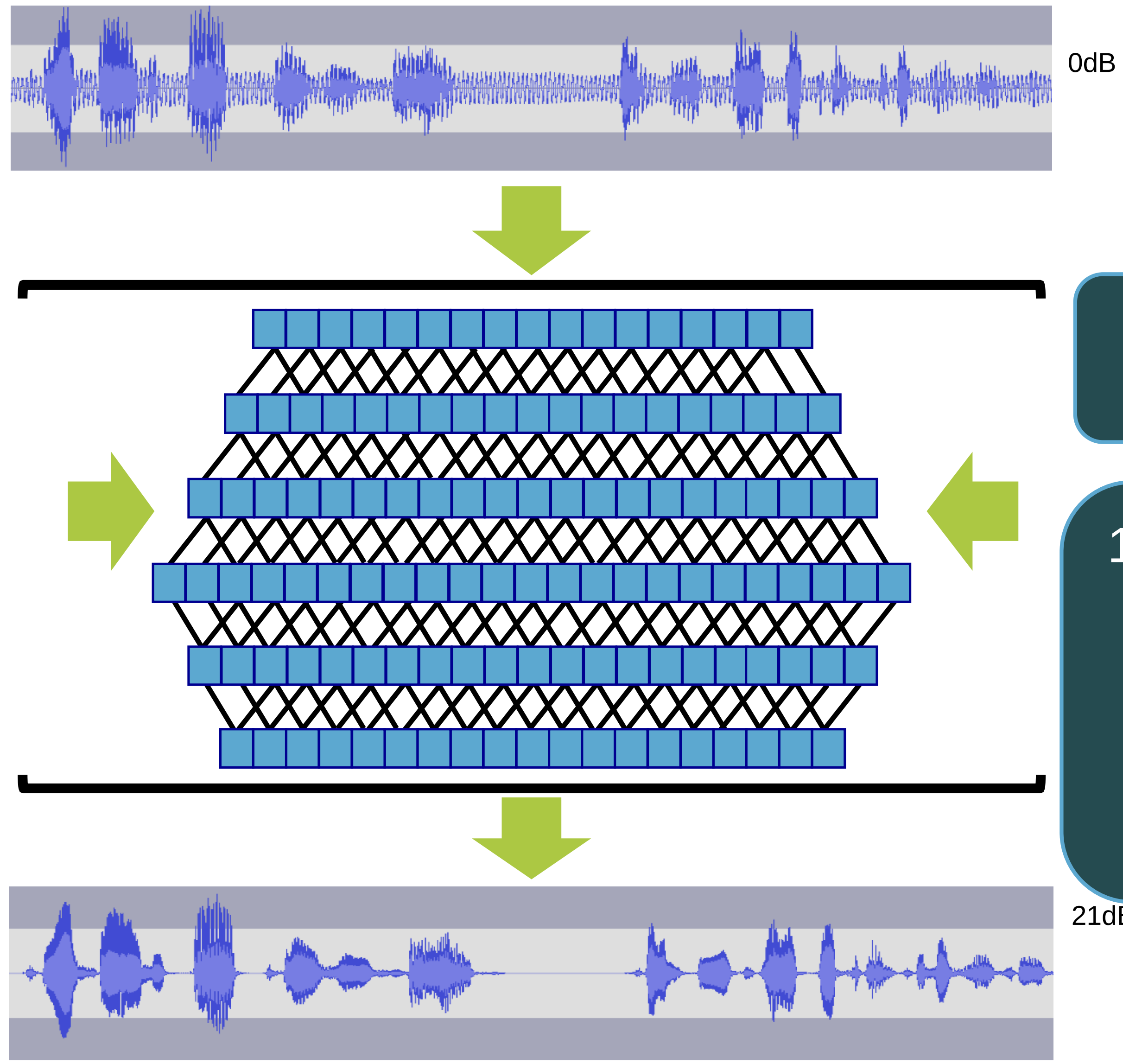
babblelabs

AI meets speech

more sophisticated models, more data, more training

Massive Compute





Massive Data Corpus

10s of 1000 hr speech
10s of 1,000 hr noise
10s of 1000 RIR
NEVER TRAIN ON THE SAME DATA TWICE

Speech Enhancement **Speech Recognition** **Speech UI Dialog**

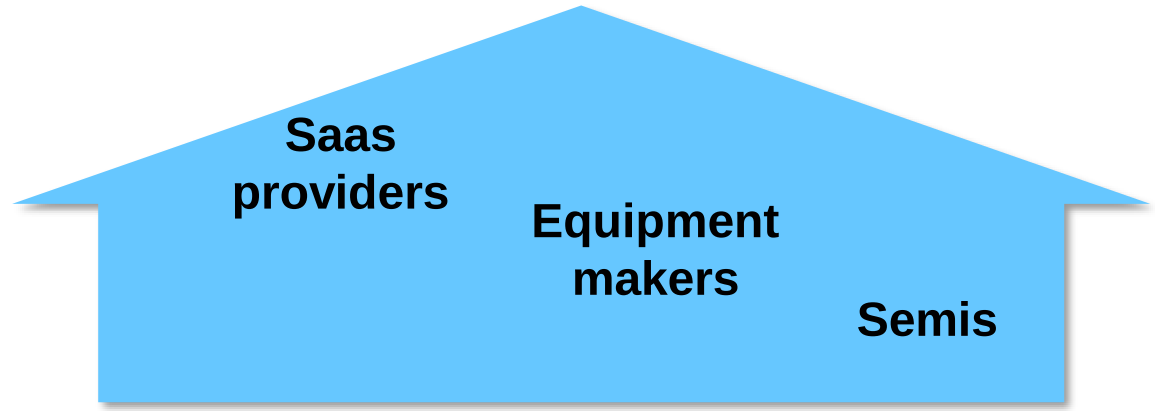
babblelabs

Technology ➔ Product ➔ Customers ➔ End Users

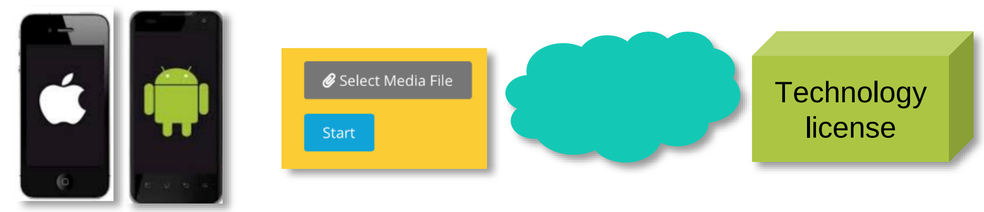
End user:
their problem

Consumer audio/video sharer:
Recording in the real world

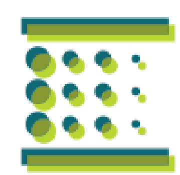
Customers



Product: Platform-optimized solutions

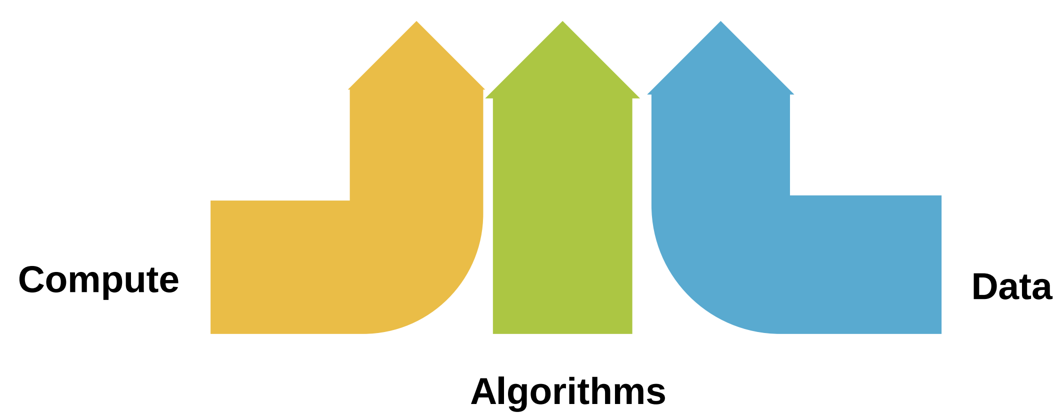


Product: Deep learning speech software



Speech Enhancement

Technology: Unique data-sets and training



babblelabs

Clear Speech Everywhere

In production for real-world video sharing, production, streaming, and audio

Foundation to product release in 28 weeks!

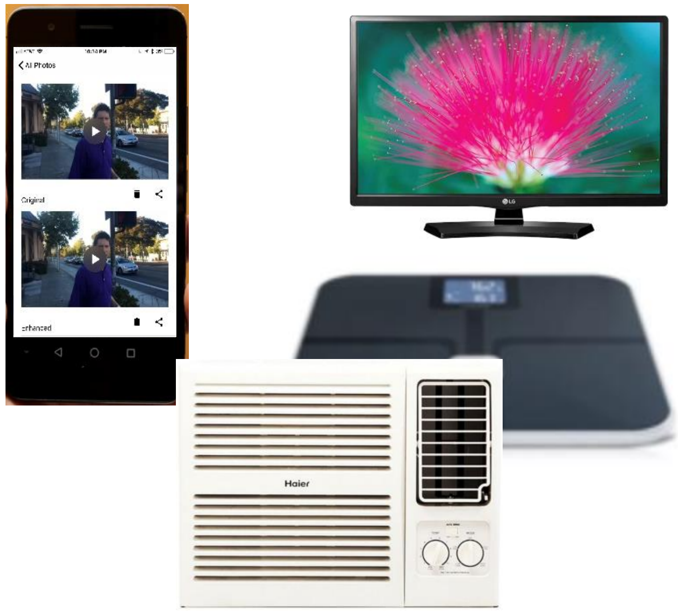
Common product delivered across platforms

Primary products

Cloud API

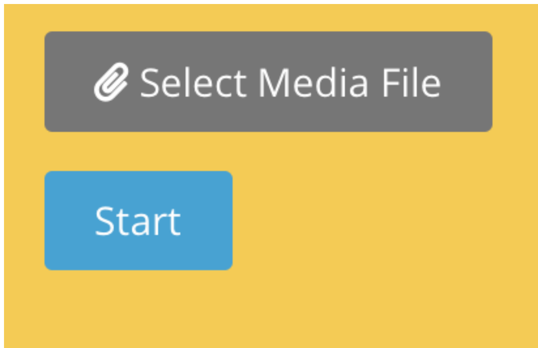
```
host=api.babblelabs.com
token=$(cat token.txt) #token.txt
created by login.sh
email="<yourEmailAddr>"
api="audio" # audio or video
contentType="audio/wav" # a normal
supported mime-type of audio or
video
for infile in *.wav # Replace this
with a different pattern, or a list
of file names (of the same type)
do
  (
  set -x
  time  curl --progress-bar \
    --data-binary @$infile \
    --dump-header $infile.hdr \
    -H "Content-Type":"$contentType"
    \
    -o $infile.denoised \
    -H "Authorization: Bearer
    $token" \

    "https://$host/audioEnhancer/api
    /$api/stream/$email"
  ) &
done
```

