

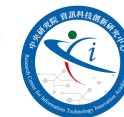
Bio-Informed Speech Processing, Modeling, and Generation

Yu Tsao (Ph.D., Professor)

Research Center for Information Technology Innovation

Academia Sinica

yu.tsao@citi.sinica.edu.tw



中央研究院
ACADEMIA SINICA

Yu Tsao, Ph.D.

Research Fellow (Professor)

- Work Experience

- Research Fellow (Professor)

Research Center for Information Technology Innovation, Academia Sinica, Taiwan

- President

The Association for Computational Linguistics and Chinese Language Processing, Taiwan

- Chair

IEEE CTSoc, Audio/Video Systems and Signal Processing Technical Committee

- Selected Honors & Services

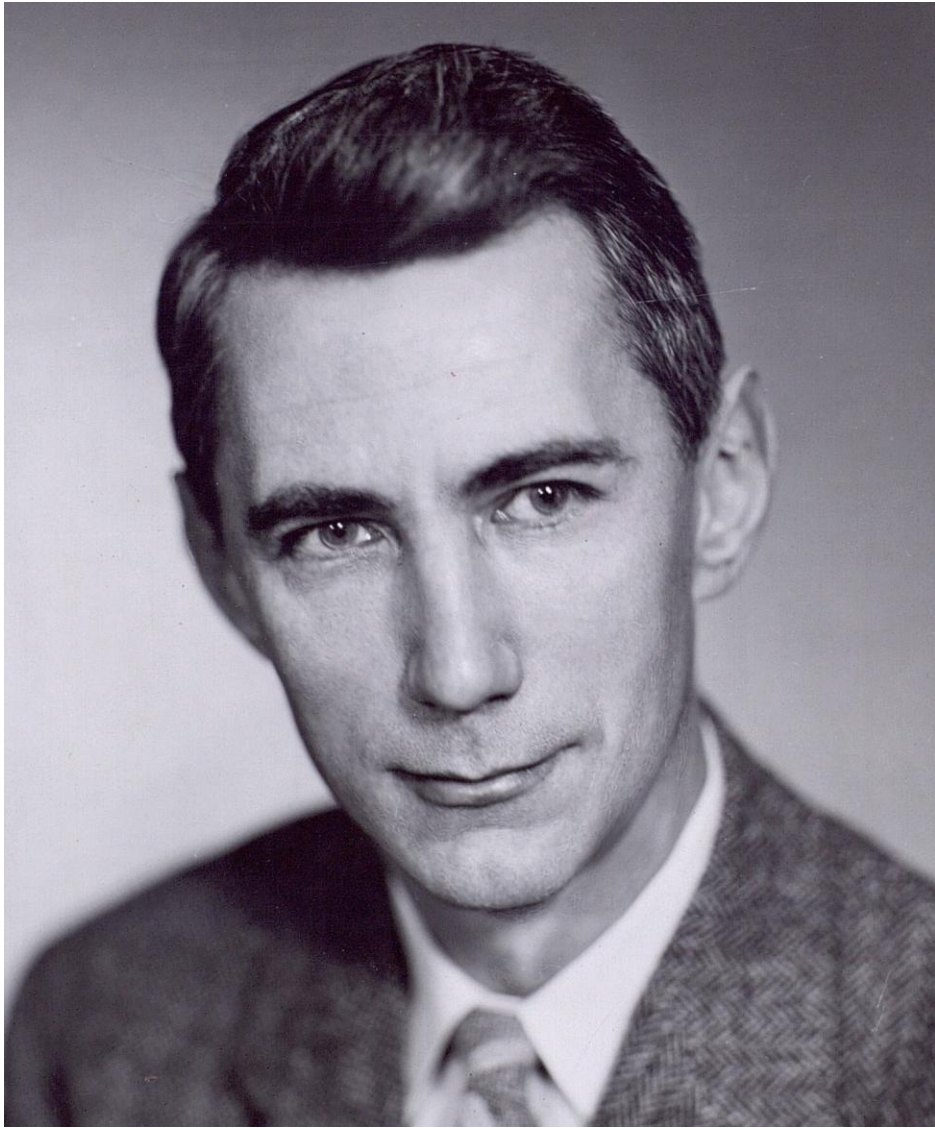
- 2025 IEEE Chester W. Sall Memorial Award
- Outstanding Research Award, National Science and Technology Council (NSTC)
- 2021 IEEE Signal Processing Society Young Author Best Paper Award (Corresponding Author)
- Associate Editor of IEEE Transactions on Consumer Electronics
- Senior Associate Editor of IEEE Signal Processing Letters

- Research Interests

- Neural speech enhancement and assessment
- Assistive oral communication technology
- Biomedical signal processing



<https://www.citi.sinica.edu.tw/pages/yu.tsao/>
<http://bio-asplab.citi.sinica.edu.tw/>

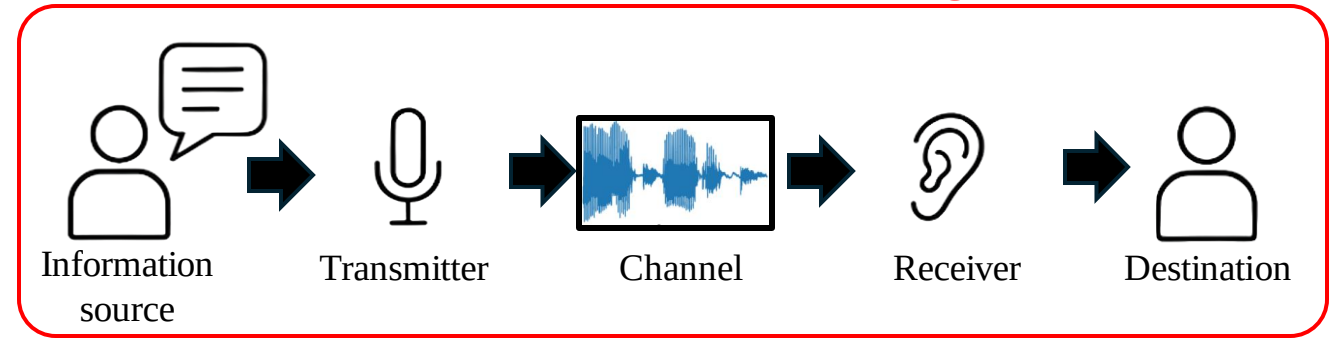


Claude E. Shannon

A Mathematical Theory of Communication (1948)

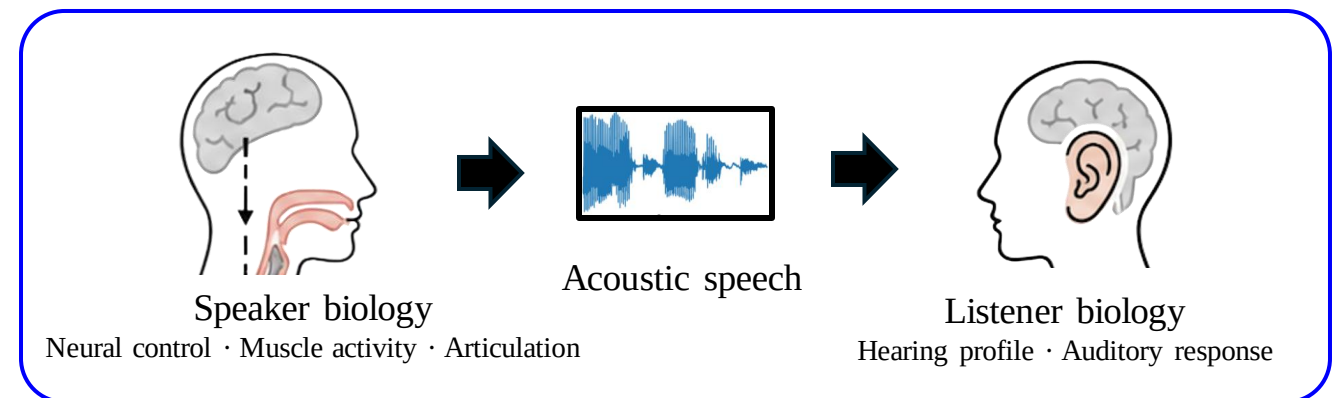
A communication system

Signal-level view



Speech communication

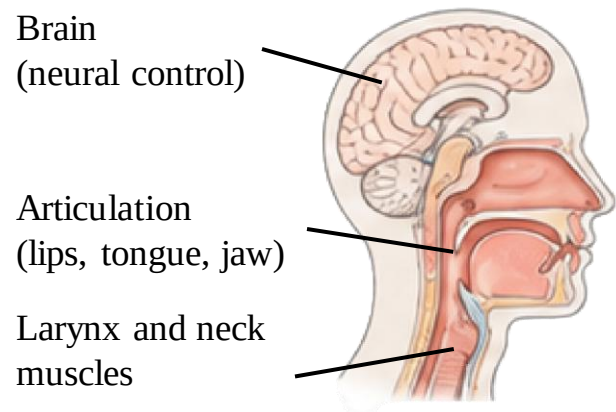
Biological processing



To better understand and to improve speech communication, we can consider biological information involved in both the **speaker** and the **listener** sides.

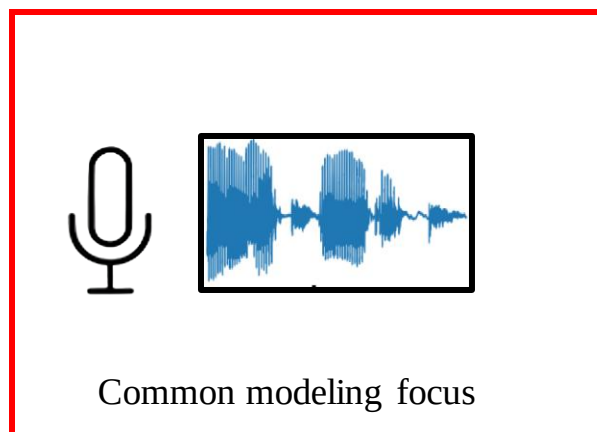
Beyond the Acoustic Signal

SPEECH PRODUCTION

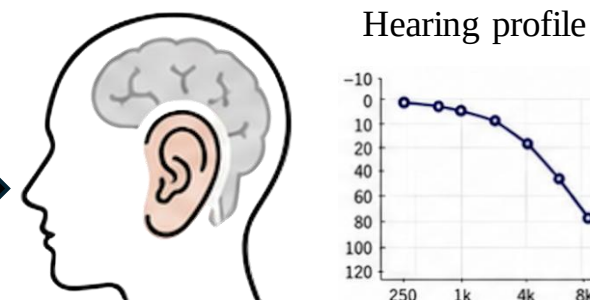


Neural control
Muscle activity
Articulatory movement

ACOUSTIC SIGNAL



SPEECH PERCEPTION



Hearing sensitivity
Auditory response
Perceptual outcome

Speaker-side Information

Visual cues · Bone conduction ·
Electrical activity

Listener-side Information

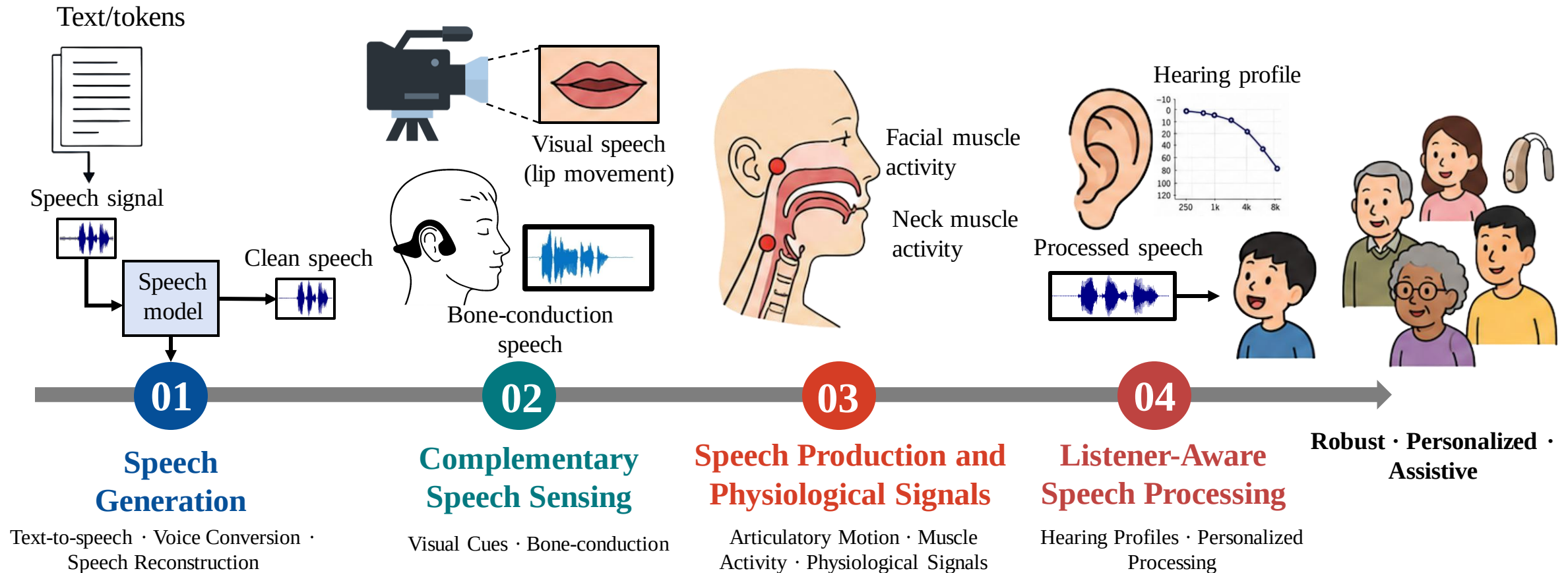
Hearing profile · Neural response

Bio-informed speech processing



Biological information **complements** the acoustic signal by describing how speech is **produced** and **perceived** and is useful for building **effective models**.

From Acoustic Modeling to Human-Centered Speech Intelligence



Speech technology moves from modeling **signals** toward modeling **human communication**.

OUTLINE

1. **Speech Generation (Acoustic Only)**

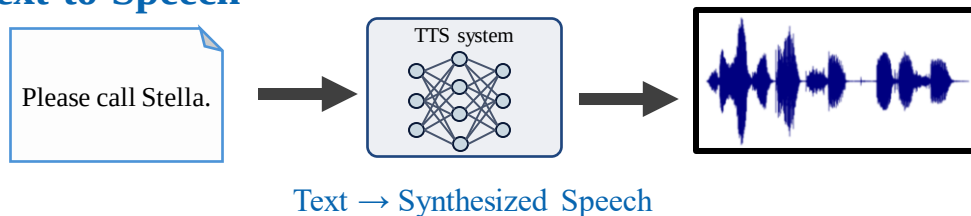
- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

Conventional Speech Generation and Reconstruction

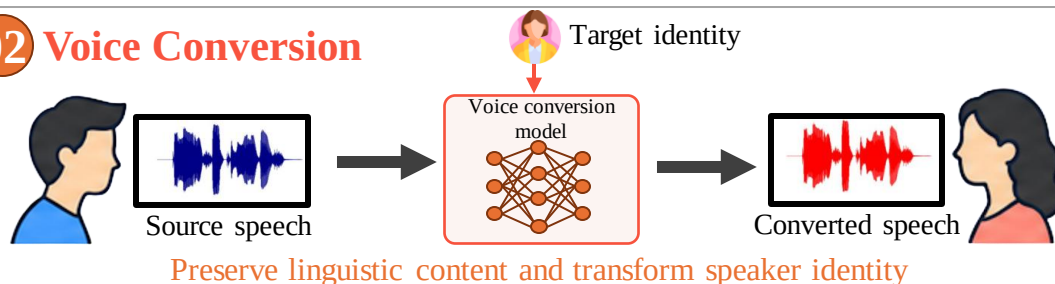
Speech can be synthesized, converted, or reconstructed from different inputs

Three common tasks

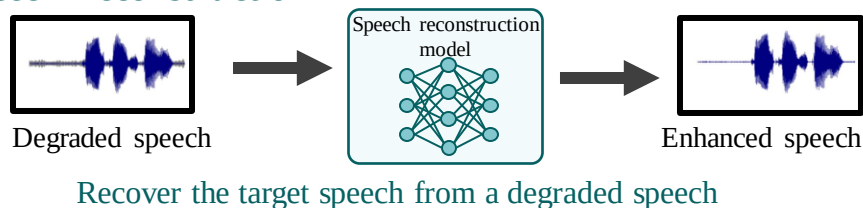
01 Text-to-Speech



02 Voice Conversion



03 Speech Reconstruction



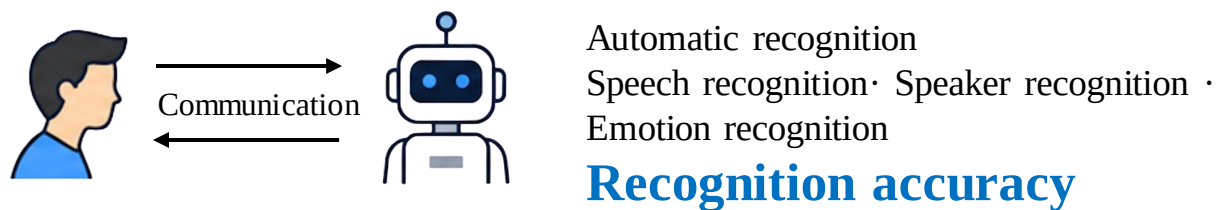
Who receives the output?