On-Device License

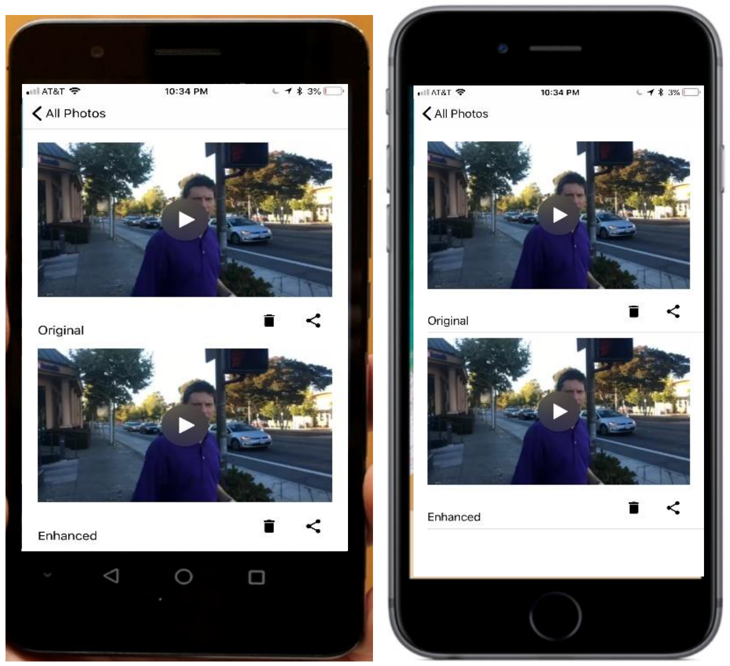


For visibility and demonstration

Web UI



Android & iPhone Apps



babblelabs

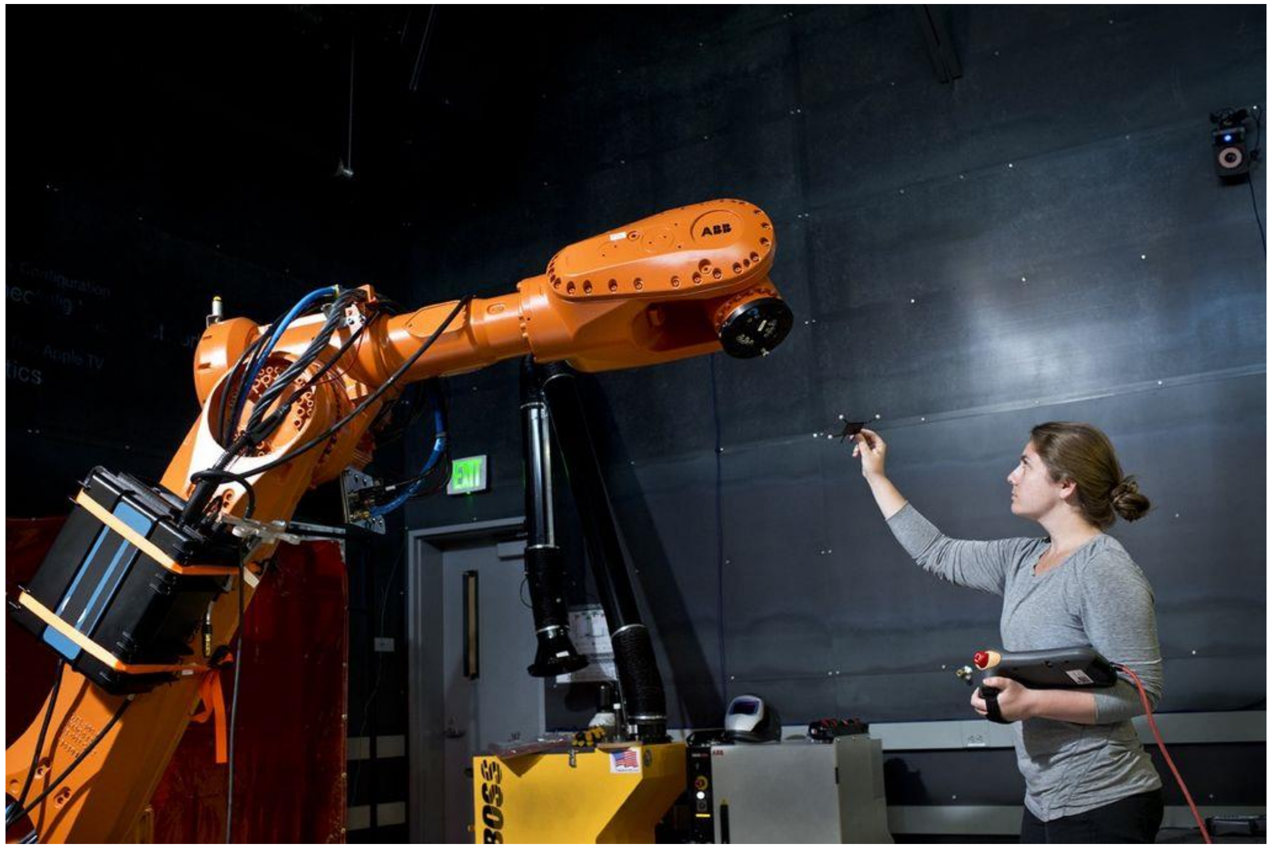
babblelabs

What is Speech Ennhancment

Human – Human Interface Challenges

babblelabs

Human – Machine Interface Challenges



babblelabs

BabbleLabs Answer to these Challenges: Clear Cloud™

Noisy

Enhanced

babblelabs

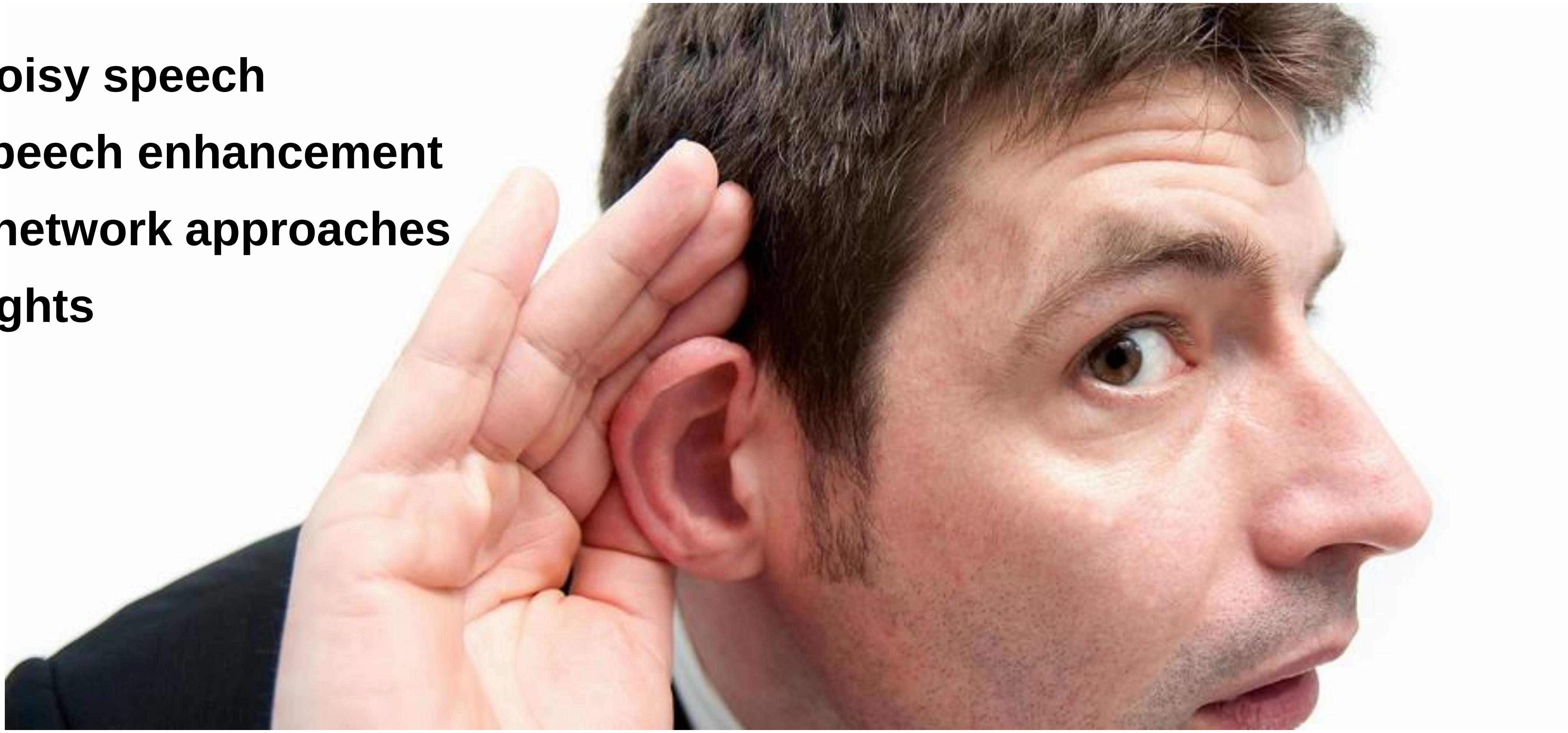