Human listener



or

Machine listener



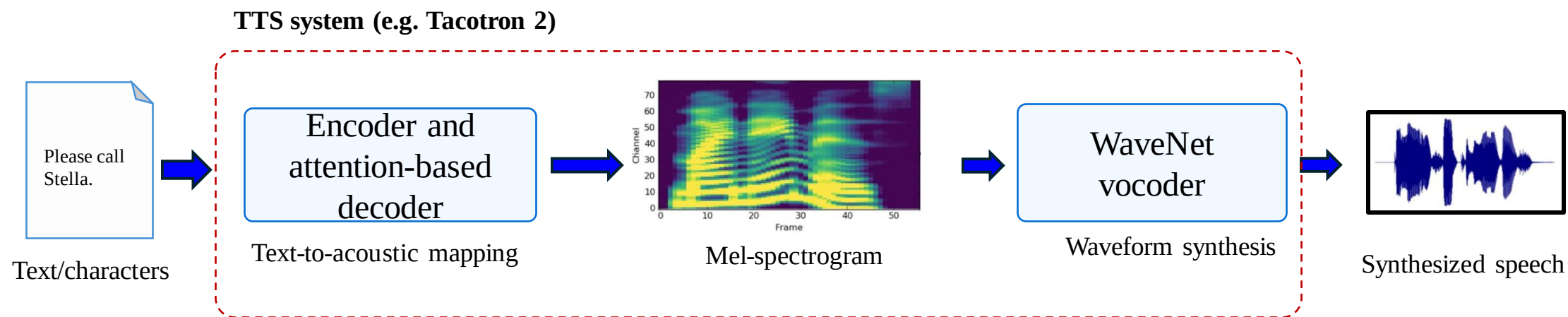
Evaluation depends on the intended receiver



Generated speech must remain useful to either **human perception** or **machine recognition**.

Text-to-Speech (TTS) Synthesis

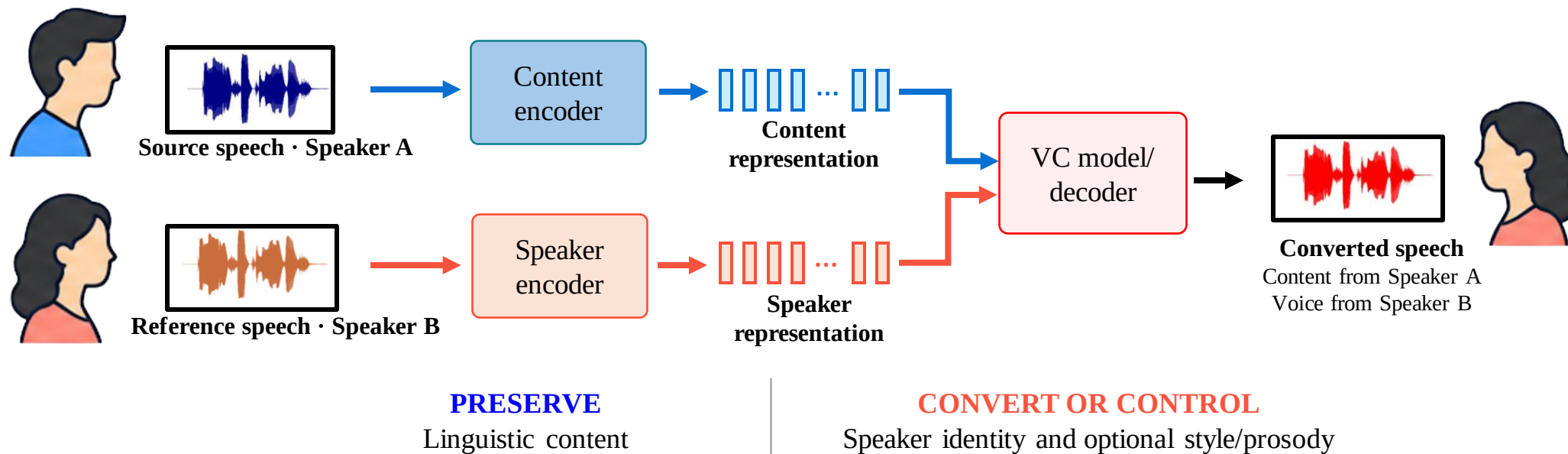
Convert text to speech [1], [2], [3]



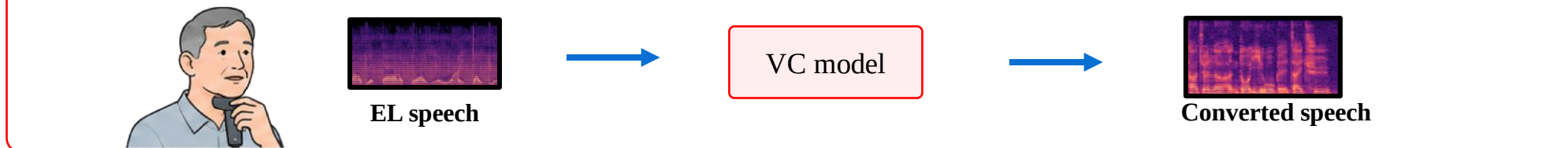
A TTS system predicts spectral features **from text** and uses vocoders **to** synthesize the **waveform**.

Voice Conversion (VC)

Convert source speech to a target voice [4], [5], [6], [7]



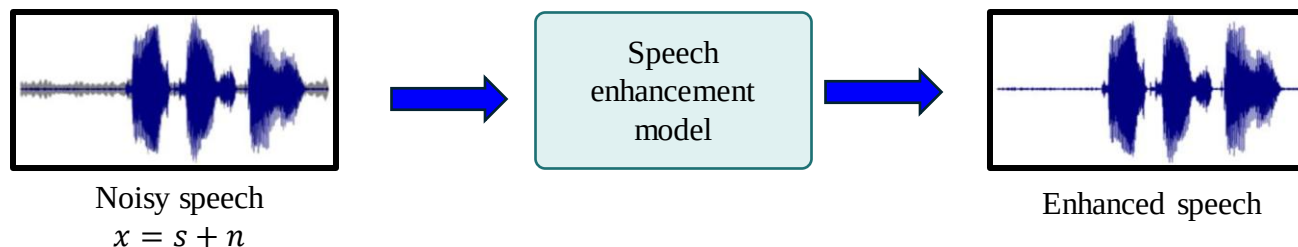
Assistive example: VC for electrolaryngeal speech



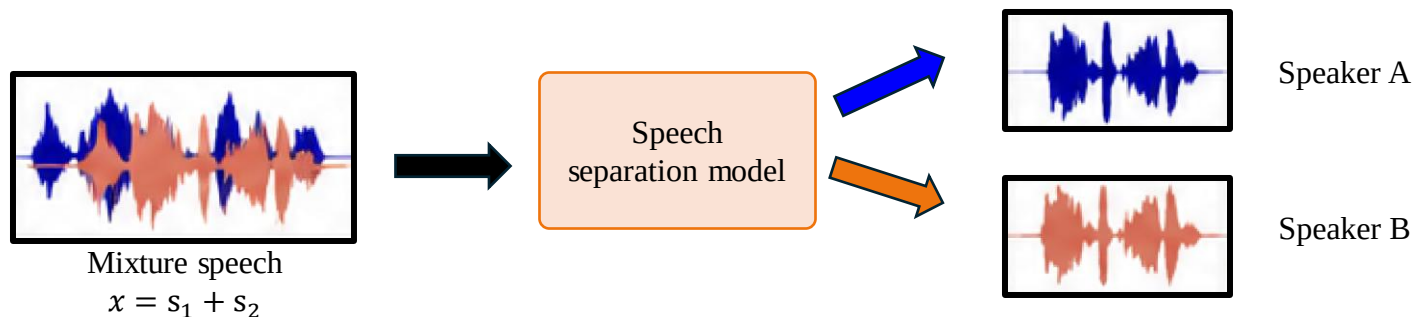
Speech Reconstruction

Recovering target speech from degraded or mixed observations [8-14]

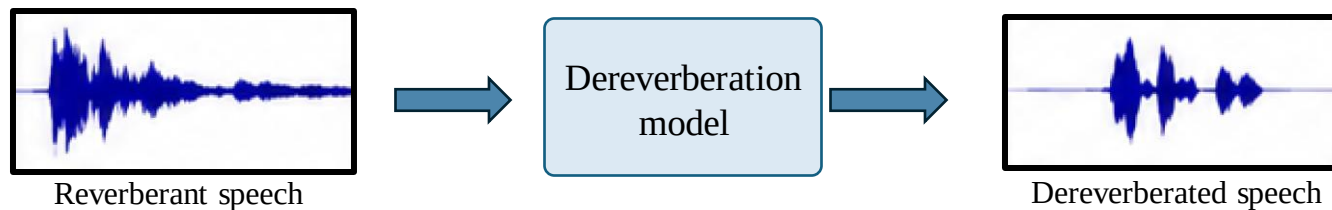
01 Speech Enhancement



02 Speech Separation



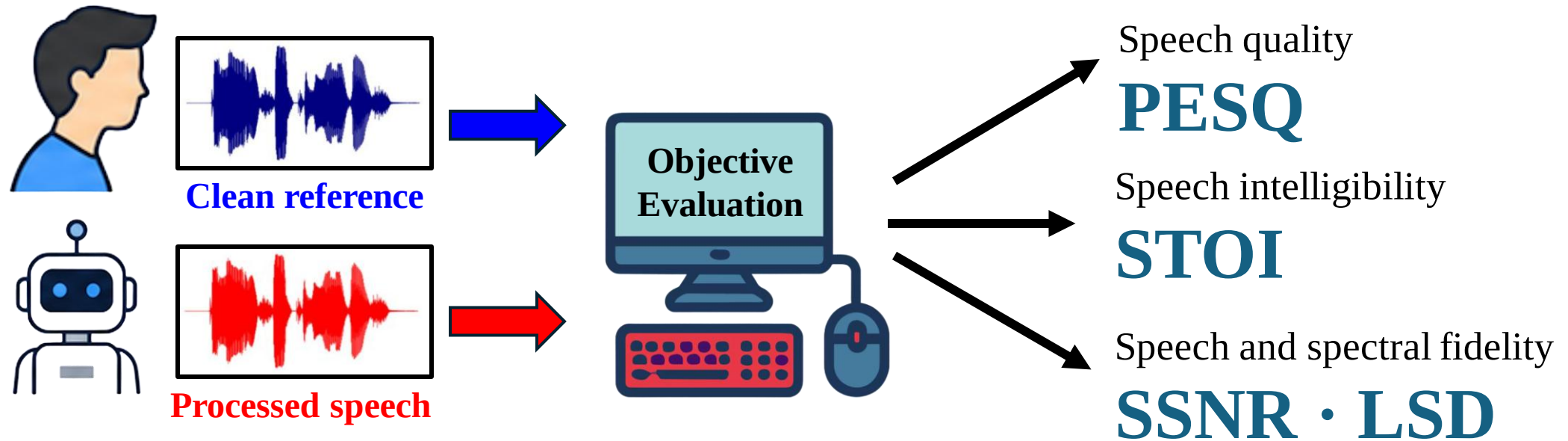
03 Speech Dereverberation



Speech reconstruction **estimates high-quality speech** from low-quality speech, such as noisy, reverberant, or multi-speaker observations.

What Do Traditional Metrics Measure?

Objective metrics to assess speech distortions, quality (PESQ) [15] and intelligibility (STOI) [16]



PESQ – perceived quality of processed speech

STOI – intelligibility of processed speech

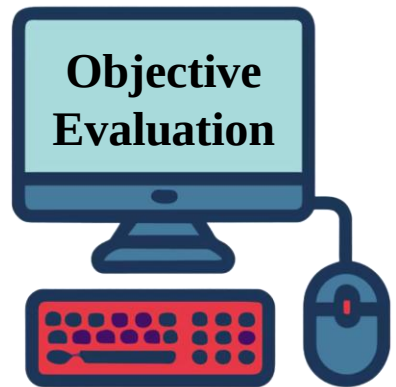
SSNR/LSD – signal and spectral fidelity to the reference



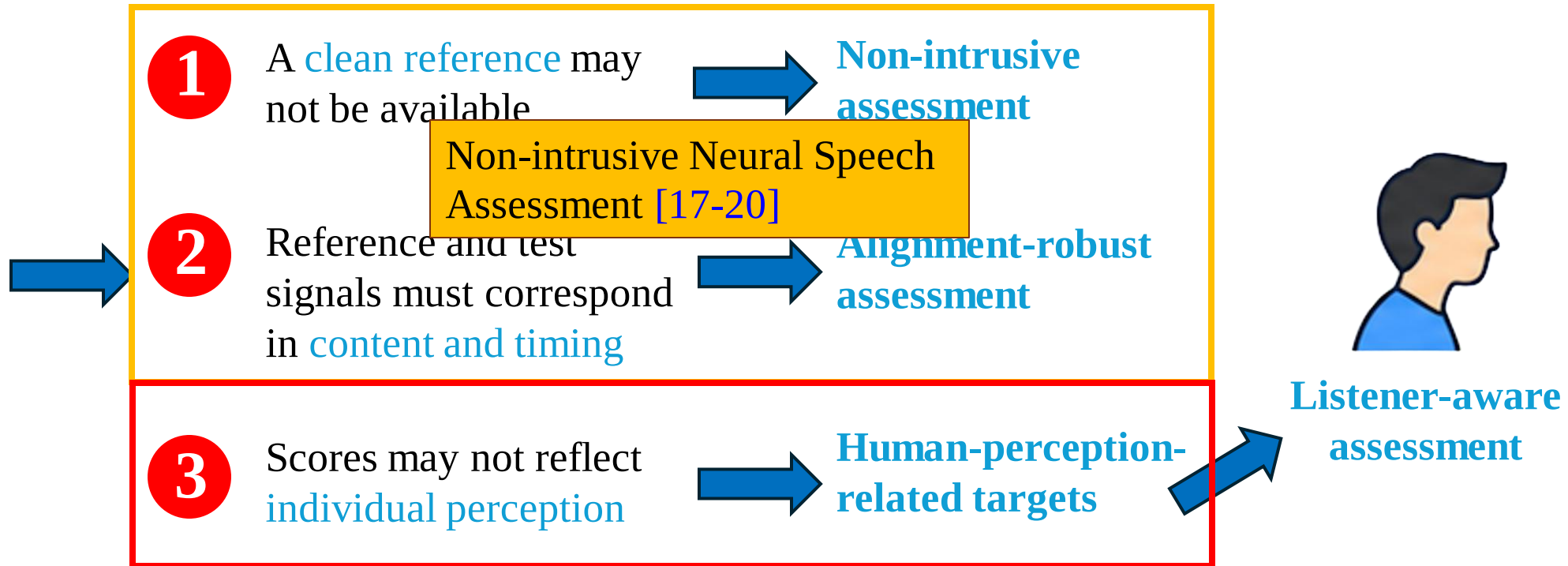
Usually we used different metrics to assess **different aspects** of generated speech.

What Do Traditional Metrics Miss?