Outline

A bit about noisy speech

Traditional speech enhancement

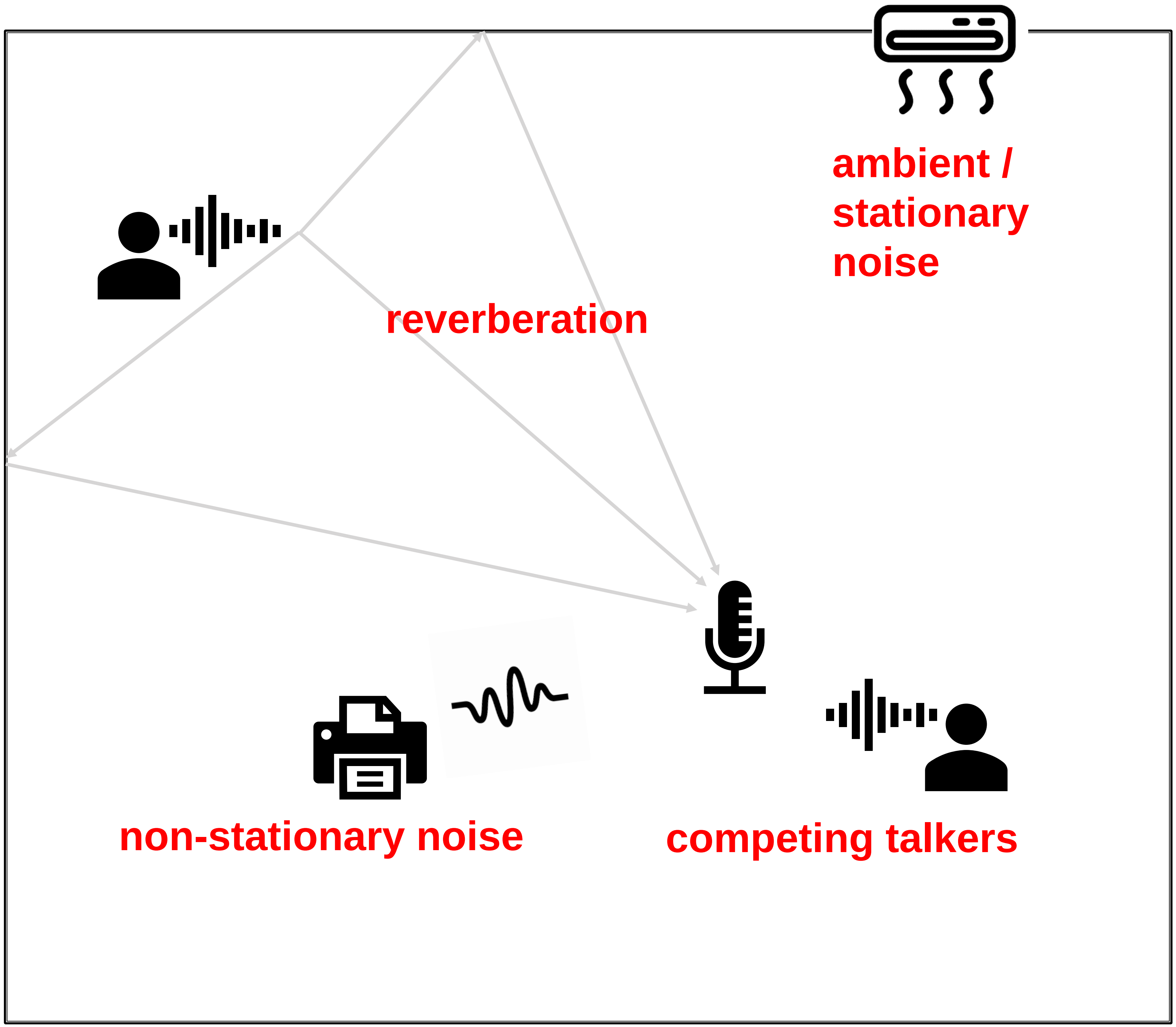
Deep neural network approaches

Closing thoughts



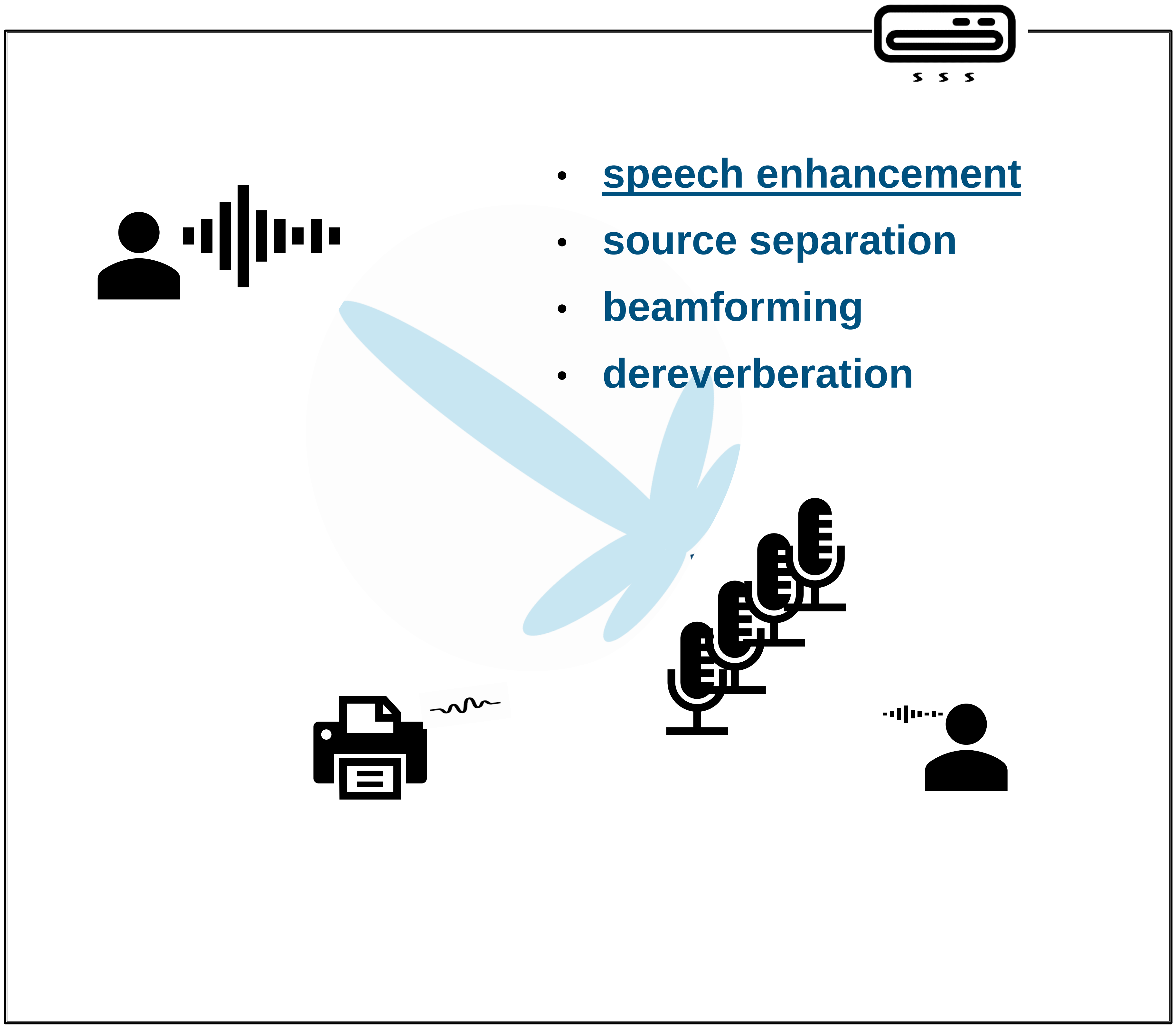
babblelabs

Acoustic Impairments Model



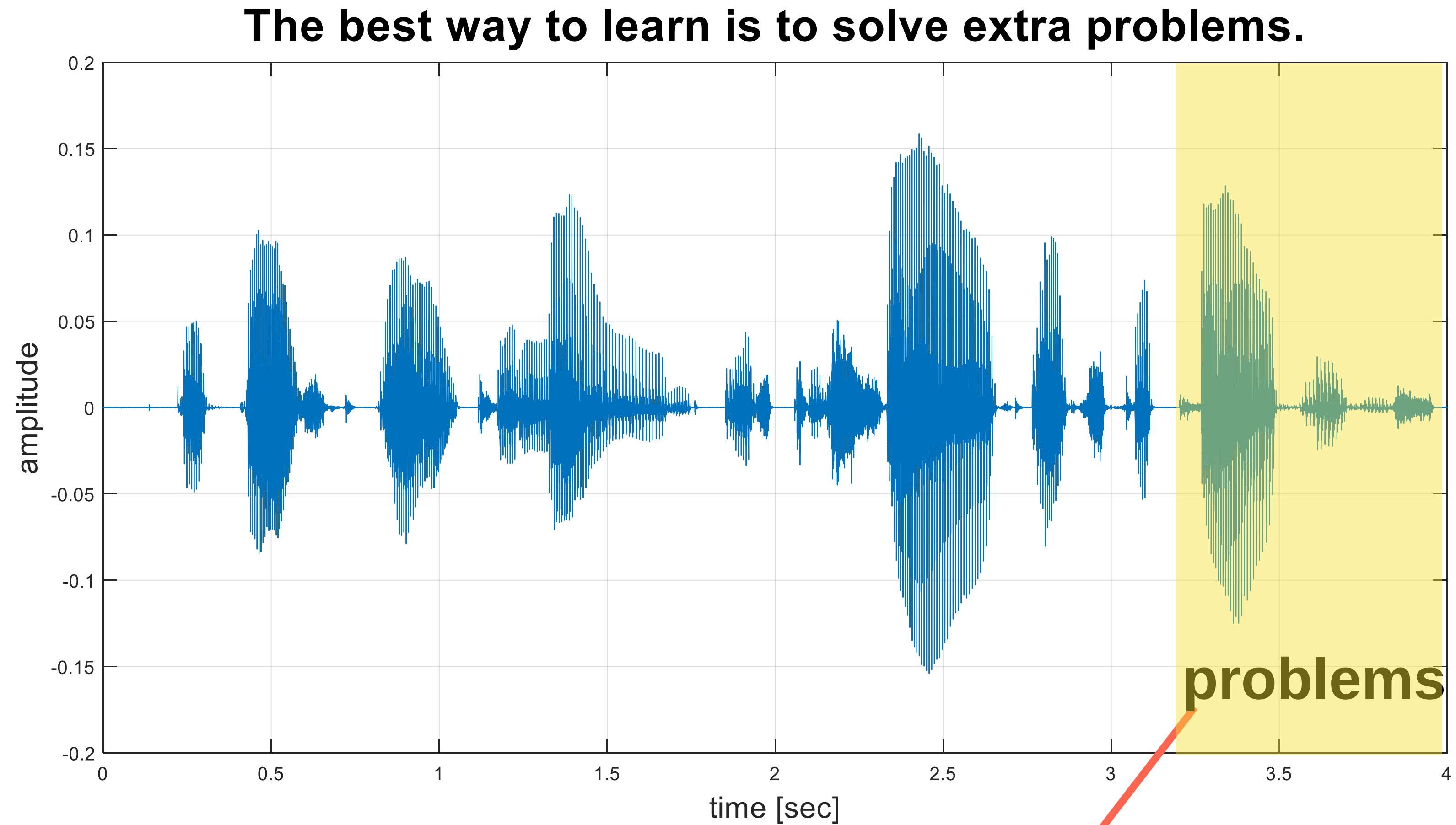
Solutions to Acoustic Impairments

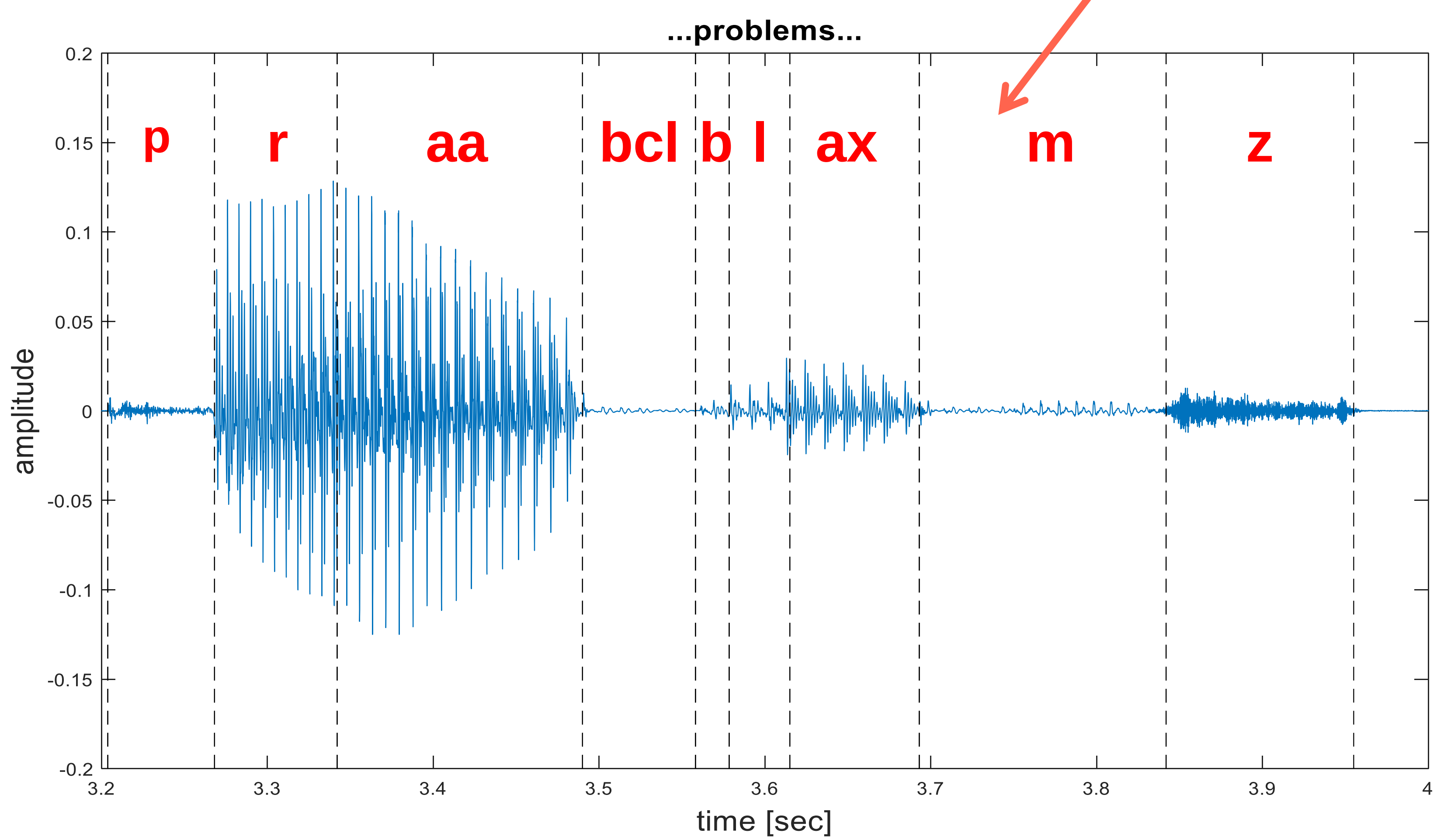
- speech enhancement
- source separation
- beamforming
- dereverberation



babblelabs

Classes of Speech Sounds





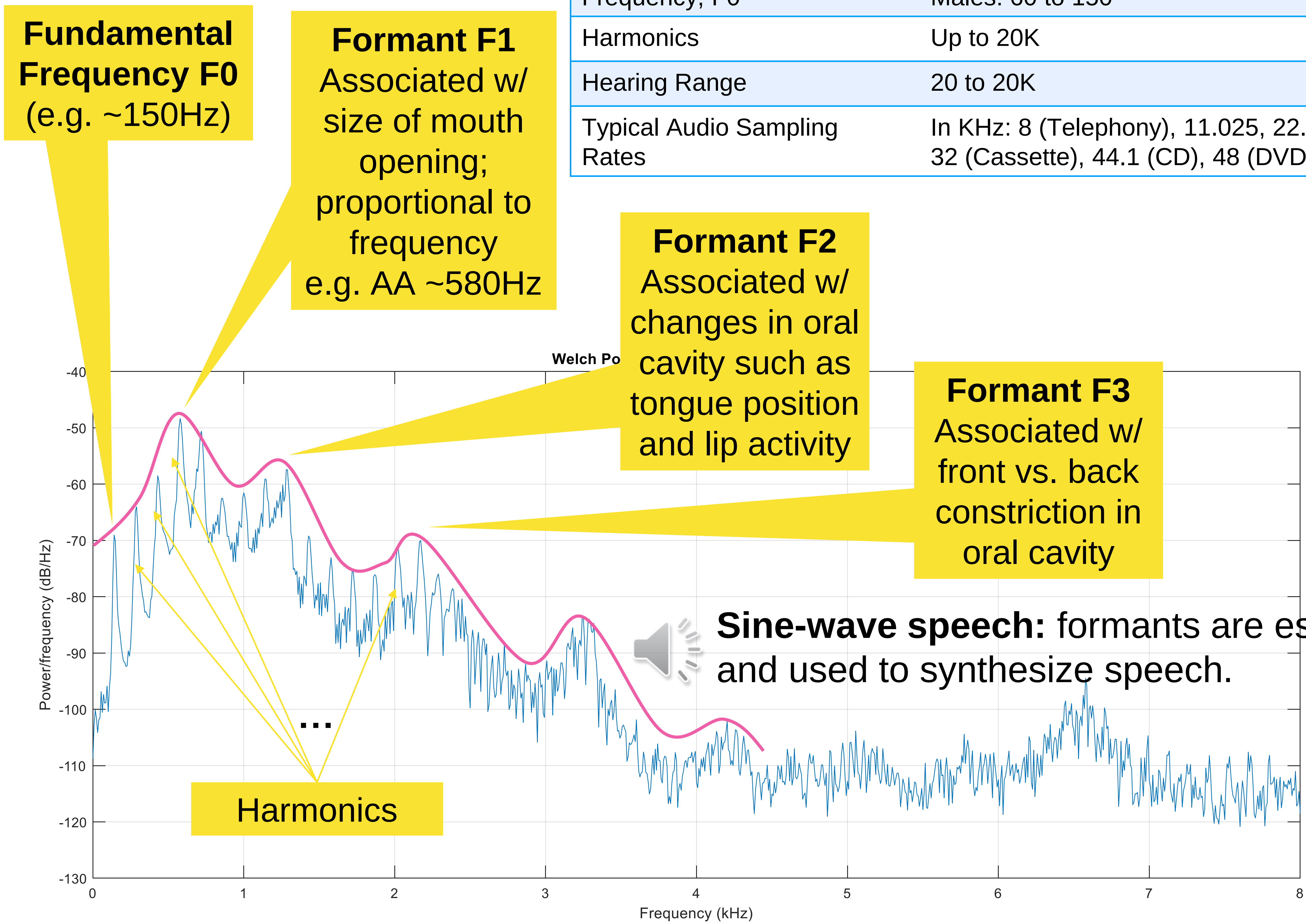
ARPABET Phonetic Symbols

Phone	Phoneme
Vowel	iy, ih, eh, ey, ae, **aa**, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, **ax**, ix, axr, axh
Semivowel	**l**, **r**, w, y, hh, hv, el
Affricate	jh, ch
Stops	**b**, d, g, **p**, t, k, dx, q
Nasal	**m**, n, ng, em, en, eng, nx
Fricative	s, sh, **z**, zh, f, th, v, dh

babblelabs

Speech Spectrum

Spectral Characteristics	Frequency Range [Hz]
Fundamental Frequency, F0	Females/Children: 200 to 400 Males: 60 to 150
Harmonics	Up to 20K
Hearing Range	20 to 20K
Typical Audio Sampling Rates	In KHz: 8 (Telephony), 11.025, 22.05 (MP3s), 32 (Cassette), 44.1 (CD), 48 (DVD)



Sine-wave speech: formants are estimated and used to synthesize speech.

Examples generated using Dan Ellis SW from: http://www.ee.columbia.edu/ln/rosa/matlab/sws/

babblelabs

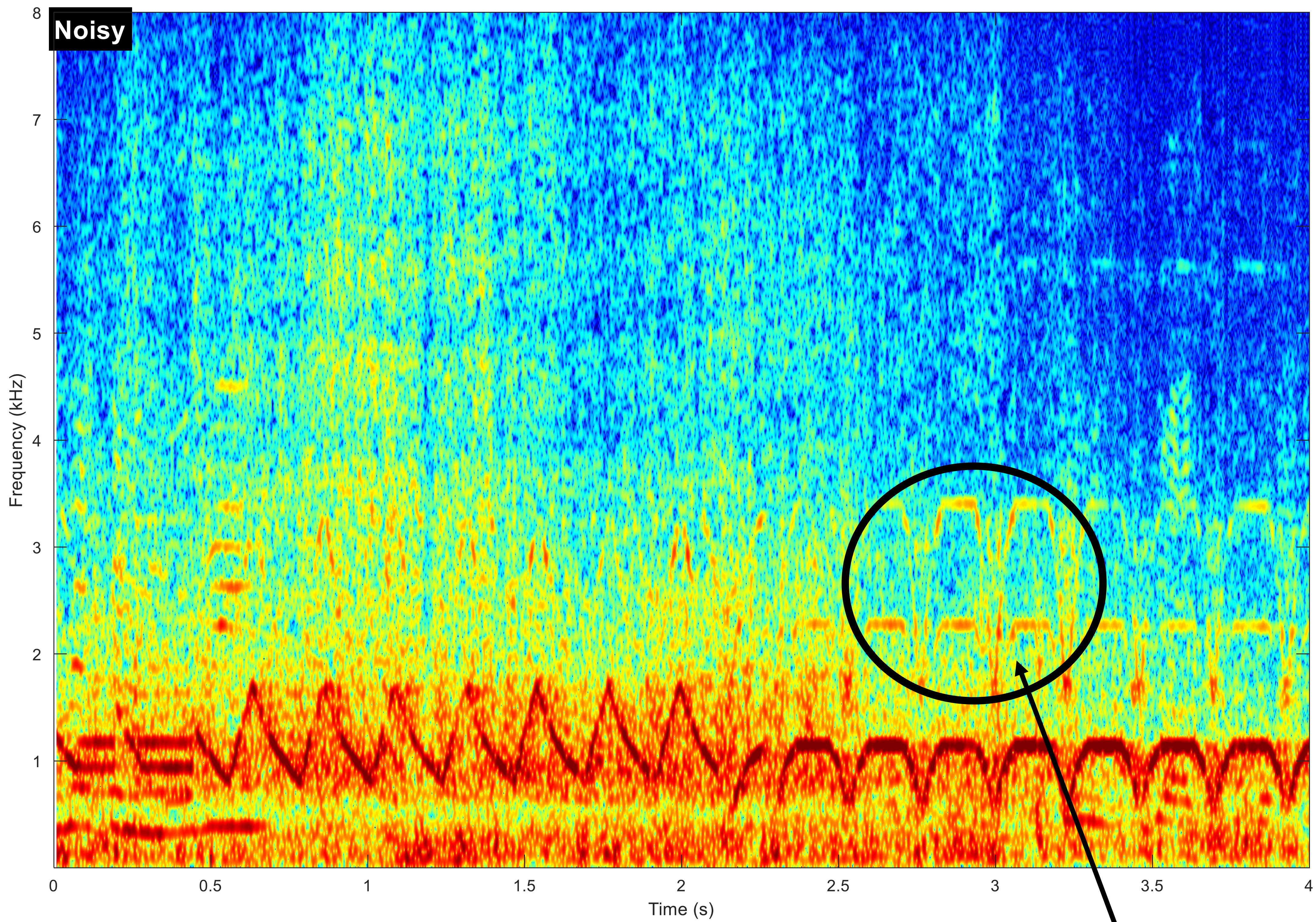
Noise and Speech Levels

Level [dB]	Classroom, Hospital Home, Store	Trains, Airplanes	Restaurants
Speech SPL	60 to 70	60 to 70	60 to 70
Noise SPL	50 to 55	70 to 75	59 to 80
SNR	+5 to +20	-15 to 0	-20 to +11

SPL: Sound Pressure Level relative to threshold of human hearing
(20 micro-Pascals (force per square meter) ~mosquito flying 3m away)