Objective metrics to assess speech distortions, quality (PESQ) [15] and intelligibility (STOI) [16]



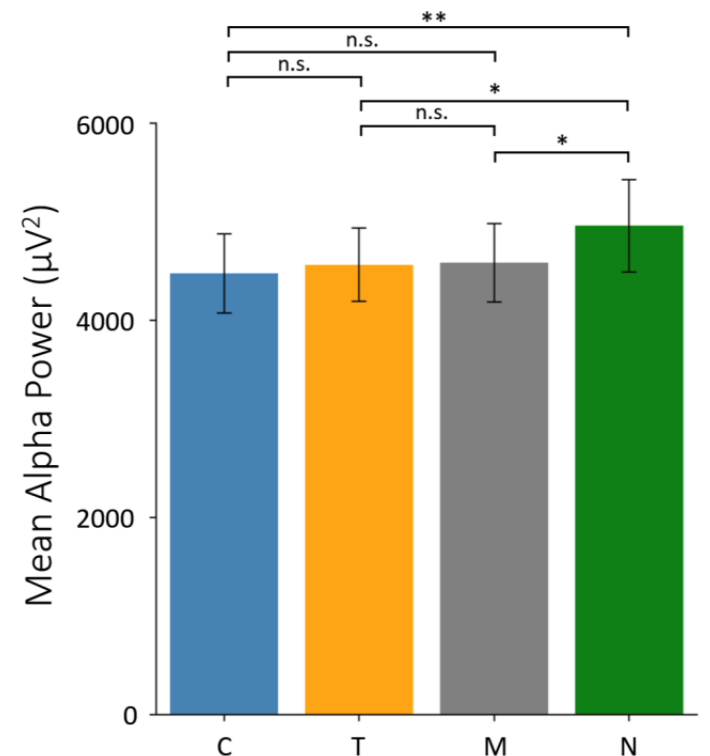
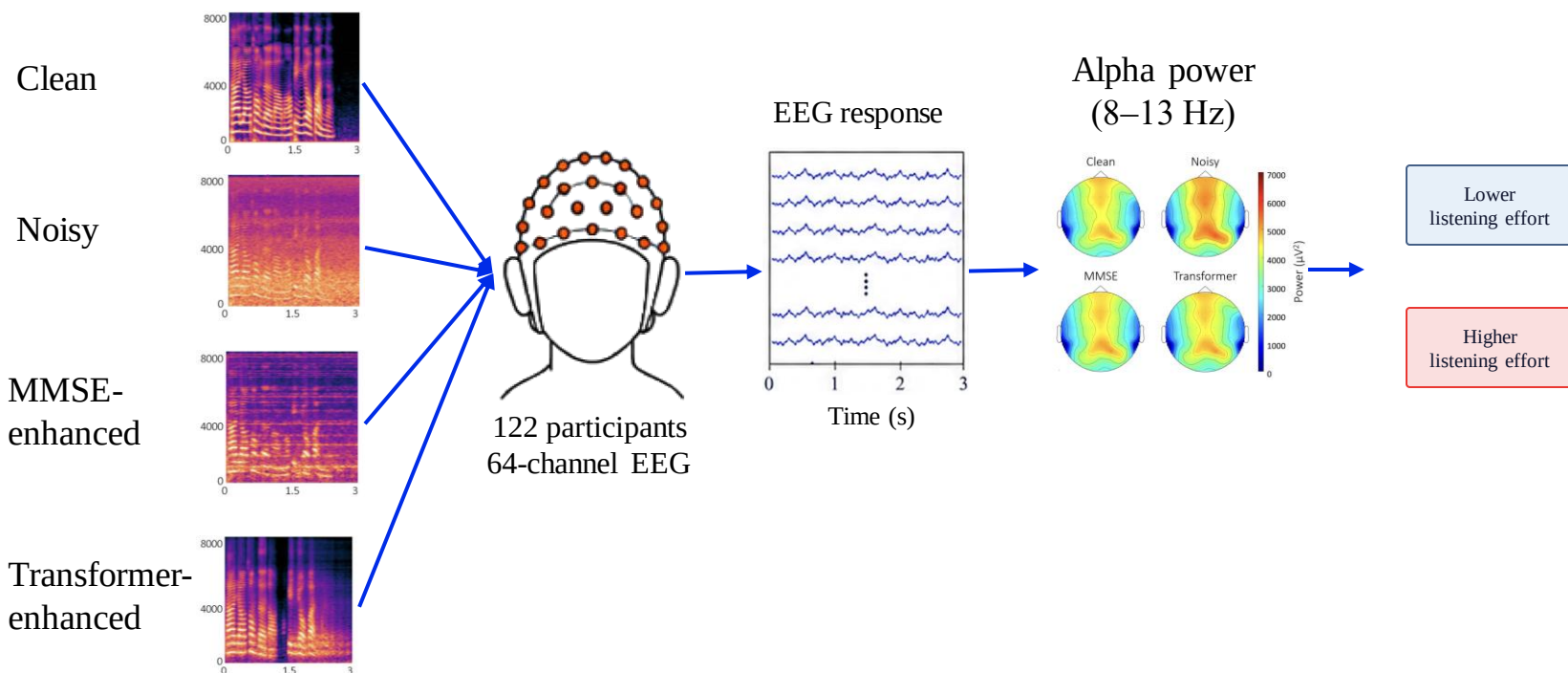
PESQ · STOI · SSNR · LSD



Objective metrics may not perfectly reflect individual listeners' perceptions:
(1) Specific **hearing profile**; (2) Variability in **behavioral test** results.

Electroencephalography (EEG) and Listening Effort

Assessing listening effort based on brain responses [21]



C = Clean; T = Transformer; M = MMSE; N = Noisy

Noisy speech produced higher alpha power than clean and enhanced speech.

Behavioural measures: comprehension accuracy and subjective intelligibility ratings

Physiological measure: EEG alpha activity

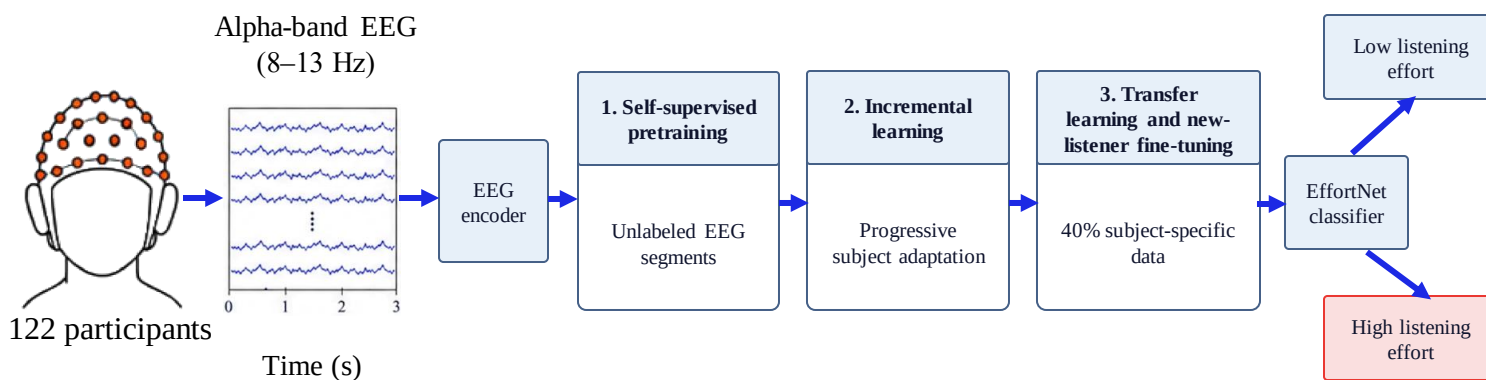


EEG responses can offer valuable insights into **listening effort**.

EffortNet Predicts Listening Effort by EEG

Assessing listening efforts based on brain responses [21]

EffortNet predicts listening conditions based on EEG



Training data

122 participants

64-channel EEG

Clean, noisy, and enhanced speech

Predicted listening effort

Probability (%)

	Clean	Transformer	MMSE	Noisy
Low effort	80.0	62.3	40.0	15.7
High effort	20.0	37.7	60.0	84.3

EEG responses to Transformer-enhanced speech were closer to the clean condition than those for MMSE-enhanced speech.

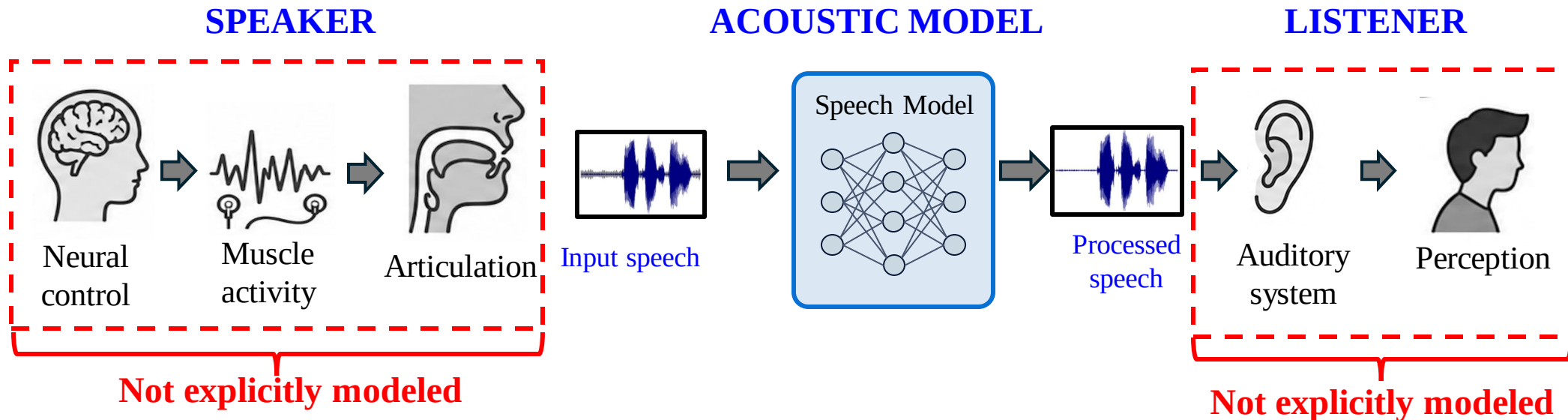
Comparing EffortNet with other models

Model	Accuracy (%)
CNN	62.3
STANet	61.1
EffortNet	80.9



Brain-response assessment can potentially **complement** objective evaluations and subjective listening test.

Limitations of Acoustic-Only Speech Processing



Requiring available and reliable acoustic signal sensing



Not directly observing speech-production mechanisms



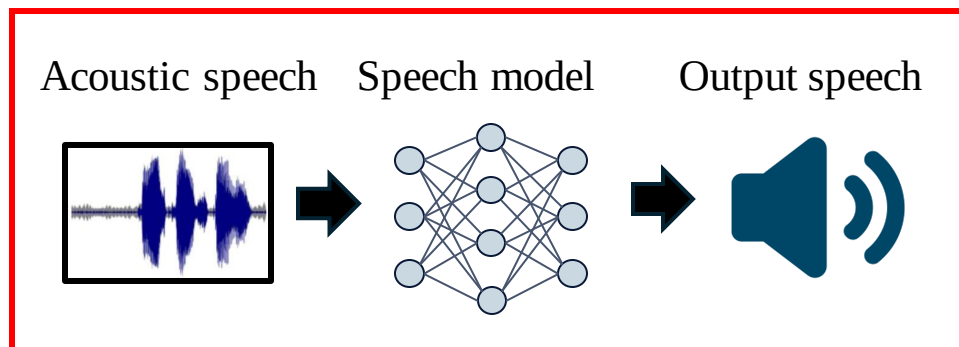
Not optimized for individual listener needs



The acoustic waveform is central to human communication, but complementary **biological information** can provide **additional insights**.

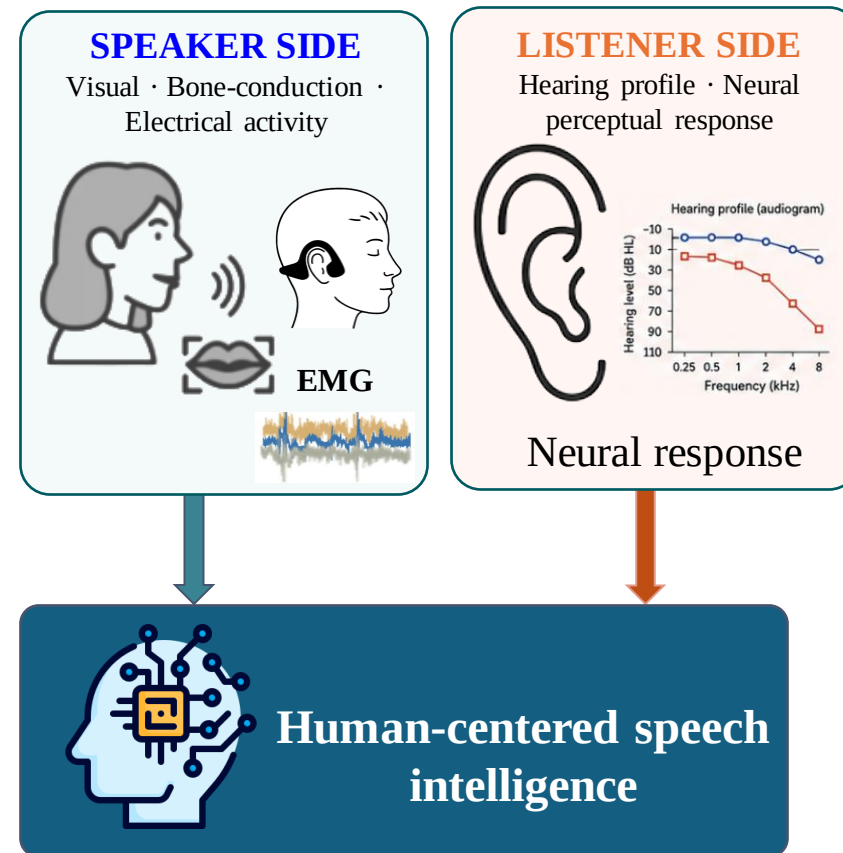
Acoustic and Biological Information for Speech Processing

ACOUSTIC-BASED MODELING

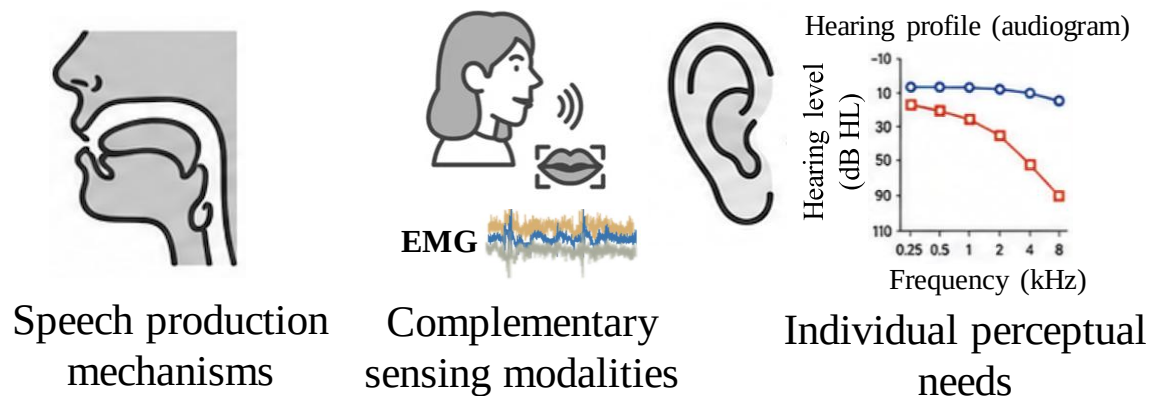


Additional
biological
information

BIO-INFORMED SPEECH MODELING



What may be missing?



Biological information **complements** the acoustic signal, **not replacing** it.