Typical target range for speech enhancement: -5 to 15dB

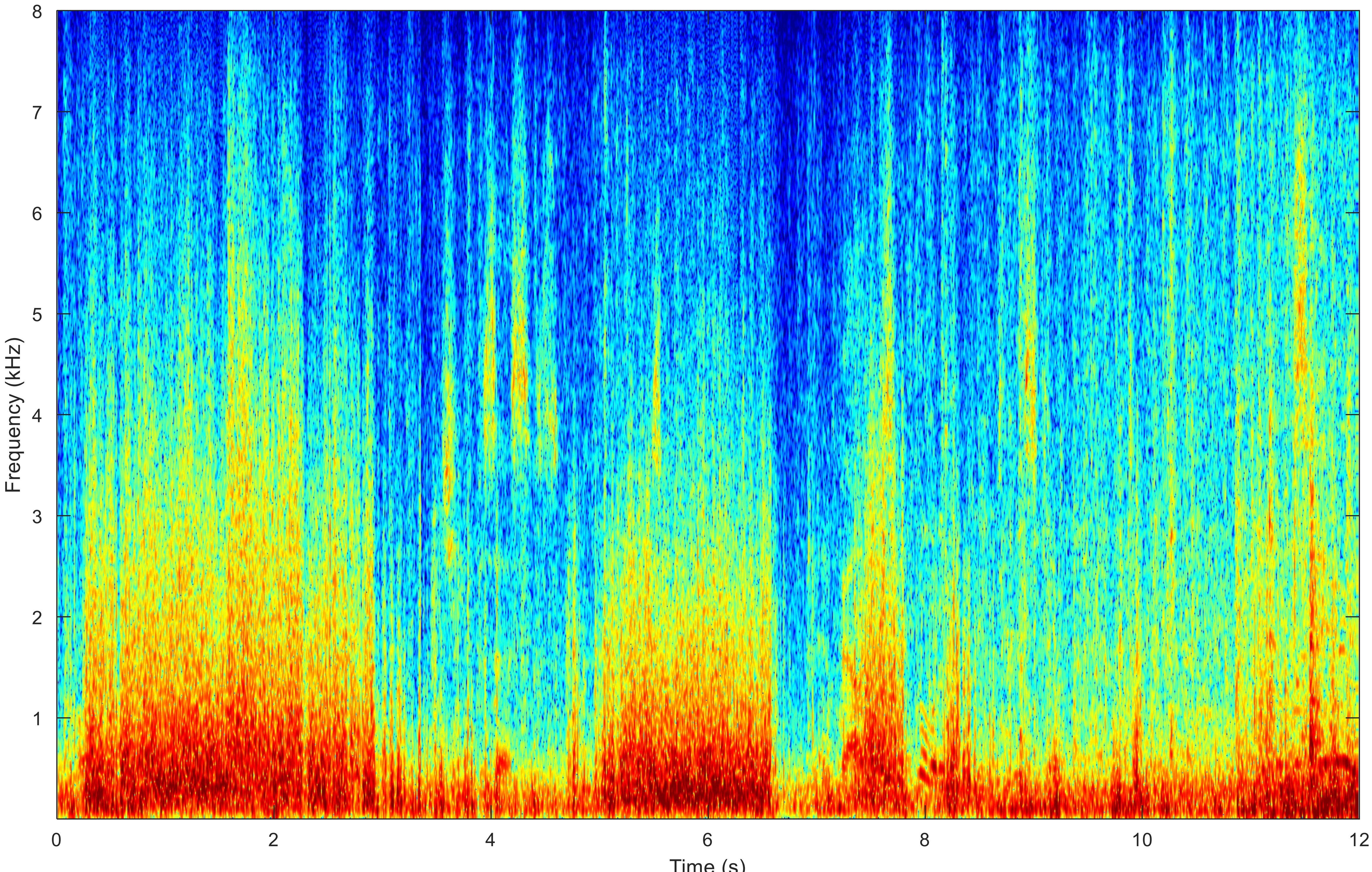
Sirens



Strong, structured frequency modulated tones & overtones

babblelabs

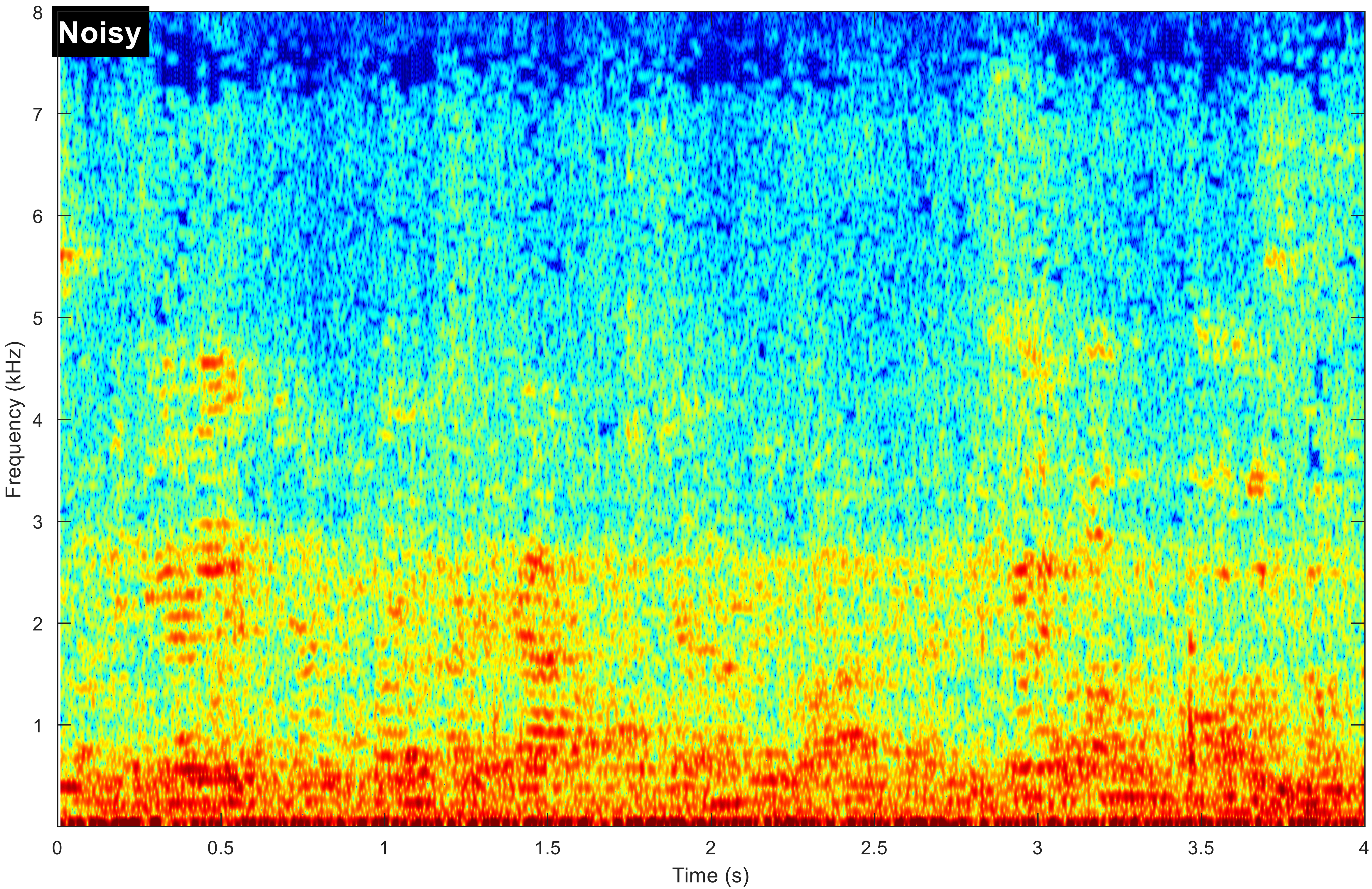
Wind Noise



Strong low frequency bursts + stationary broad spectrum

babblelabs

Crowd



Broad, non-stationary spectrum in speech range

babblelabs

Evaluating Performance of Speech Enhancers

- **Quality** measures assess **how** a speaker produces an utterance.
    - Is the utterance “natural”, “raspy”, “hoarse”, “scratchy”?
    - Does is sound good or bad?
- **Intelligibility** measures **what** a speaker said.
    - What did you understand?
    - What is the word error rate?

babblelabs

Subjective Measures of Quality

ITU-T P.835 Standard for Speech Enhancement Quality Assessment

Rating	Signal Distortion (SIG)	Background Distortion (BAK)	Overall Quality (OVL) Based on Mean Opinion Score Rating Scale (MOS)
5	Very natural, no degradation	Not noticeable	Excellent: Imperceptible
4	Fairly natural, little degradation	Somewhat noticeable	Good: Just perceptible, but not annoying
3	Somewhat natural, somewhat degraded	Noticeable but not intrusive	Fair: Perceptible and slightly annoying
2	Fairly unnatural, fairly degraded	Fairly conspicuous, somewhat intrusive	Poor: Annoying, but not objectionable
1	Very unnatural, very degraded	Very conspicuous, very intrusive	Bad: Very annoying and objectionable

babblelabs

Objective Measures of Quality and Intelligibility

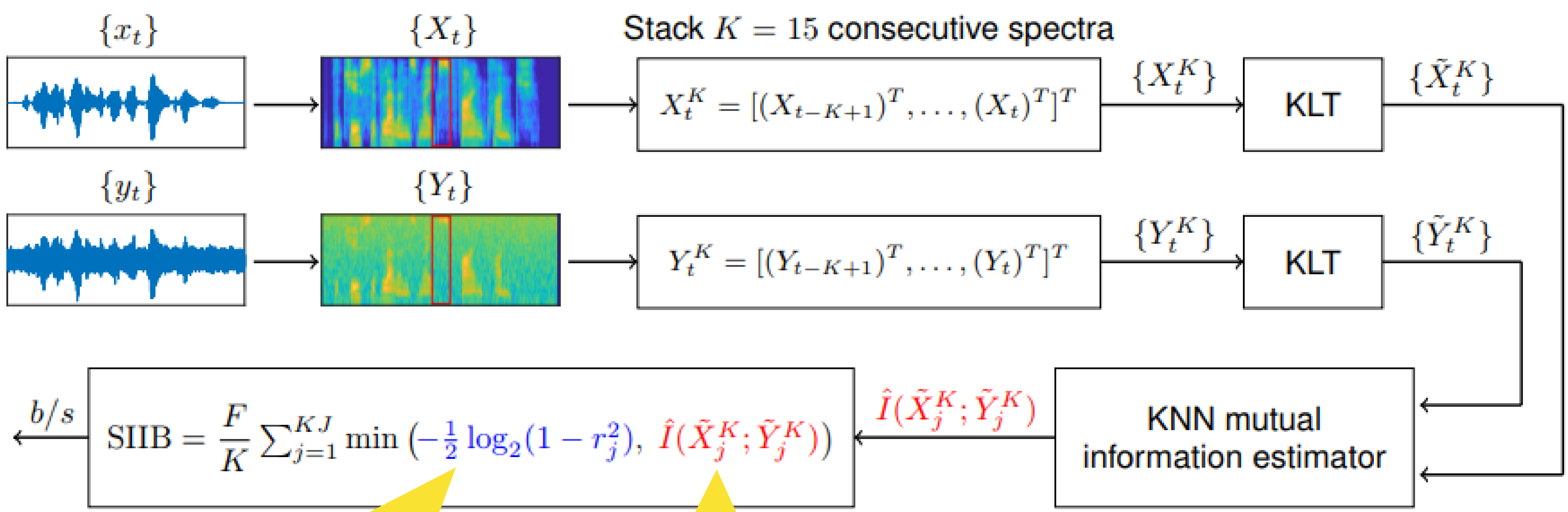
Quality	**Intelligibility**
Segmental SNR (SNRseg) Frequecy Weighted Segmental SNR (fwSNRseg) Weighted Spectral Slope (WSS) Log-likelihood Ratio (LLR) Itakura-Saito (IS) Cepstral Distance (CEP) Hearing Aid Speech Quality Index (HASQI) Perceptual Evaluation of Video Quality (PEVQ) Perceptual Evaluation of Audio Quality (PEAQ) Perceptual Evaluation of Speech Quality (PESQ) **Perceptual Objective Listening Quality Analysis (POLQA)** Composite Metrics	Normalized Covariance Metrics (NCM) Speech Intelligibility Index (SII) High-energy Glimpse Proportion Metric Coherence and Speech Intelligibility Index (CSII) Quasi-stationary Speech Transmission Index (QSTI) Short-time Objective Intelligibility Measure (STOI) Extended STOI Measure (ESTOI) Hearing-Aid Speech Perception Index (HASPI) K-Nearest Neighbor Mutual Information Intelligibility Measure (MIKNN) Speech Intelligibility Prediction based on a Mutual Information Lower Bound (SIMI) **Speech Intelligibility in Bits (SIIB)** Speech-based Envelop Power Spectrum Model with Short-Time correlation (sEPSM) Automatic Speech Recognition (ASR)

Effectiveness of metrics is evaluated by measuring correlation of metric predictions against subjective test data

babblelabs

Speech Intelligibility in Bits (SIIB)

- Measures amount of information between speaker and listener.
- Linguistic models for “clean” speech communication measure 50-100 bps typical information rate.