OUTLINE

1. **Speech Generation (Acoustic Only)**

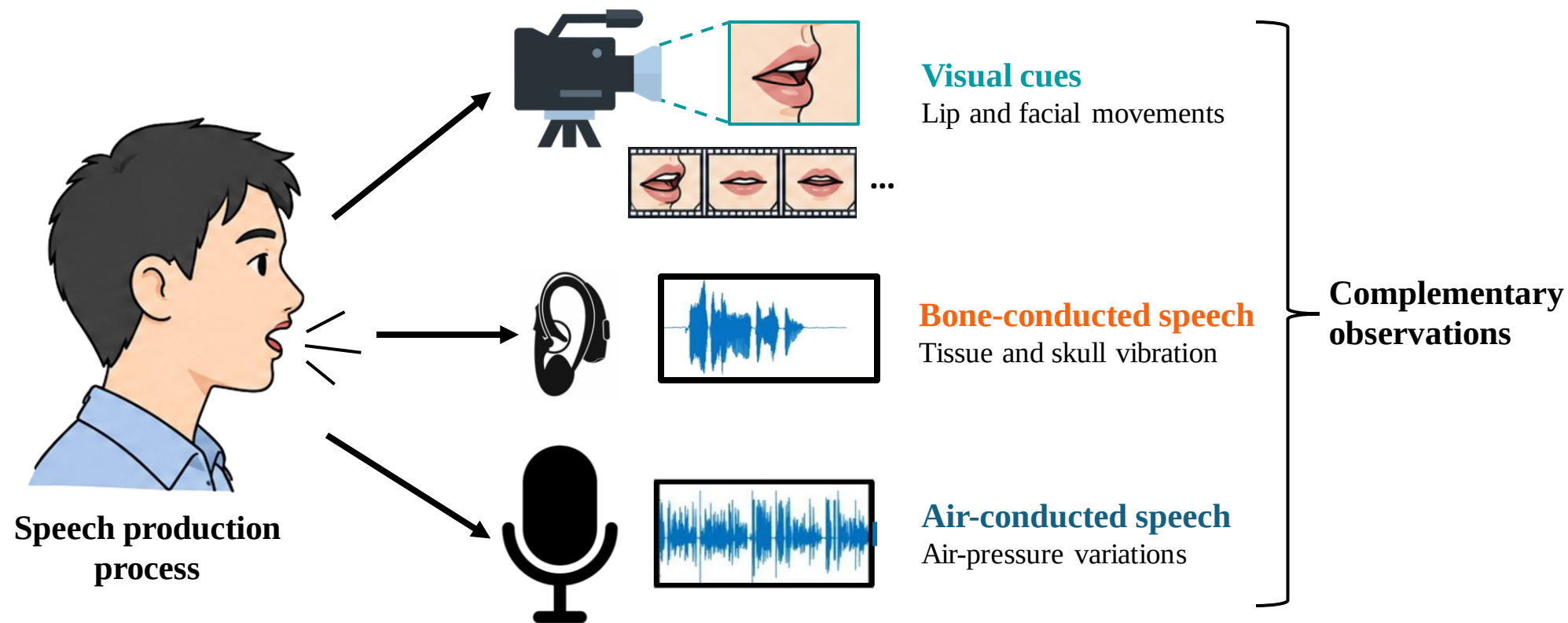
- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

2. **Complementary Speech Sensing**

- Audio-Visual Speech Enhancement
- Air and Bone Conducted Speech

Complementary Speech Sensing

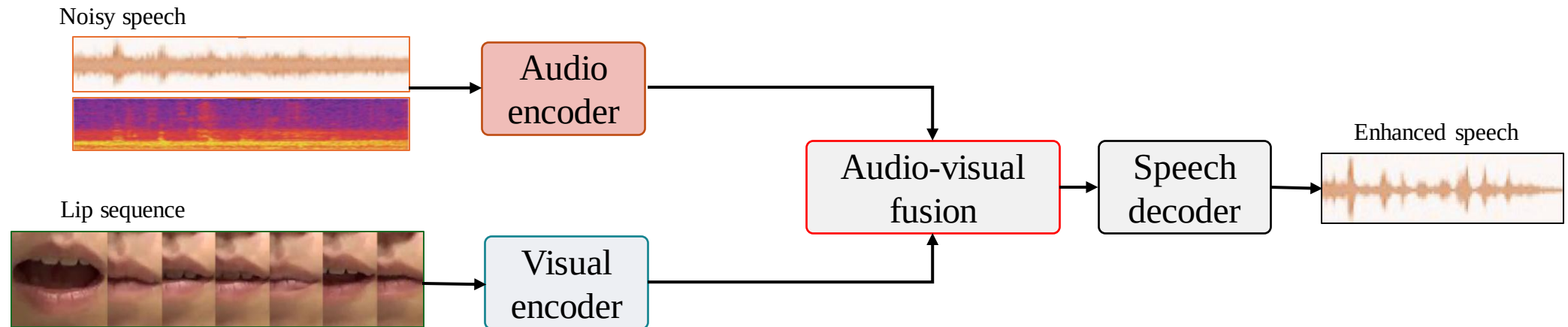
Multiple observations of the same speech event



Different sensors capture **complementary observations** of the same speech-production process.

Audio-Visual Speech Enhancement

Complementary information provided by visual cues [22-25]



What does the visual stream contribute?



Articulatory information



Temporal information



Robustness to acoustic noise

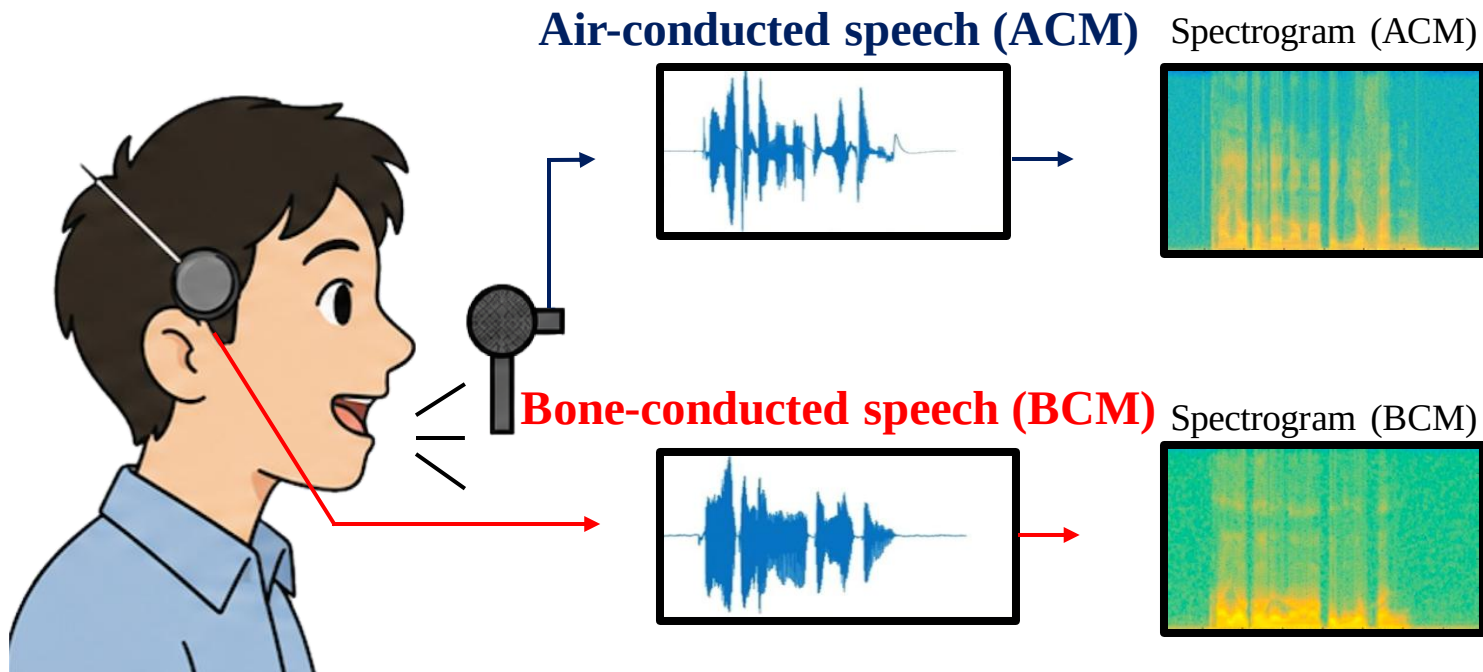
Visual cues are not affected by acoustic noise, but their usefulness depends on visibility and audio-visual synchronization.



Visual information is particularly useful when the acoustic signal is **severely degraded**.

Air- and Bone-Conducted Speech

Complementary information provided by vibration signals [26-29]



Same utterance with synchronized recording

Air-conducted speech

- Air-pressure variations
- Broader spectral bandwidth
- Clearer high-frequency detail
- More easily distorted by ambient noise

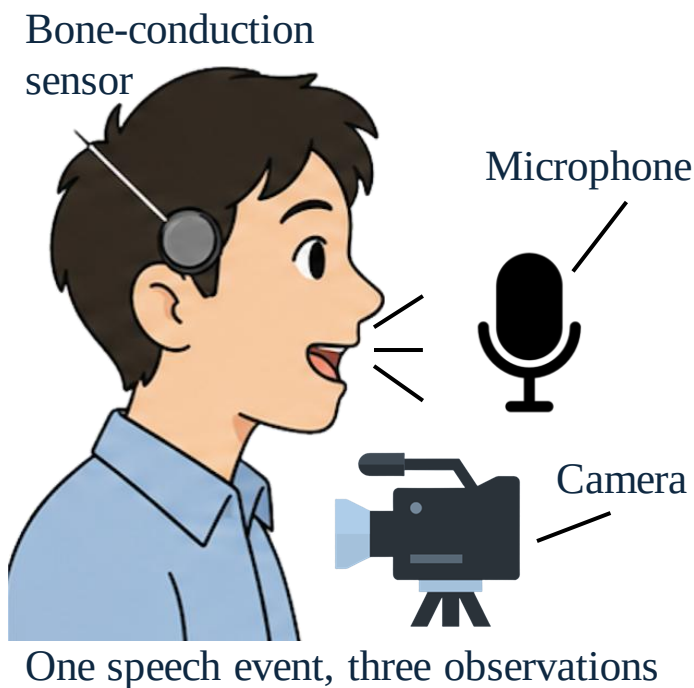
Bone-conducted speech

- Tissue and skull vibration
- Less affected by airborne noise
- Strong low-frequency and voicing information
- Limited high-frequency detail



Air-conducted speech provides **broader spectral detail**, whereas bone-conduction signals provide complementary information that is more **robust to noise**.

Air-Conducted Speech, Visual Cues, and Bone-Conducted Speech



Modality	Physical observation	Information retained	Main limitation
Air-conducted speech	Air-pressure variations	Broadband acoustic detail	Sensitive to ambient noise
Visual cues	Lip and facial movements	Visible articulation and timing	Occlusion and audio-visual misalignment
Bone-conduction signals	Tissue and skull vibrations	Voicing and low-frequency structure	Reduced high-frequency detail



These sensors are **not** simply giving us three versions of the **same information**. They represent different **physical consequences** of the same speech-production process.

OUTLINE

1. **Speech Generation (Acoustic Only)**

- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

2. **Complementary Speech Sensing**

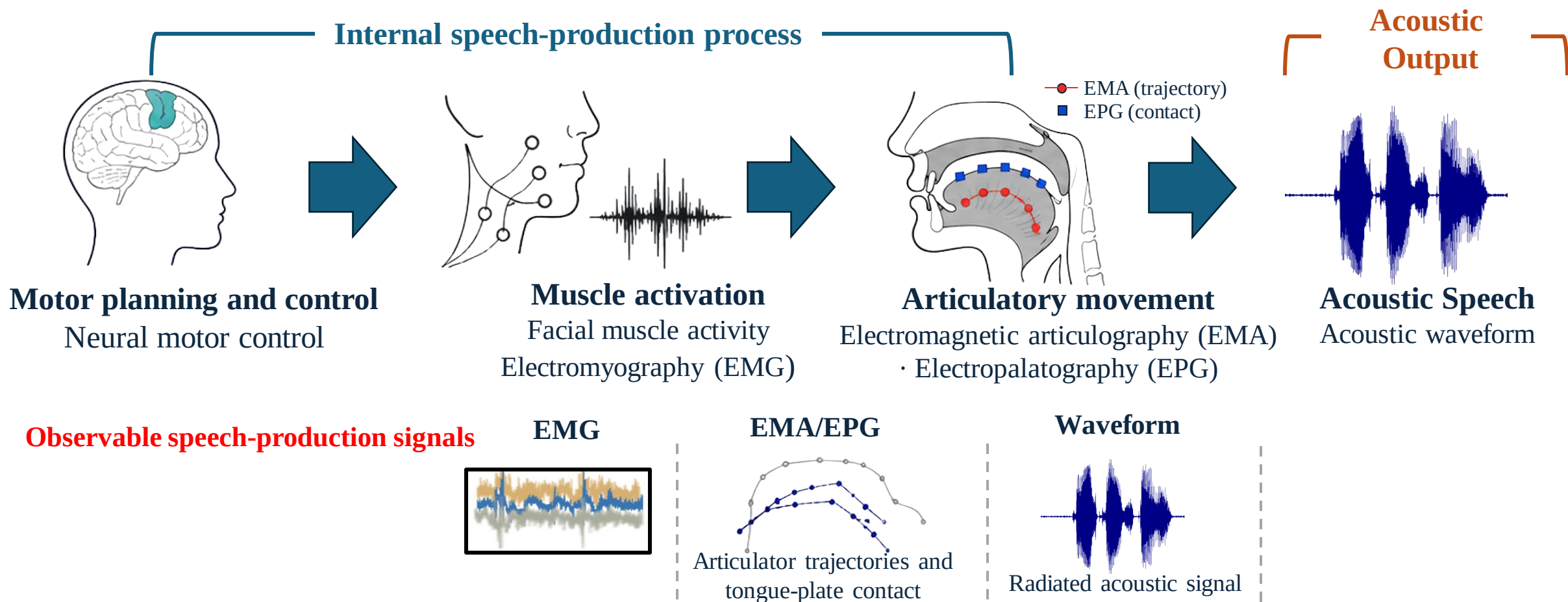
- Audio-Visual Speech Enhancement
- Air and Bone Conducted Speech

3. **Speech Production and Physiological Signals**

- Physiological and Articulatory Observations of Speech Production

Physiological and Articulatory Signals in Speech Production

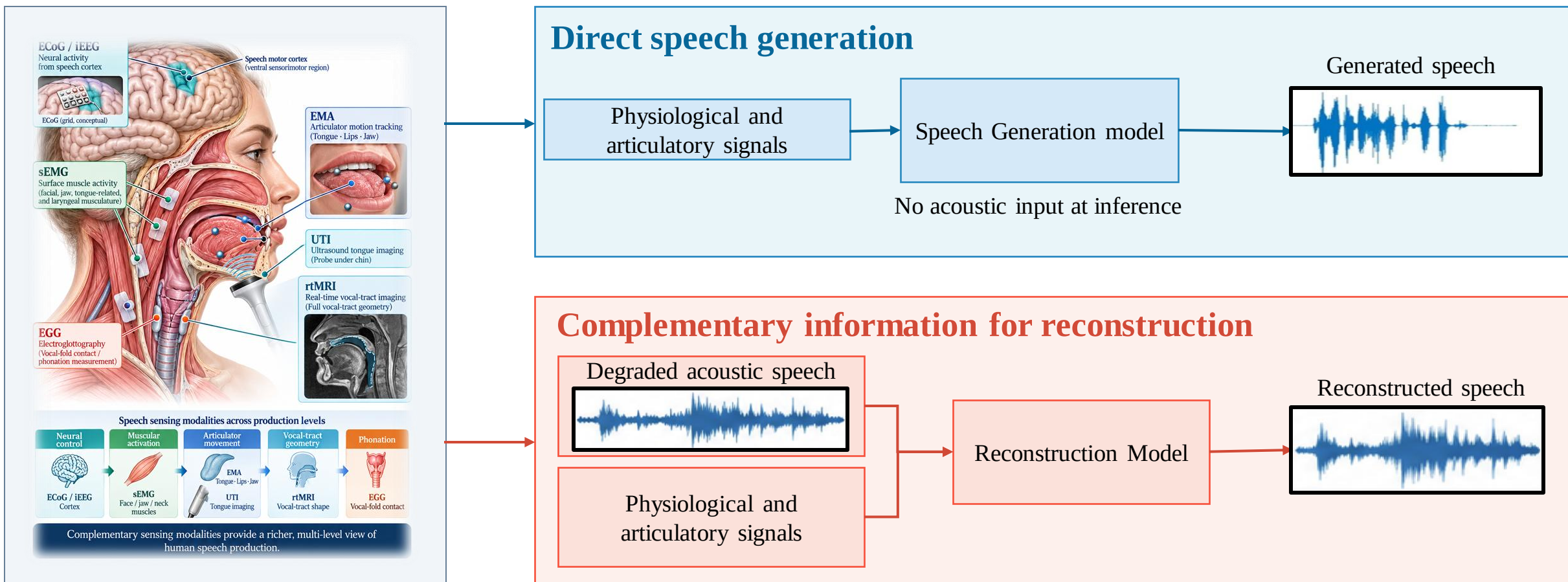
Physiological and articulatory signals [30-54]



⚠️ Physiological and articulatory signals can support **speech generation**, **speech enhancement**, and **silent speech interfaces**.

Generation and Reconstruction

Physiological and articulatory signals [30-54]



Types of Sensors

Categorized based on sensed information [30-54]

STAGE OBSERVED	SENSORS	WHAT THEY RECORD
 Neural control	→ EEG · iEEG / ECoG / sEEG	Electrical population activity
 Muscle activity	→ EMG	Speech-related muscle activation
 Articulation	→ Video · EMA · rtMRI · ultrasound	External motion, point motion or internal views
 Phonation / vibration	→ EKG · bone-conduction	Vocal-fold contact or mechanically transmitted vibration

⚠ These signals are **complementary**, but each provides an **incomplete** view. The right choice depends on which **functions** remain available, what the user can **tolerate**, and what the **task requires**.

Modalities: Audio, Movement and Vibration

Categorized by sensor types [30-54]

	Modality	What it adds	Main burden	Real-time potential
	Audio	Acoustic content / prosody	Noise or absent voice	High, with causal model
	Video [22-25]	Visible speech timing	Occlusion / synchrony	High, with camera delay
	Vibration [26-29]	A different sound path	Bandwidth / coupling	Promising for wearables
	EMA [30, 31]	Sparse point trajectories	Attached coils	Fast, constrained setup
	MRI [40-42]	Internal tract views	Scanner / slices / noise	Acquisition $\neq$ wearable use
	UTI [43, 44]	Tongue surface motion	Probe stabilization	Model and setup dependent



Qualitative engineering judgments of **information** vs. **acquisition burden** — not every cited model runs in real time.

Modalities: Muscle, Physiology and Hearing

Categorized by sensor types [\[30-56\]](#)

	Modality	What it observes	Main burden	Real-time potential
	sEMG [45-48]	Muscle activation	Electrodes / calibration	Silent-speech potential
	EGG [49]	Vocal-fold contact	Neck contact / phonation	Fast signal acquisition
	Scalp EEG [50, 51, 52]	Mixed neural activity	Artifacts / low specificity	Evidence window matters
	iEEG / ECoG [53, 54]	Local cortical activity	Surgery / coverage	Online studies exist
	Audiogram [55, 56]	Hearing thresholds	Incomplete hearing profile	Static model conditioning

 Qualitative engineering judgments of **information** vs. **acquisition burden** — not every cited model runs in real time.

OUTLINE

1. **Speech Generation (Acoustic Only)**

- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

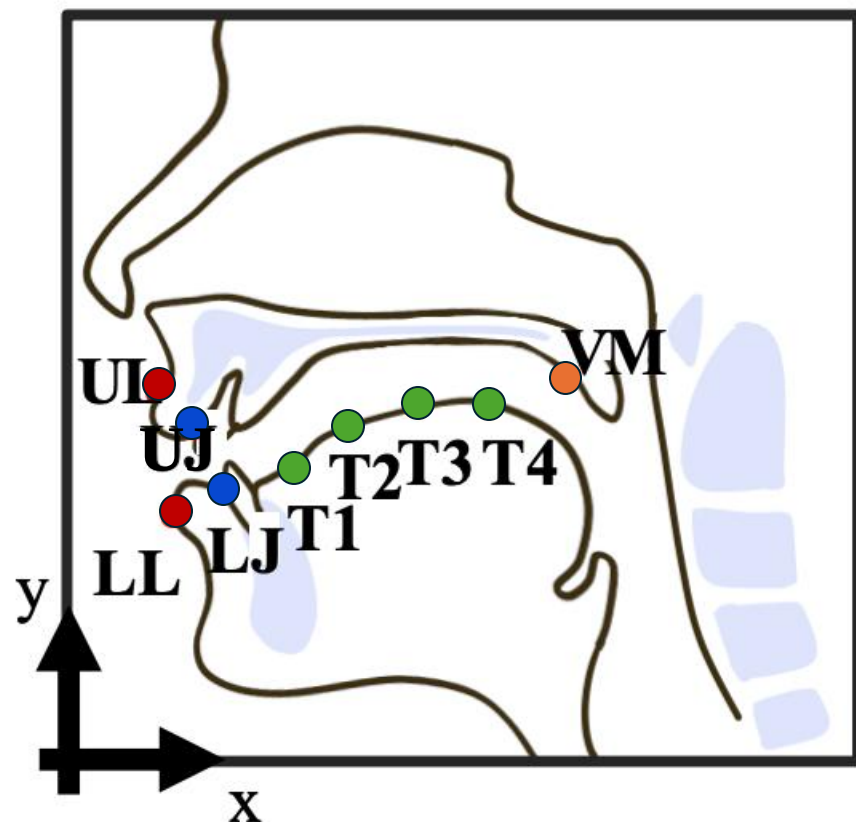
2. **Complementary Speech Sensing**

- Audio-Visual Speech Enhancement
- Air and Bone Conducted Speech

3. **Speech Production and Physiological Signals**

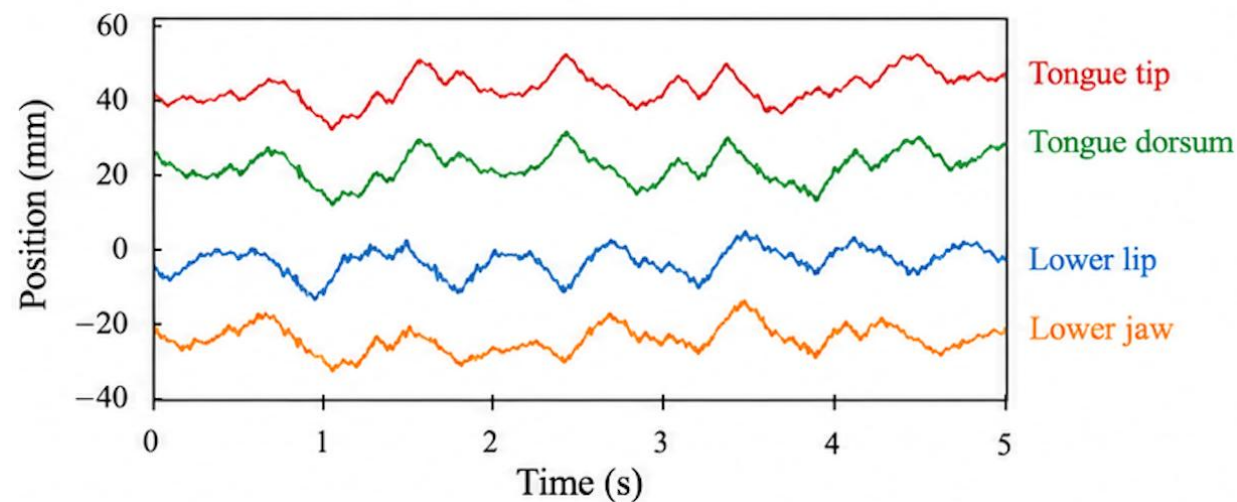
- Physiological and Articulatory Observations of Speech Production
- Electromagnetic Articulography (EMA)
- Electropalatography (EPG)
- Surface EMG (sEMG)

Electromagnetic Articulography (EMA)



What does EMA record?

- Cartesian position of each sensor
- Articulator movement over time
- Audio recorded in synchrony



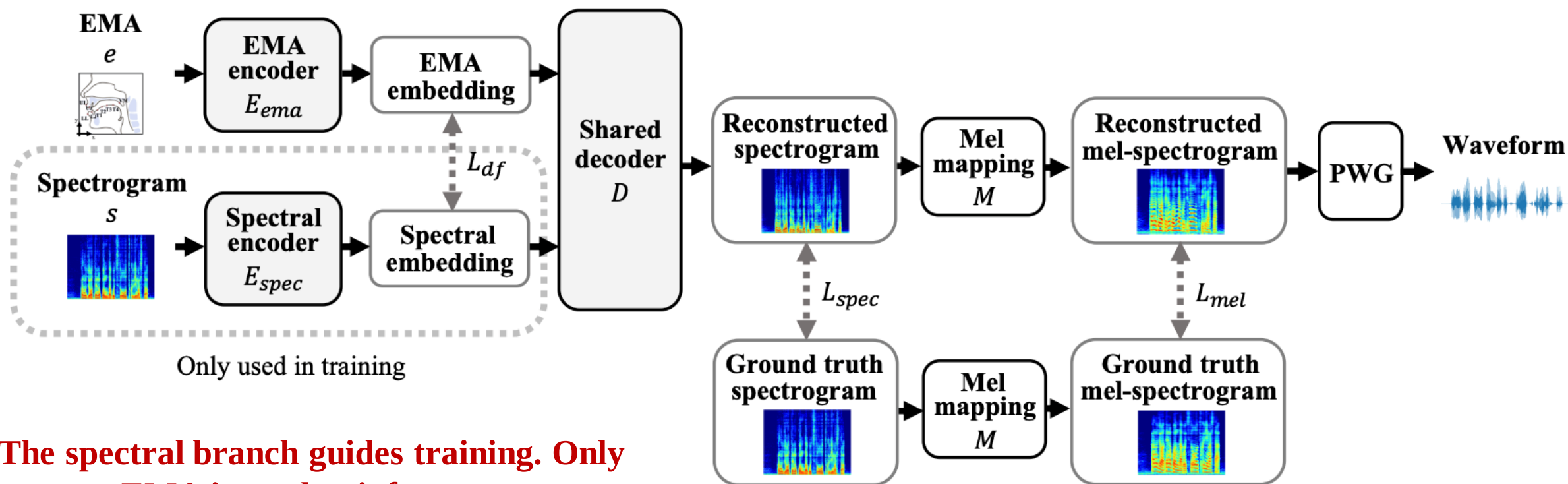
EMA2S data

9 sensor locations · EMA: 250 Hz · Acoustic: 16 kHz

⚠ EMA records articulator **positions and trajectories** during speech-generation process.

Multimodal Learning for Articulatory-to-Speech

EMA2S: An End-to-End Multimodal Articulatory-to-Speech System [30]



The spectral branch guides training. Only EMA is used at inference.

Training: EMA and spectrogram are jointly used.

Representation: Deep feature loss aligns the EMA and spectral embeddings.

Generation: The decoder predicts a spectrogram, and PWG synthesizes the waveform.

Inference: Only EMA signals are required.

EMA2S: Objective and Subjective Evaluation

EMA2S: An End-to-End Multimodal Articulatory-to-Speech System [30]

Objective evaluation

System	MCD ↓	PESQ ↑	STOI ↑	CCR ↑
Baseline	7.815	1.279	0.696	0.818
S-I	8.264	1.259	0.679	0.796
S-II	7.334	1.320	0.702	0.841
S-III	8.445	1.303	0.697	0.831
EMA2S	7.176	1.350	0.716	0.868

S-I: spectrogram loss

S-II: spectrogram + mel-spectrogram loss

S-III: multimodal training + deep feature loss

EMA2S: multimodal training with the combined loss

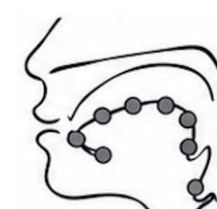
Listening test



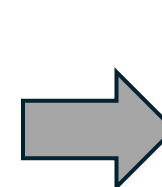
10 listeners

15 utterance pairs per listener

Fewer EMA sensors



9 sensors



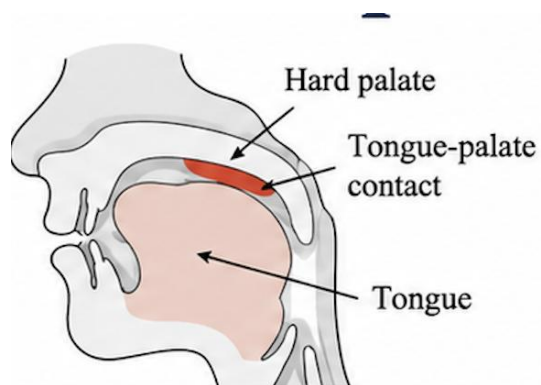
4 sensors: UL, LL, LJ, T1

The four-sensor configuration remained better than Baseline.

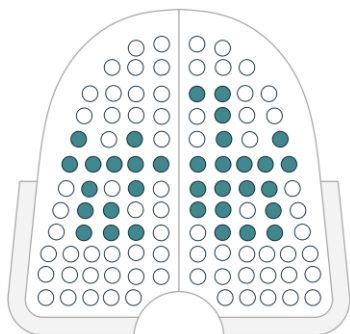


EMA trajectories can effectively support **direct speech generation** and that the system can retain its advantage with **fewer sensor locations**.

Electropalatography (EPG)



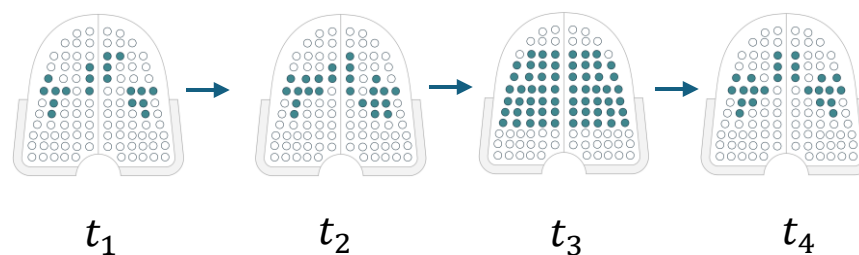
Custom artificial palate



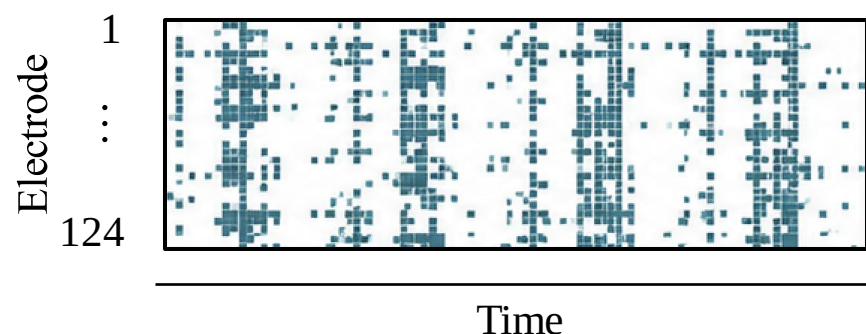
124 electrodes

- Contact = 1
- No Contact = 0

Contact pattern over time



Binary contact map sequence



What does EPG measure?

- Contact between the tongue and hard palate
- Spatial contact pattern during articulation
- Changes in contact over time

EPG2S dataset

One male speaker
319 paired EPG and speech recordings
Personalized system

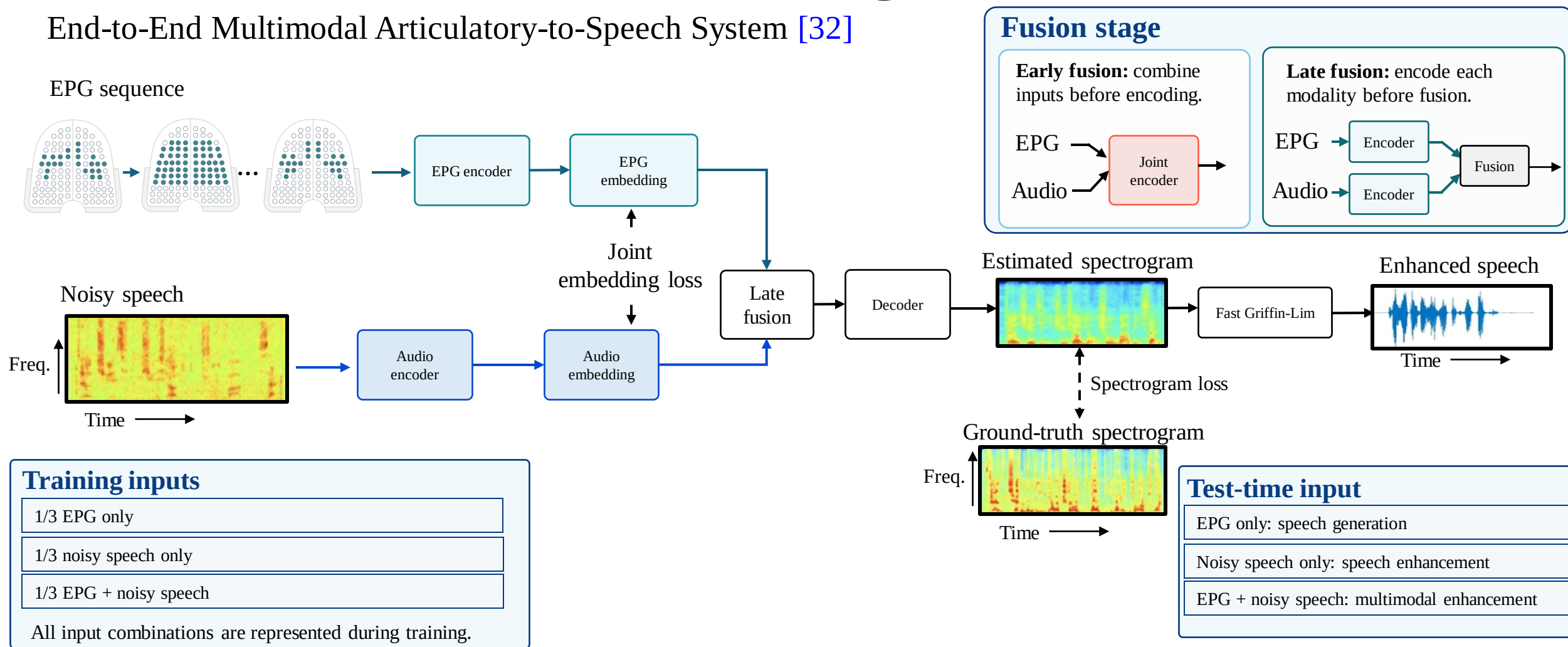
EPG does not measure the full tongue shape or continuous tongue position.