Mutual Info between “text” message and clean speech

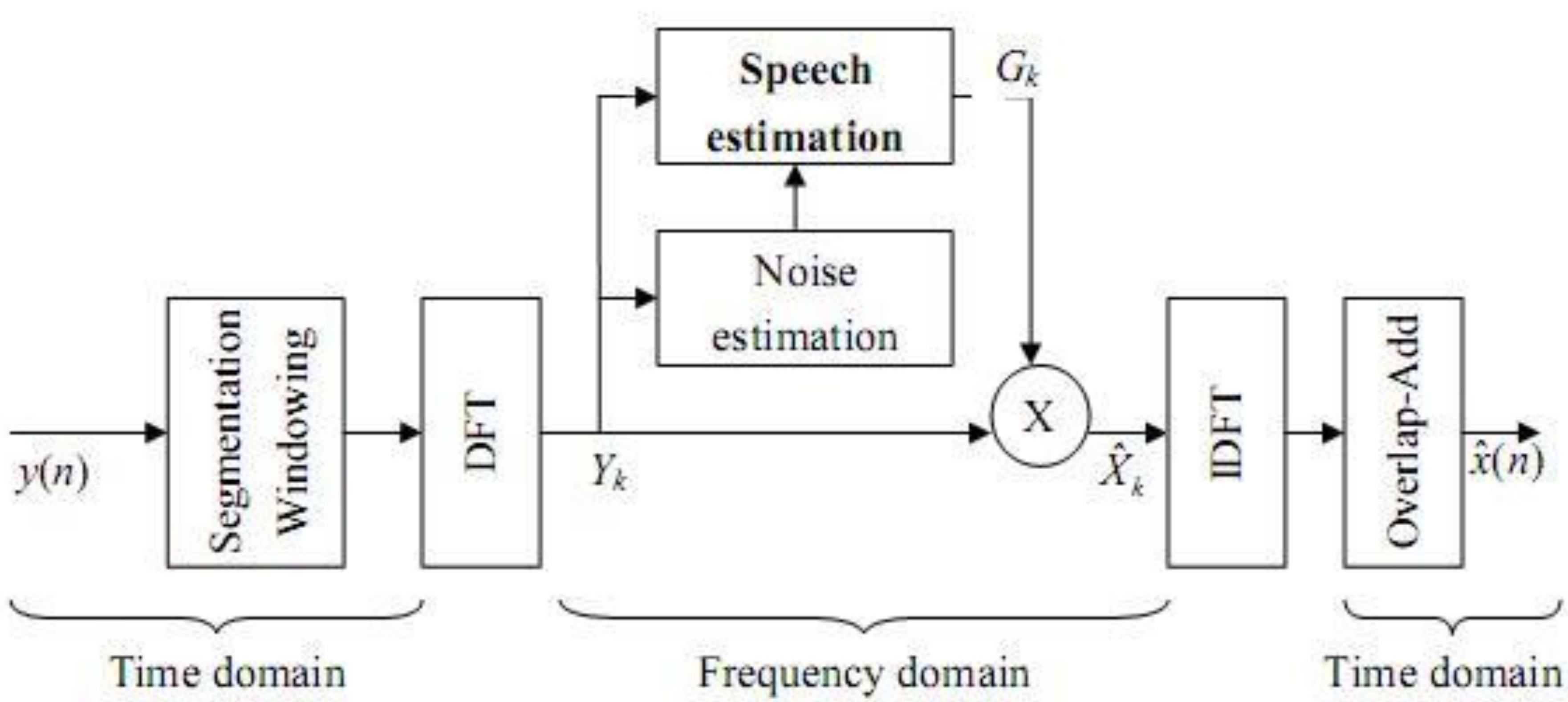
Mutual Info between clean and noise speech

From: S. Van Kuyk; W. B. Kleijn; R. C. Hendriks; “An instrumental intelligibility metric based on information theory,” in IEEE Signal Processing Letters, 2018

babblelabs

Traditional Methods of Speech Enhancement

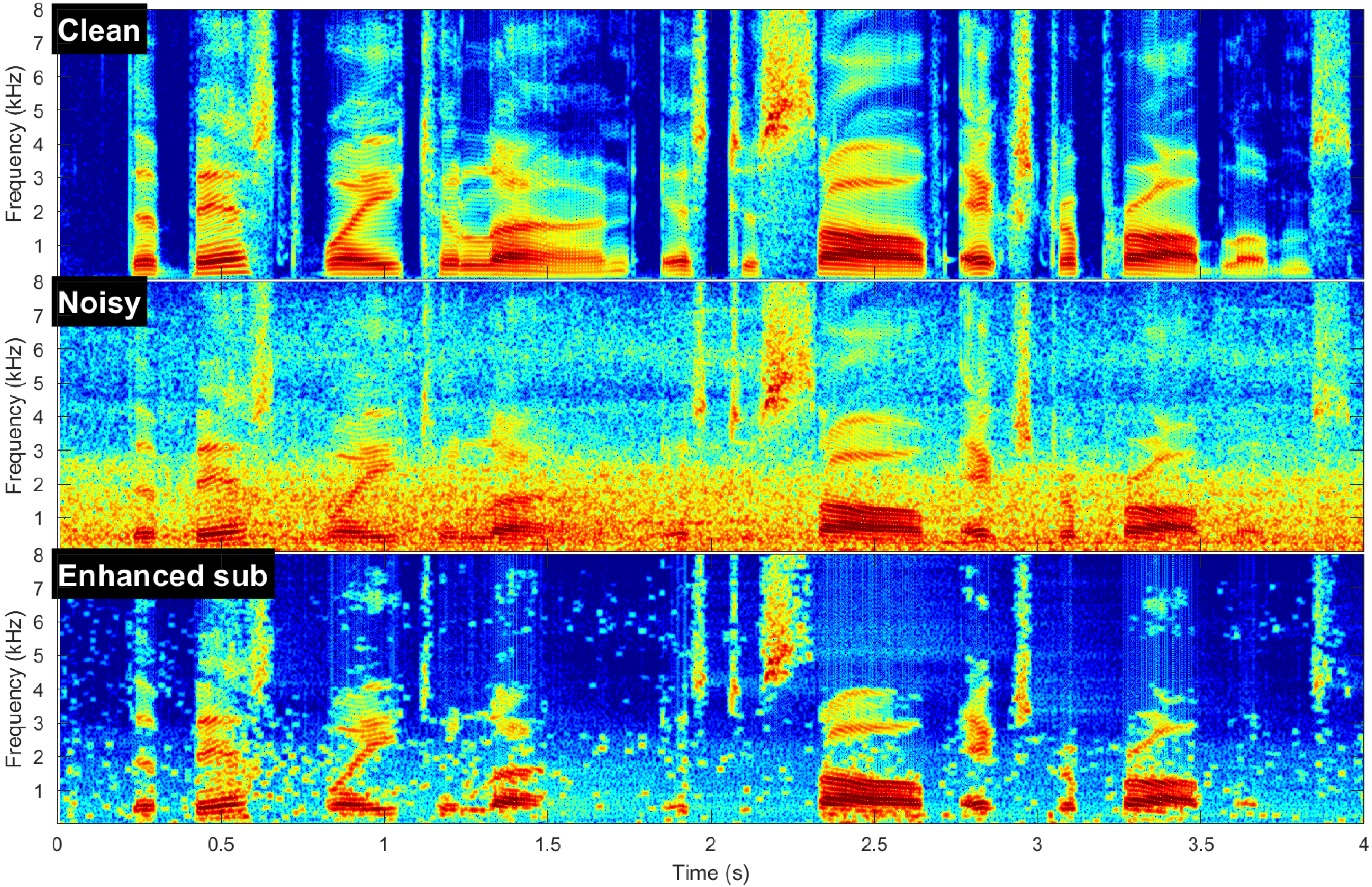
- Most commonly employ a short-time Fourier transform based analysis-modification-synthesis framework
- Frequency dependent noise suppression function
- Noises suppression based on estimates of speech and noise statistics