EPG records the location and timing of **tongue-palate contact** during speech.

EPG2S: Multimodal Learning with EPG and Audio

End-to-End Multimodal Articulatory-to-Speech System [32]



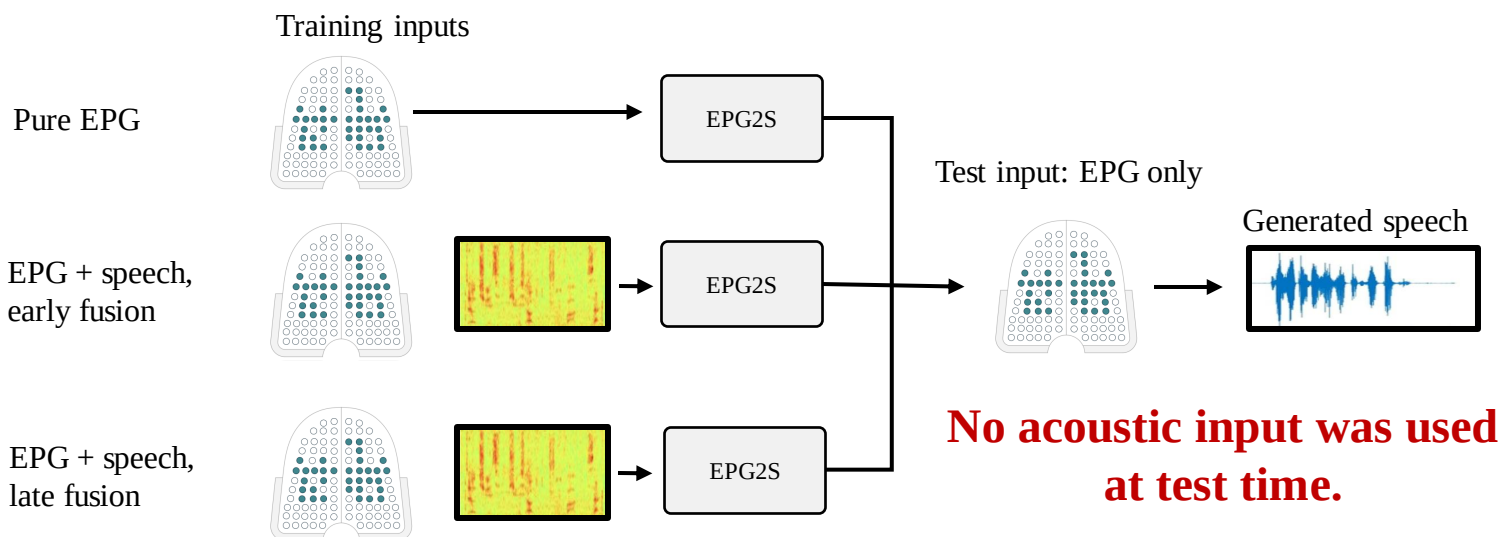
One personalized EPG2S model supports **three input** settings.

EPG2S: Speech Generation from EPG

End-to-End Multimodal Articulatory-to-Speech System [32]

Speech-generation scores

Training Input	PESQ ↑	STOI ↑	ESTOI ↑
Pure EPG	2.039	0.583	0.321
EPG + speech, EF	1.996	0.588	0.346
EPG + speech, LF	2.048	0.592	0.341



Observations

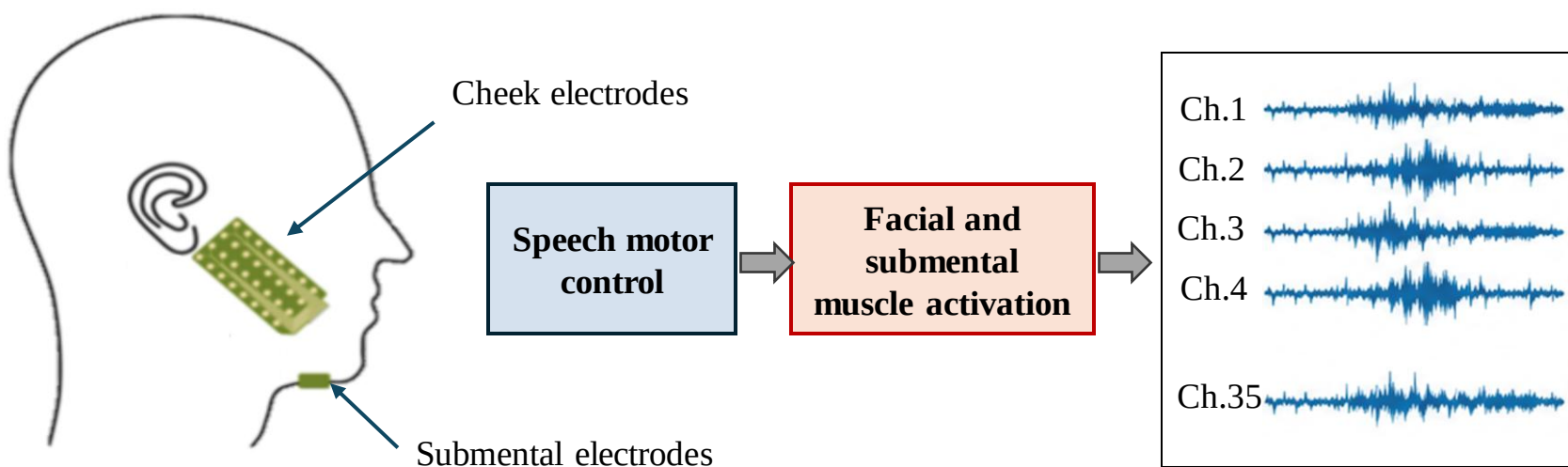
- The model generated speech from EPG alone
- Audio during training improved selected quality and intelligibility scores.
- Late-fusion training achieved the best PESQ and STOI, while early fusion achieved the best ESTOI



EPG and **acoustic** signals can provide complementary information during **multimodal learning** to improve EPG-to-speech generation.

Surface Electromyography (sEMG) for Speech

EMGSE: An end-to-end multimodal acoustic and EMG learning [45]



What does sEMG capture?

- Electrical activity associated with speech-related muscle activation
- Pre-acoustic information about articulation
- A set of multichannel signals that change over time

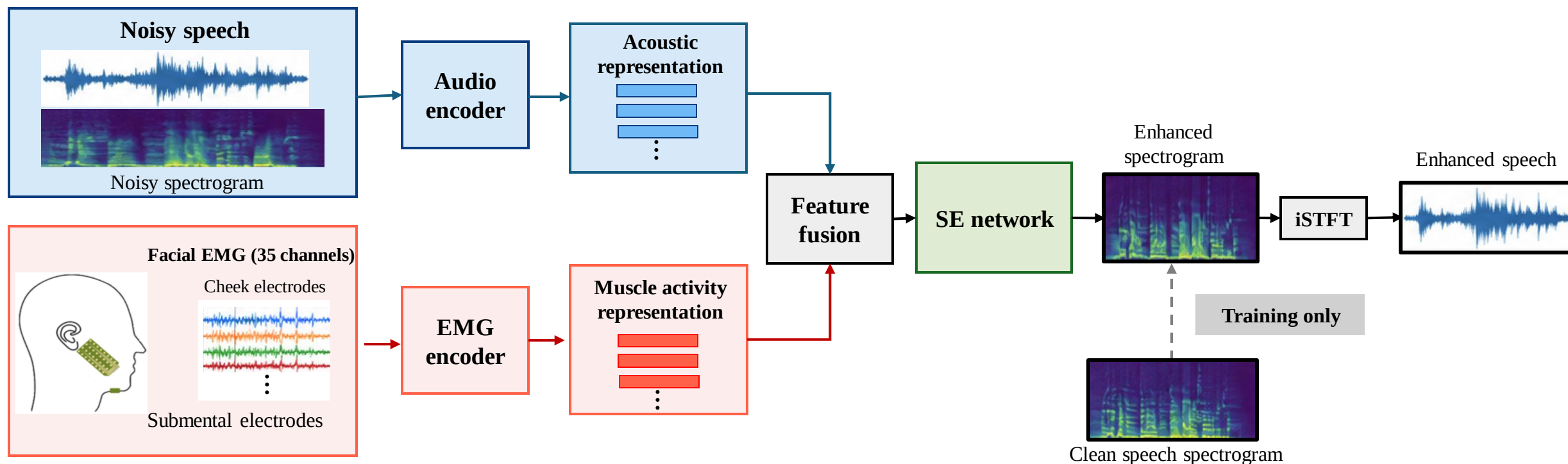
Practical characteristics

- Non-invasive surface recording
- Not directly capture airborne acoustic noise
- Sensitive to electrode placement and motion artifacts

⚠ Surface EMG provides a pre-acoustic physiological observation of speech-related **muscle activity**.

Acoustic and sEMG Fusion for Speech Enhancement

EMGSE: An end-to-end multimodal acoustic and EMG learning [45]



Acoustic: spectral information from the noisy recording.

Facial sEMG: synchronized muscle activity during articulation.



Acoustic signal provides **detailed spectral** information, while the EMG provides complementary **muscle-activity** information that is not directly affected by airborne acoustic noise.

EMGSE: Performance Across SNRs and EMG Configurations

EMGSE: An end-to-end multimodal acoustic and EMG learning [45]

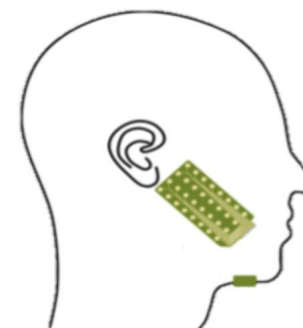
Performance at Different SNRs

	Noisy		SE(A)		EMGSE	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
-11dB	0.923	0.394	1.197	0.446	1.452	0.553
-6dB	1.138	0.481	1.608	0.579	1.829	0.658
-1dB	1.448	0.579	2.03	0.697	2.207	0.751
4dB	1.722	0.663	2.333	0.764	2.476	0.801
Avg.	1.308	0.529	1.792	0.621	1.991	0.691

EMGSE improves PESQ and STOI at every tested SNR.

Larger gains occur under the lowest SNR conditions.

Effect of EMG Channel Configuration



Full configuration: 35 channels

Cheek-only configuration: 28 channels

Configurations	PESQ ↑	STOI ↑
35 channels	2.046	0.709
28 cheek channels	2.039	0.711

The cheek subset gives comparable performance to the full configuration.



(1) **EMG+SE** outperforms audio-only SE(A) under all conditions. (2) Using only **28 cheek** EMG channels achieves performance comparable to that of the full configuration.

OUTLINE

1. **Speech Generation (Acoustic Only)**

- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

2. **Complementary Speech Sensing**

- Audio-Visual Speech Enhancement
- Air and Bone Conducted Speech

3. **Speech Production and Physiological Signals**

- Physiological and Articulatory Observations of Speech Production
- Electromagnetic Articulography (EMA)
- Electropalatography (EPG)
- Surface EMG (sEMG)
- Silent Speech Generation

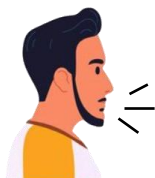
Silent Speech Interfaces

Speech generation from non-acoustic speech-production signals [35], [48], [57]

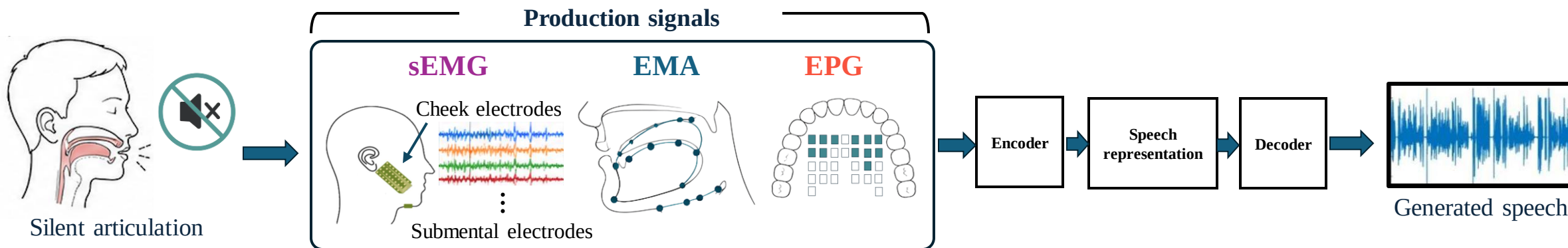
Potential uses



Communication when phonation is unavailable



Intentional silent communication



Practical limitations:

Speaker and session variability

Sensor placement and calibration requirements

Silent and vocalized articulation mismatch

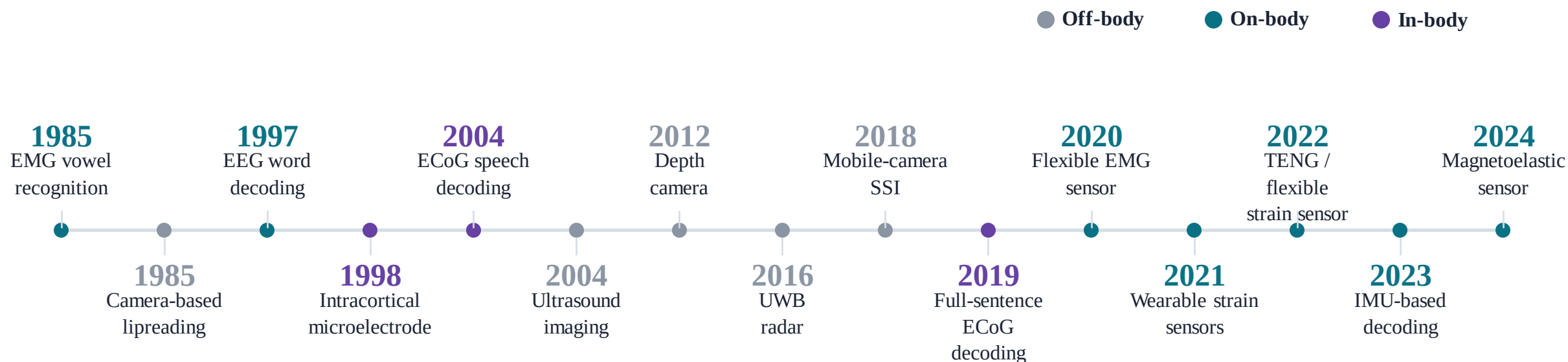
Limited speech intelligibility and naturalness



Silent speech interfaces generate **speech** from **articulatory** movements or speech-related **muscle activity**. However, there are practical limitations to further address:

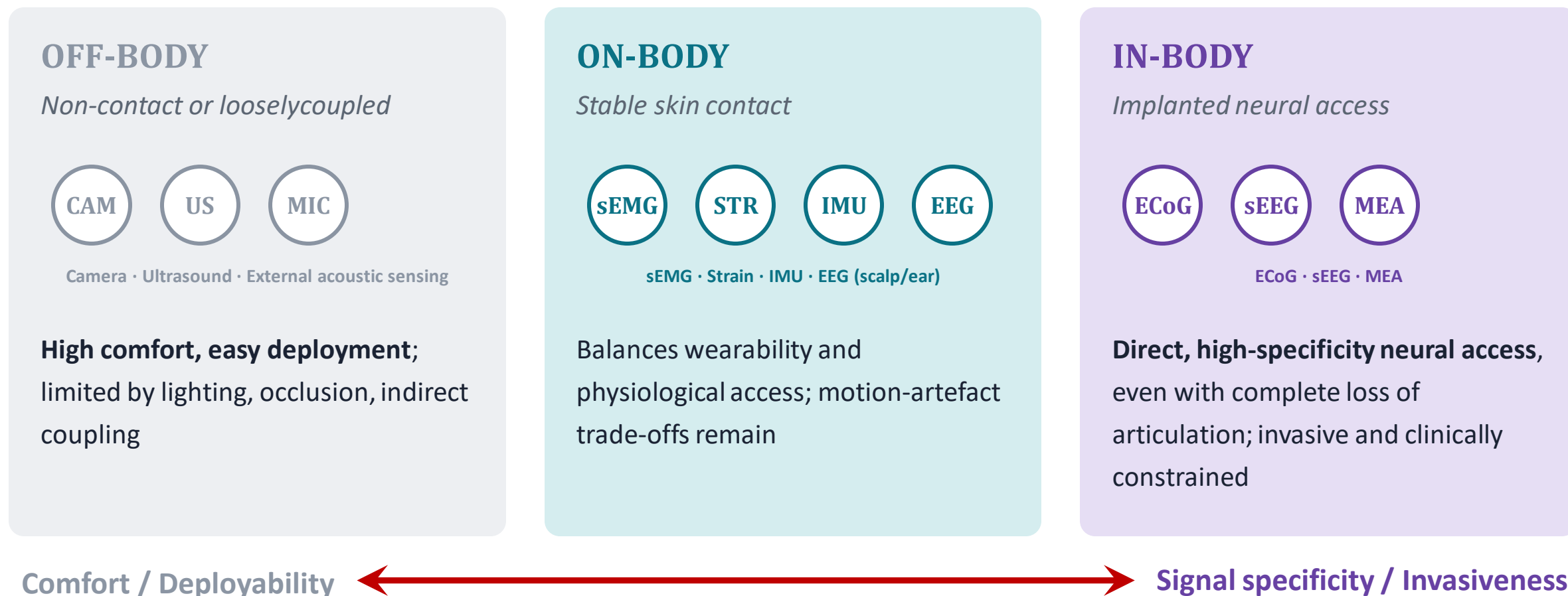
History of Silent Speech Interfaces

Timeline and milestones sourced from [57] (Tang et al., Nature Sensors, 2026), Figure 2.



Off-body / On-body / In-body Sensing

Timeline and milestones sourced from [57] (Tang et al., Nature Sensors, 2026), Figure 1.



Summary of Sensors

Timeline and milestones sourced from [57] (Tang et al., Nature Sensors, 2026), Table 1.

Sensor	Biological Signal	Spatial Res.	Tempo Res.	Strength	Limitation
Optical	Visible/IR lip & face motion	sub-mm–mm	30–60 Hz	Non-contact; commodity hardware	Lighting, occlusion, angle sensitivity
Ultrasound	Vocal-tract acoustic time-of-flight	mm-scale	30–100 Hz	Deep vocal-tract access; robust to light	Bulky probe; airflow noise
IMU	Motion-scale kinematics	Motion-scale	100–200 Hz	Micro-motion capture; low cost/size	Motion artefacts; head-movement coupling
Strain	Piezoresistive/capacitive deformation	mm-scale	1–1000 Hz	Skin-conformal; comfortable	Thermal/mechanical drift
sEMG	Surface myopotential	mm-scale	1–2 kHz	Direct muscle intent; mature tech	Impedance drift; motion/sweat artefacts
TENG	Triboelectric charge transfer	mm-scale	Event-triggered (~ μ s)	Self-powered; high-SNR pulses	Session-to-session variability
EEG (scalp)	Scalp neural potentials	cm-scale	1–2 kHz	Non-invasive brain intent	Low SNR; setup time
ECoG	Subdural cortical LFP	mm-scale	1–5 kHz	High bandwidth; low noise	Requires craniotomy
sEEG	Depth-electrode stereo-EEG	mm-scale	1–5 kHz	3-D access to speech network	Hemorrhage/infection risk
MEA	Intracortical spikes	Cellular-scale	30 kHz	Highest information rate	Most invasive; foreign-body response



These signals are **complementary**, but each provides an **incomplete** view. The right choice depends on which **functions** remain available, what the user can **tolerate**, and what the **task requires**.

OUTLINE

1. **Speech Generation (Acoustic Only)**

- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

2. **Complementary Speech Sensing**

- Audio-Visual Speech Enhancement
- Air and Bone Conducted Speech

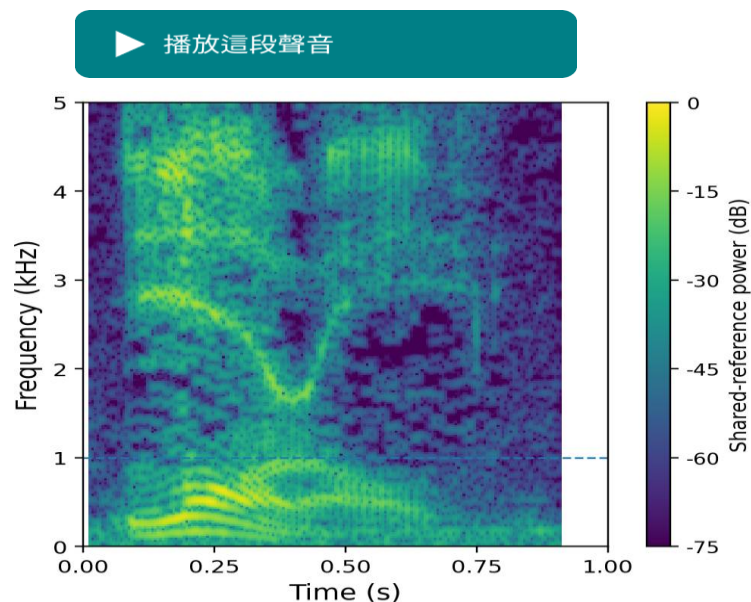
3. **Speech Production and Physiological Signals**

- Physiological and Articulatory Observations of Speech Production
- Electromagnetic Articulography (EMA)
- Electropalatography (EPG)
- Surface EMG (sEMG)
- Silent Speech Generation

4. **Listener-Aware Speech Processing**

- Hearing Profiles and Listener-Aware Speech Processing
- Audiogram and Hearing Loss Profiles

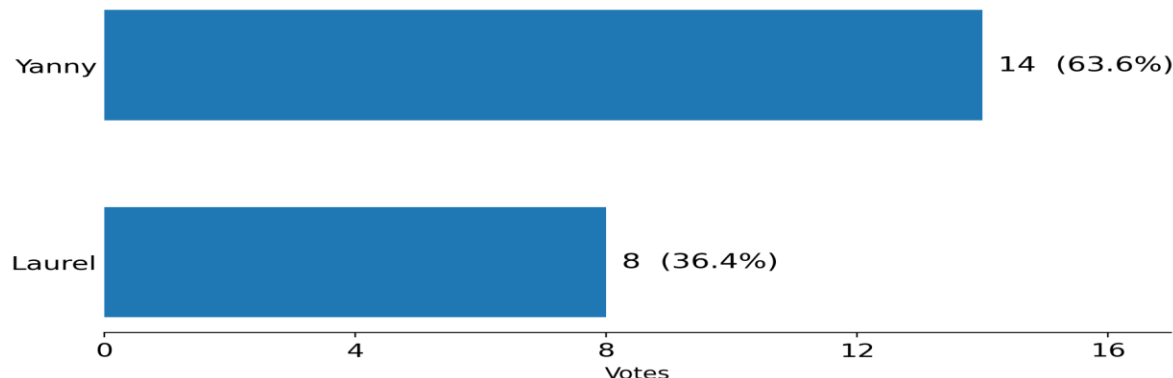
Laurel or Yanny ?



In the study [58], by modifying the preceding sound, 276 varied in whether they heard “Laurel” or “Yanny,” while 108 heard almost exclusively one word.

Findings

- (1) Listeners sensitive to **higher frequencies** tend to hear “Yanny,” while those sensitive to **lower frequencies** tend to hear “Laurel” instead.
- (2) The **preceding sound changed** how some listeners heard the same recording. Some listeners **switched** between the two words, while others **consistently heard** only one.

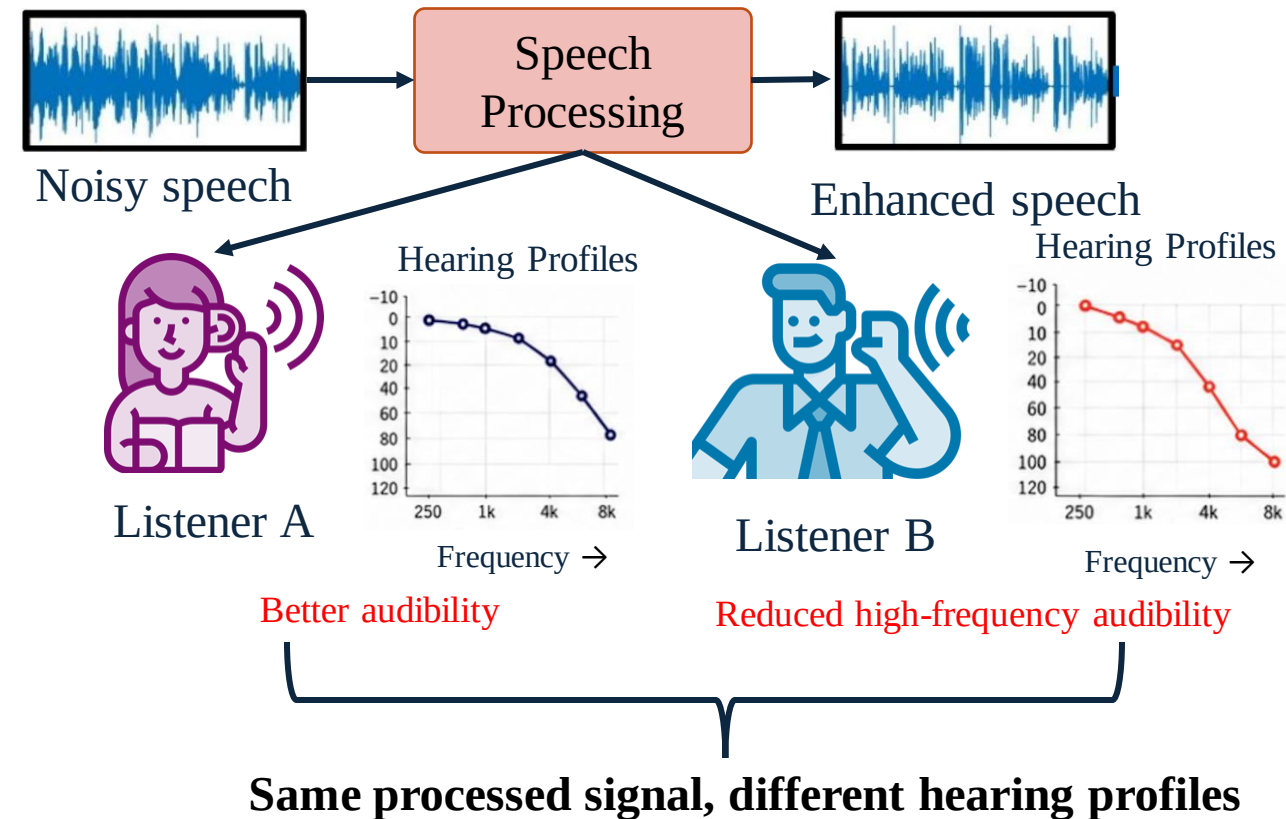


⚠ The same acoustic signal may be **perceived differently** by different listeners.

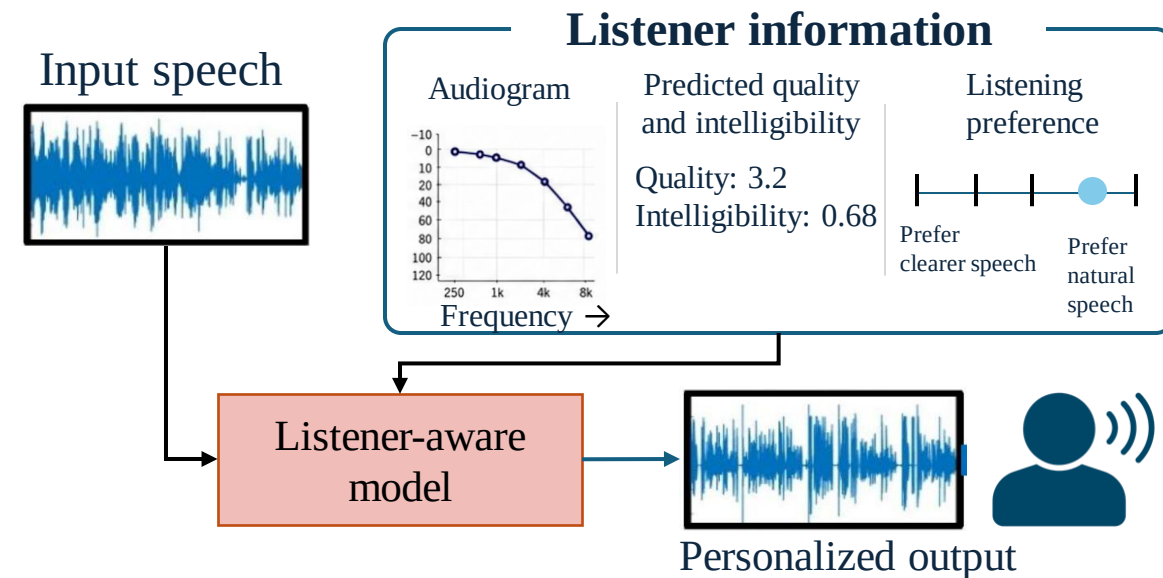
Hearing Profiles and Listener-Aware Speech Processing

Listener-aware speech processing [37, 51, 56, 59]

Conventional Speech Processing



Listener-Aware Speech Processing



Related research:

- **NeuroAMP**: audiogram-conditioned amplification
- **Preference-based SE**: enhancement using listener preference

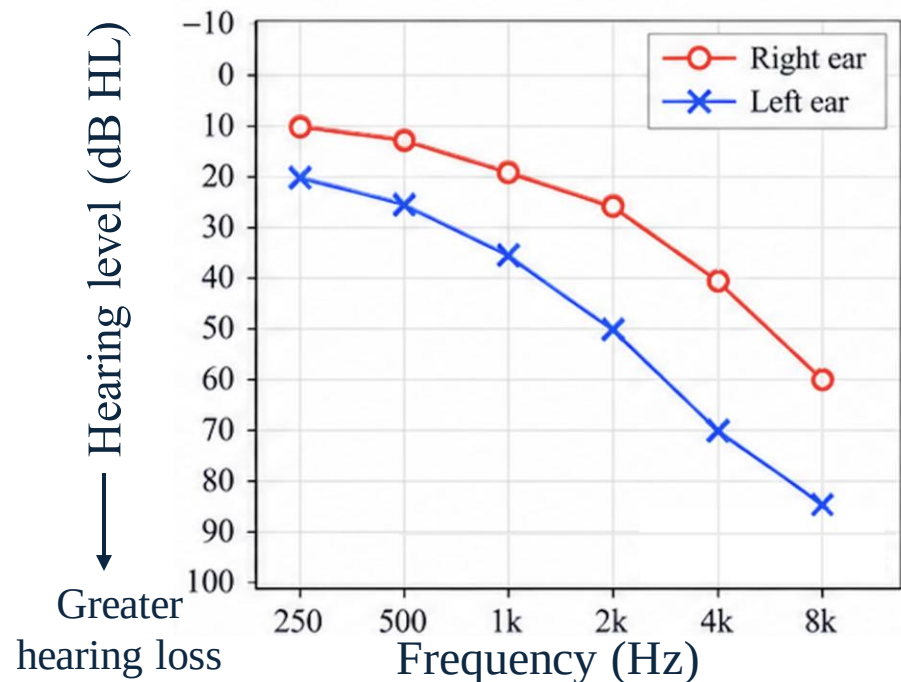


Listener information can **condition speech processing** or perceptual **assessment**.