Traditional Methods: Spectral Subtraction

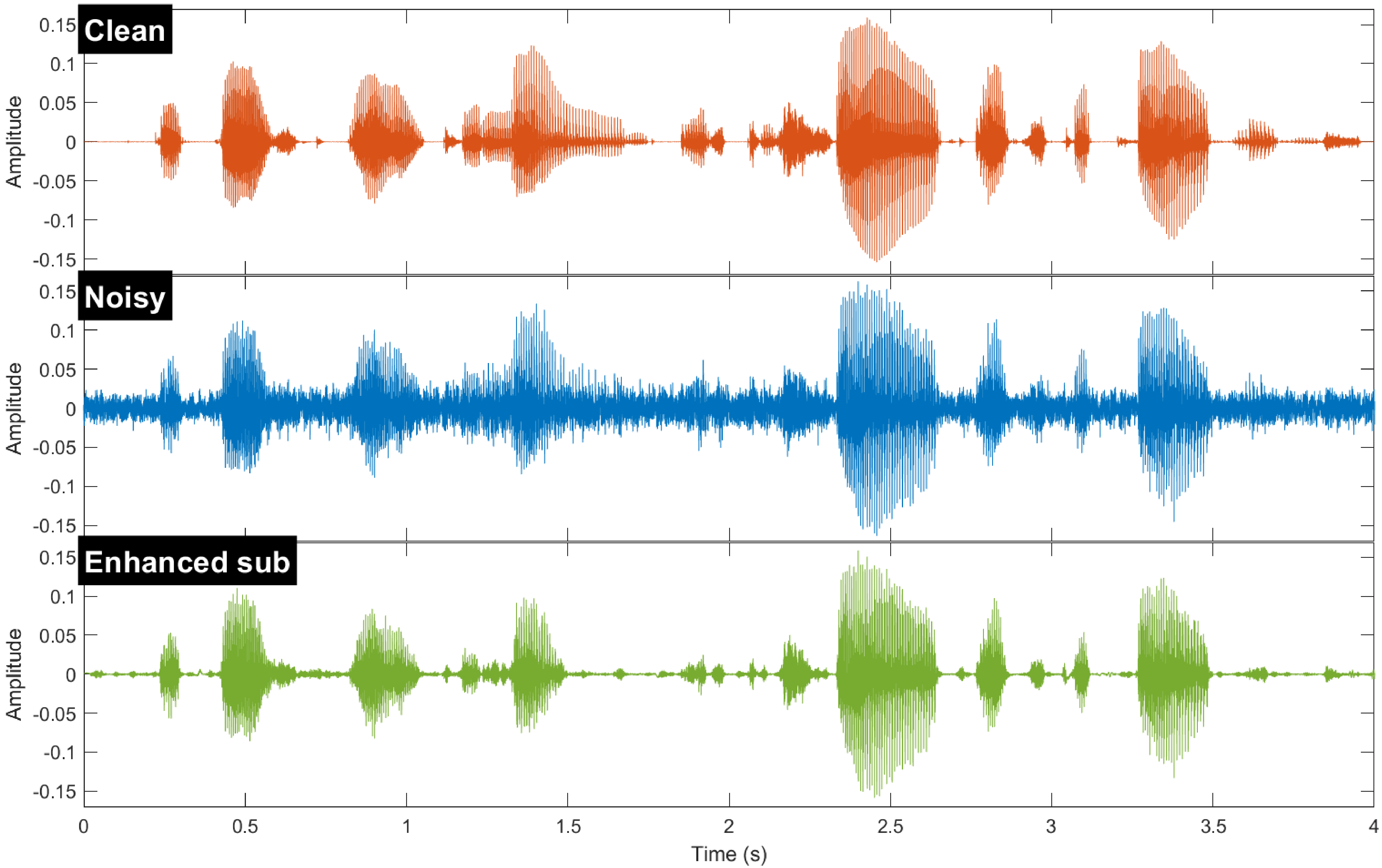
$\underbrace{R(\omega)}_{\substack{\text{noisy}\\\text{speech}}} = \underbrace{S(\omega)}_{\substack{\text{clean}\\\text{speech}}} + \underbrace{D(\omega)}_{\text{noise}}$	Noisy speech model
$\lvert\widehat{D}\rvert^2 = \mathrm{E}\{\lvert R\rvert^2\} = \mathrm{E}\left\{\lvert\widehat{D}\rvert^2\right\}$ $when\ S = 0$	**Noise magnitude estimate** measured during period of speech inactivity using Voice Activity Detector
$\lvert R\rvert^2 = \lvert S\rvert^2 + \lvert D\rvert^2 + \underbrace{2\mathrm{Re}\{SD^*\}}_{\substack{\text{ignore}\\\text{this term!!}}}$	Noisy speech magnitude Cross term is ignored because clean speech and noise are uncorrelated
$\lvert\hat{S}\rvert^2 = \lvert R\rvert^2 - \lvert\widehat{D}\rvert^2$	**Clean speech magnitude estimate**
$\hat{S}(\omega) = \underbrace{\lvert\hat{S}(\omega)\rvert}_{\substack{\text{clean}\\\text{magnitude}\\\text{estimate}}} \exp\left\{j\underbrace{\Phi_r(\omega)}_{\substack{\text{noisy}\\\text{phase}}}\right\}$	**Clean speech synthesized** from noisy phase and magnitude estimate Difference in noisy and clean phase not perceptible for SNRs > 8dB

Spectral Subtraction: Spectrograms



babblelabs

Spectral Subtraction: Waveforms



babblelabs

Deep Neural Networks for Speech Enhancement

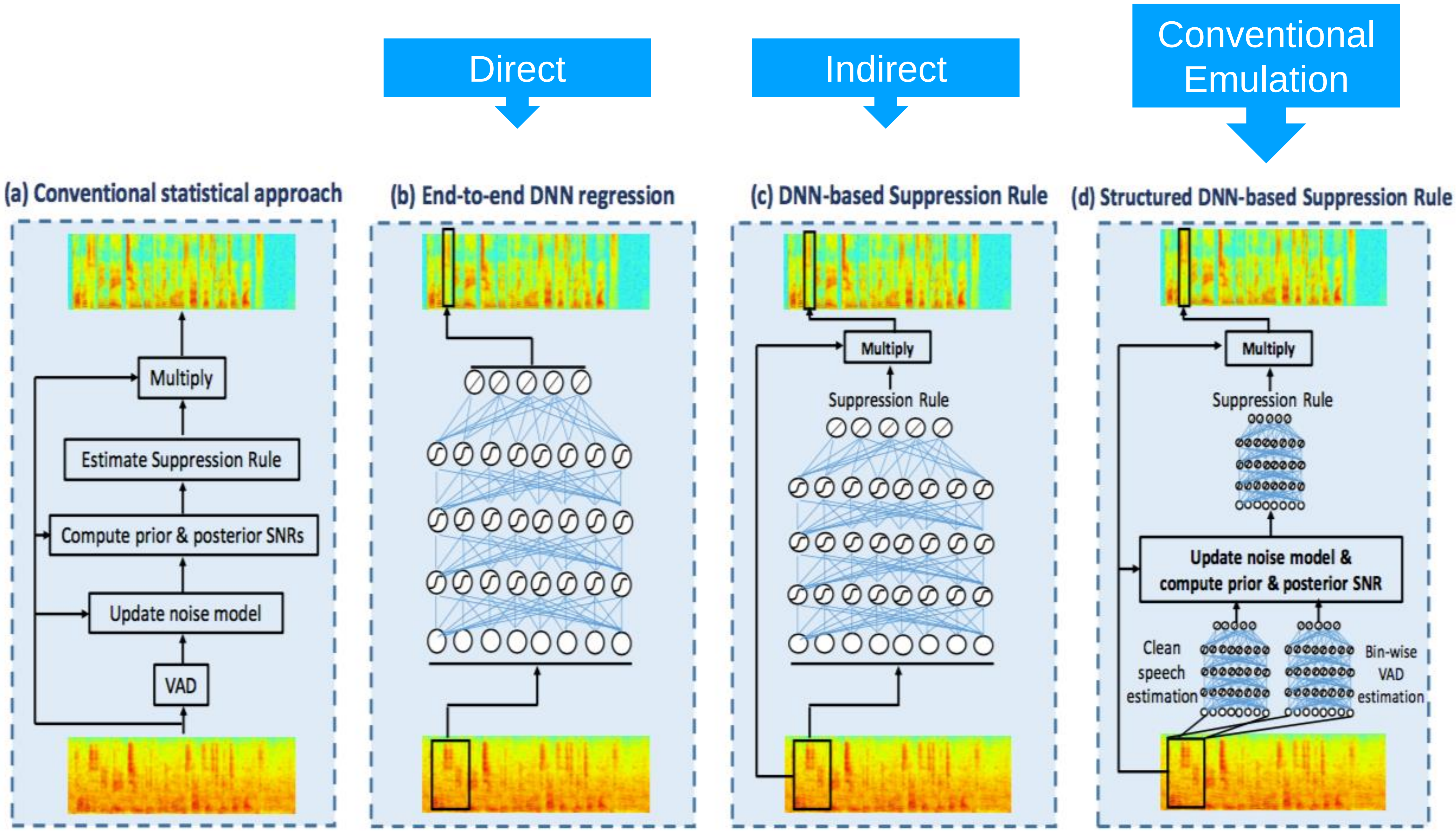


Figure 1: Different architectures for DNN-based speech enhancement: (a) Statistical noise suppression. (b) Enhancement by end-to-end DNN regression from noisy spectral features to clean features. (c) Estimating binwise suppression gain directly by a DNN. (d) Employing separate DNNs replacing the different components of conventional suppression gain estimation.

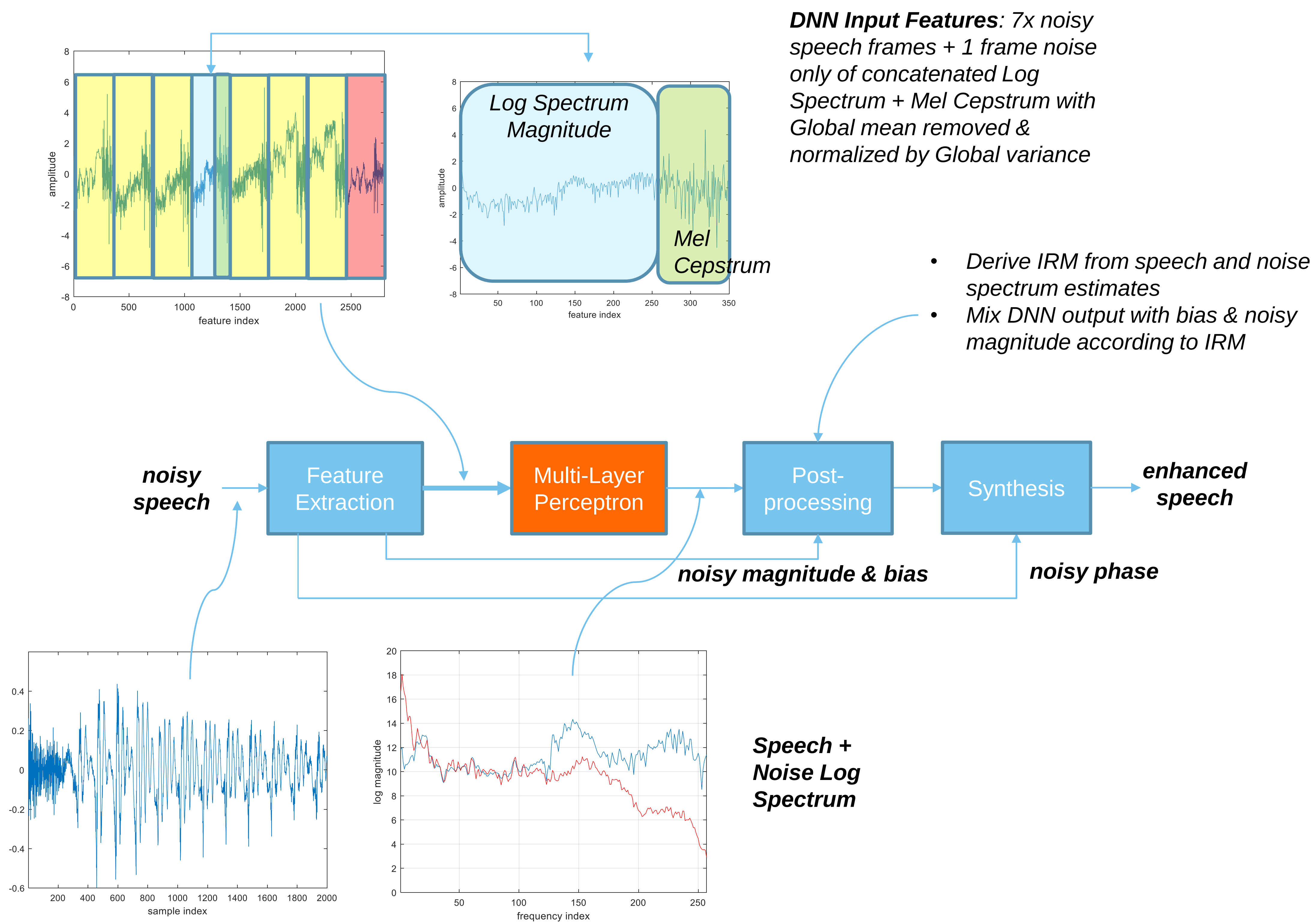
Mirsamadi, Seyedmahdad, and Ivan Tashev. "Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation." INTERSPEECH. 2016.

Common Ideal Target Masks

target mask/filter	formula	optimality principle
IBM:	$a^{\mathrm{ibm}} = \delta(\lvert s\rvert > \lvert n\rvert),$	max SNR $a \in \{0, 1\}$
IRM:	$a^{\mathrm{irm}} = \dfrac{\lvert s\rvert}{\lvert s\rvert + \lvert n\rvert},$	max SNR $\theta_{\mathrm{s}} = \theta_{\mathrm{n}},$
"Wiener like":	$a^{\mathrm{wf}} = \dfrac{\lvert s\rvert^2}{\lvert s\rvert^2 + \lvert n\rvert^2},$	max SNR, expected power
ideal amplitude:	$a^{\mathrm{iaf}} = \lvert s\rvert / \lvert y\rvert,$	exact $\lvert\hat{s}\rvert$, max SNR $\theta_{\mathrm{s}} = \theta_{\mathrm{y}}$
phase-sensitive filter:	$a^{\mathrm{psf}} = \dfrac{\lvert s\rvert}{\lvert y\rvert}\cos(\theta),$	max SNR given $a \in \mathbb{R}$
ideal complex filter:	$a^{\mathrm{icf}} = s/y,$	max SNR given $a \in \mathbb{C}$

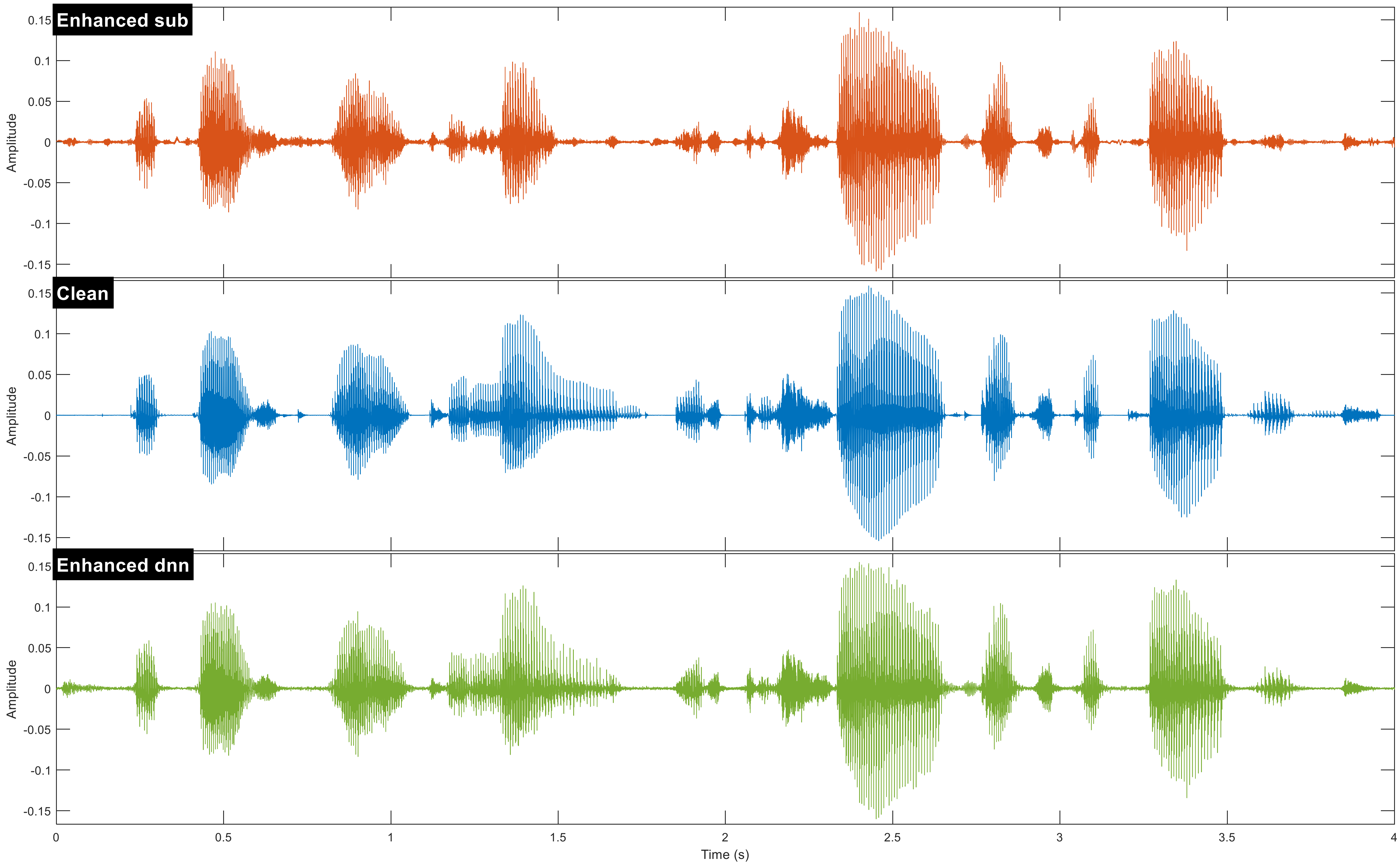
H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in Proc. Int. Conf. Acoust., Speech, Signal Process., 2015, pp. 708–712

GeorgiaTech System

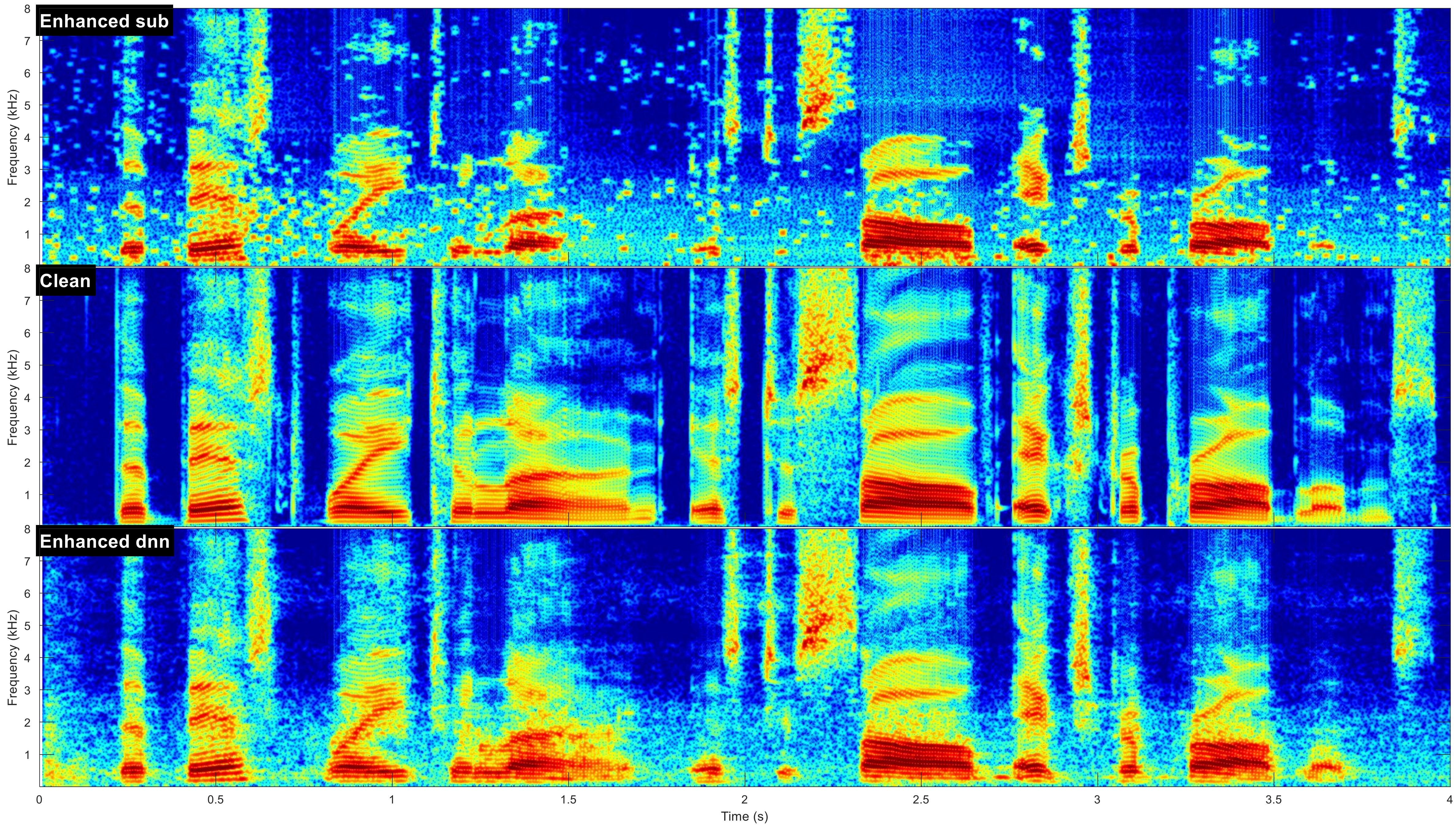


Xu, Yong, et al. "A regression approach to speech enhancement based on deep neural networks." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23.1 (2015): 7-19.

Spectral Subtractive vs. BabbleLabs DNN

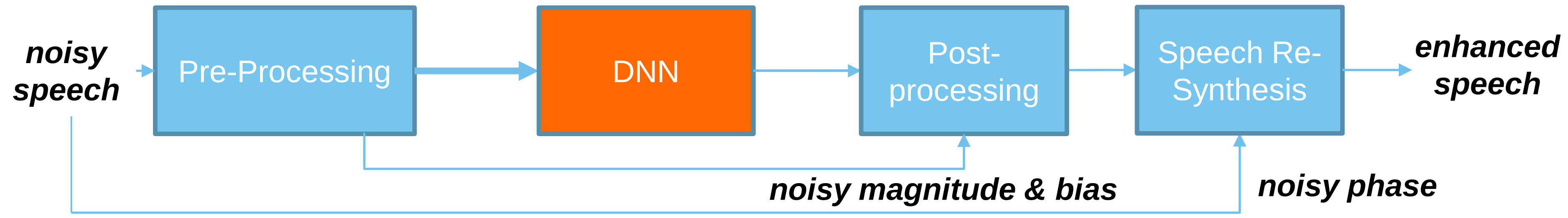


Spectral Subtractive vs. BabbleLabs DNN



Metric	Noisy	Subtractive	BabbleLabs DNN
SNR [dB]	5.98	10.53	13.57
PESQ	1.18	1.42	2.01
ESTOI	0.44	0.49	0.71
SIIB_Gauss [bps]	65	56	93

BabbleLabs Production Flow



- 90% of the code in the blue boxes
- 90% of the compute in the orange box
- Prototyping is in blocking format, while deployment is in streaming format.
- Using Matlab and the GPU coder, we were able to covert from reference to deployment code in 6 man-weeks.
- Currently we are porting the DNN using other open source tools.
  - Exploring the migration to GPU coder to unify the flow if possible.

babblelabs

References

- Loizou, Philipos C. Speech enhancement: theory and practice. CRC press, 2007.
- Van Kuyk, Steven, W. Bastiaan Kleijn, and Richard C. Hendriks. "An evaluation of intrusive instrumental intelligibility metrics." arXiv preprint arXiv:1708.06027 (2017).
- Mirsamadi, Seyedmahdad, and Ivan Tashev. "Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation." INTERSPEECH. 2016.
- Xu, Yong, et al. "A regression approach to speech enhancement based on deep neural networks." *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 23.1 (2015): 7-19.
- H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in Proc. Int. Conf. Acoust., Speech, Signal Process., 2015, pp. 708–712
- *From http://www.vision.huji.ac.il/visual-speech-enhancement/*
- https://looking-to-listen.github.io/

babblelabs

s p e a k   y o u r   m i n d