Audiograms and Hearing Loss Profiles

Personalized hearing profile [56, 60, 63]

An Audiogram

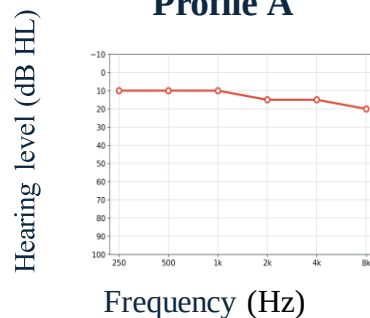


Frequency: pitch of the tone

Threshold: lowest level detected at each frequency

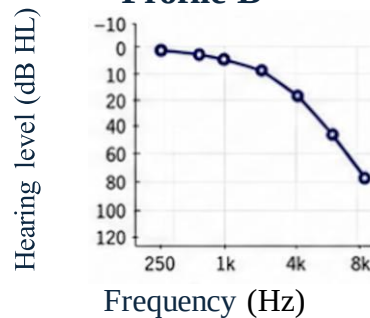
Different Hearing Profiles

Profile A



Near-flat thresholds

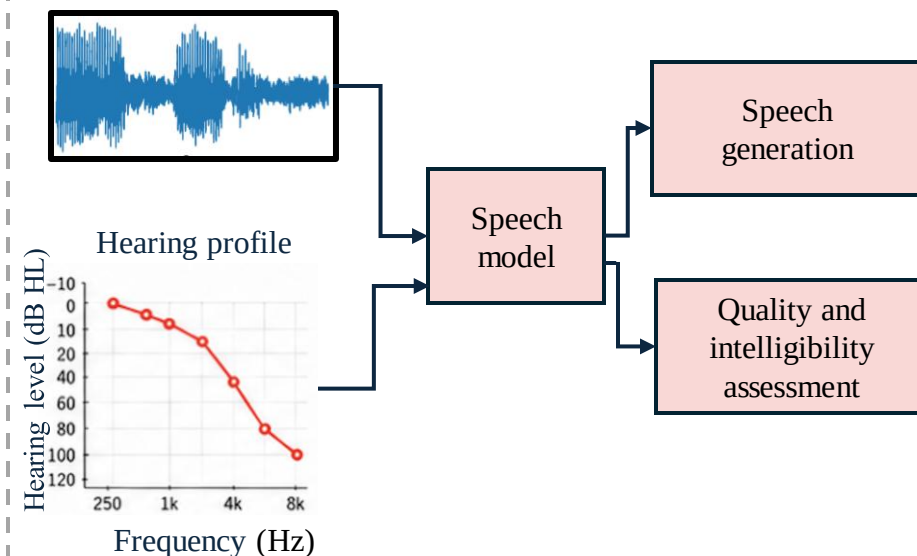
Profile B



High-frequency hearing loss

The same speech signal may not be equally audible to both listeners.

Audiogram-conditioned speech model



⚠ An **audiogram** indicates **frequency-specific hearing thresholds** for an individual listener.

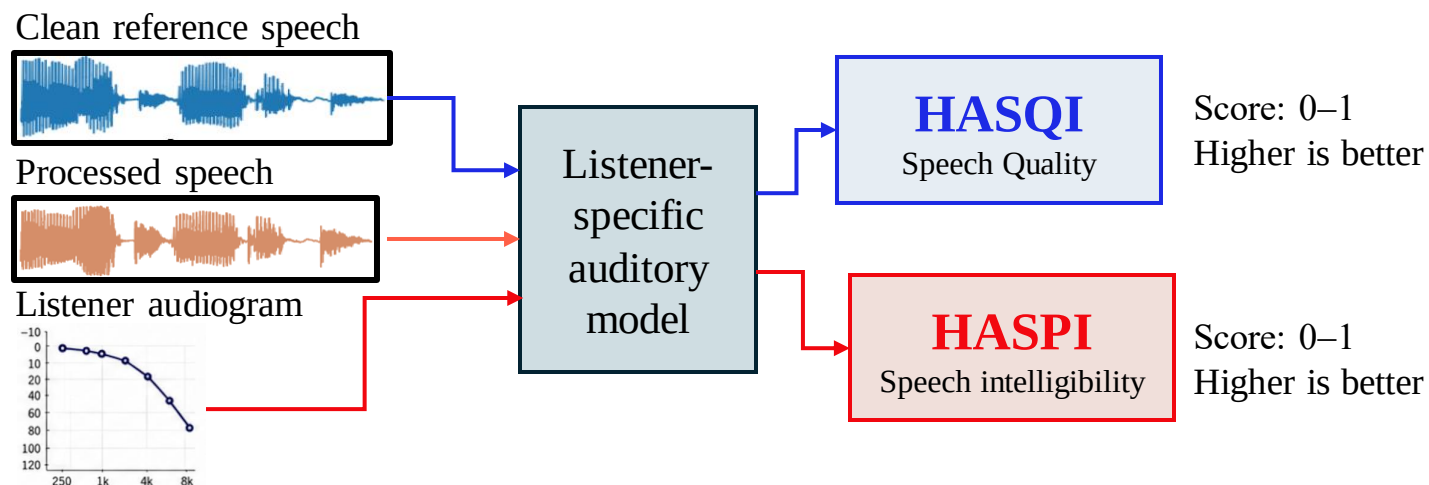
HASQI and HASPI

Reference-based metrics for hearing-aid speech assessment [61], [62]

Objective: Speech quality/intelligibility metrics for **normal hearing**, such as PESQ, STOI, and ESTOI do not fit to hearing-loss individuals

Solutions: Develop **hearing-profile-oriented** speech quality/intelligibility metrics

Reference-based Evaluation



What do the scores represent?

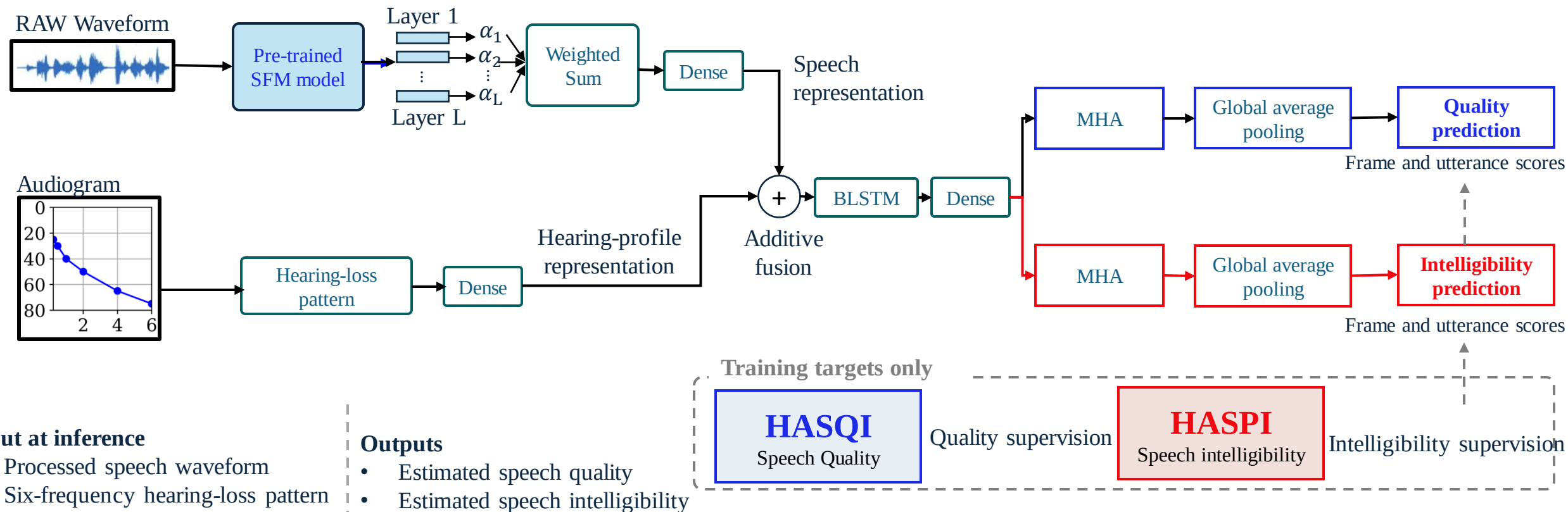
Metric	Assessment focus	Listener information
HASQI	Quality and signal distortion	Audiogram-dependent auditory model
HASPI	Speech intelligibility	Audiogram-dependent auditory model



HASPI and HASQI are widely used to **assess** hearing-aid performance, but both require a **reference signal**.

Non-Intrusive Hearing-Aid Speech Assessment

HASA-Net: Non-Intrusive deep-learning-based hearing-aid speech quality and intelligibility assessment [63]



We believe this framework could be **generalized** to account for **normal-hearing people** with different hearing profiles among.

OUTLINE

1. Speech Generation

- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

2. Complementary Speech Sensing

- Audio-Visual Speech Enhancement
- Air and Bone Conducted Speech

3. Speech Production and Physiological Signals

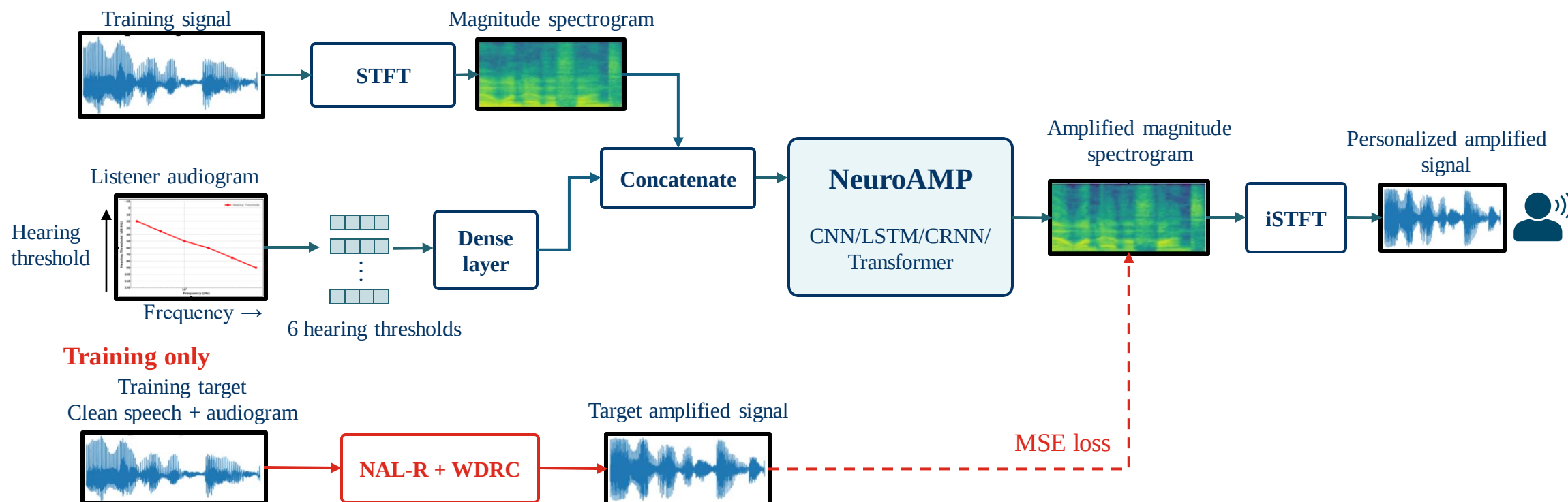
- Physiological and Articulatory Observations of Speech Production
- Electromagnetic Articulography (EMA)
- Electropalatography (EPG)
- Surface EMG (sEMG)
- Silent Speech Generation

4. Listener-Aware Speech Processing

- Hearing Profiles and Listener-Aware Speech Processing
- Audiogram and Hearing Loss Profiles
- Hearing-Profile-Conditioned Speech Generation (End-to-end Hearing Aids)

NeuroAMP

An end-to-end general-purpose deep neural amplifier for personalized hearing aids (HA) [56]

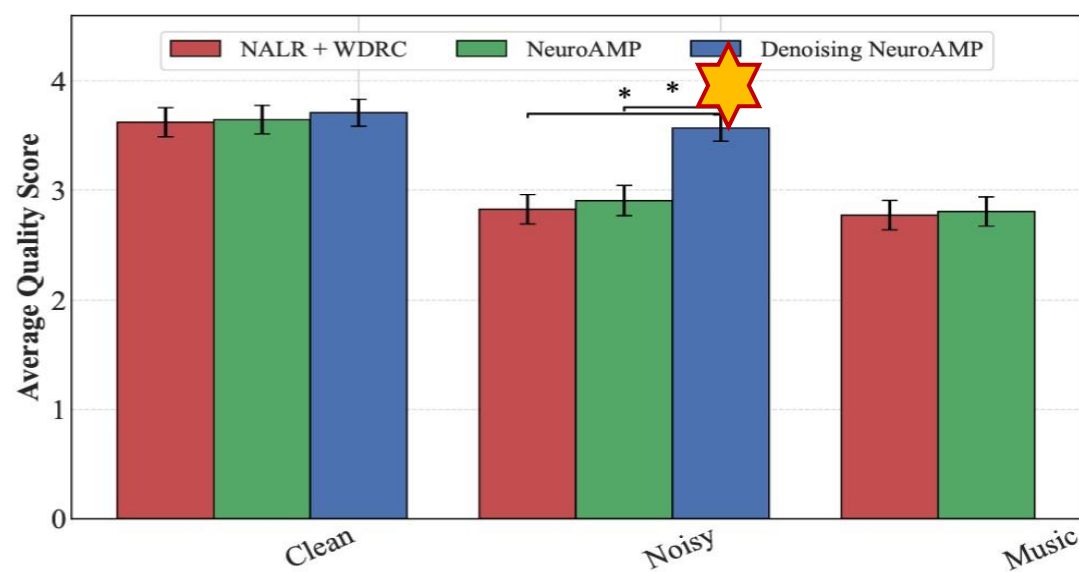


The audiogram directly **conditions** listener-specific **speech generation** (enhancement and amplification). NeuroAMP combines speech enhancement and amplification in a **single model**, reducing the potential **mismatch** and **complexity** of traditional two-stage hearing-aid (HA) systems.

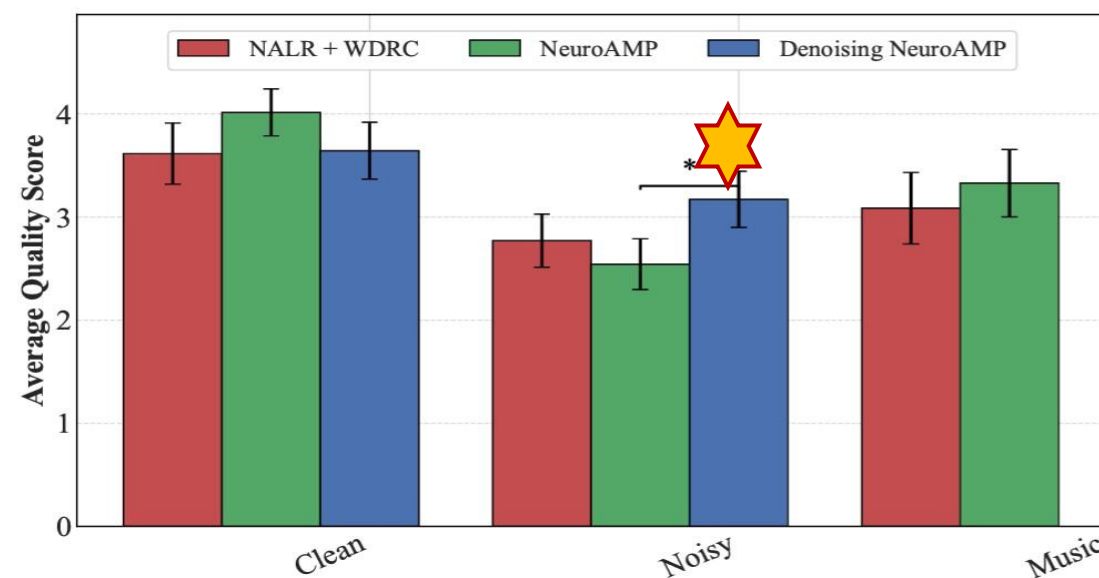
NeuroAMP versus Traditional HA Methods

An end-to-end general purpose deep neural amplifier for personalized hearing aids [56]

Simulation Results



Subjective Test Results from HA Users



(1) Simulation Results: NeuroAMP **performs comparably** to conventional HA in **clean** conditions, and **outperforms** conventional HA in **noisy** environments (by data augmentation).

(2) Subjective Tests: **Consistent trends** were observed, aligning with the simulation outcomes.

OUTLINE

1. Speech Generation

- Text-to-Speech (TTS) Synthesis
- Voice Conversion (VC)
- Speech Reconstruction

2. Complementary Speech Sensing

- Audio-Visual Speech Enhancement
- Air and Bone Conducted Speech

3. Speech Production and Physiological Signals

- Physiological and Articulatory Observations of Speech Production
- Electromagnetic Articulography (EMA)
- Electropalatography (EPG)
- Surface EMG (sEMG)
- Silent Speech Generation

4. Listener-Aware Speech Processing

- Hearing Profiles and Listener-Aware Speech Processing
- Audiogram and Hearing Loss Profiles
- Hearing-Profile-Conditioned Speech Generation (End-to-end Hearing Aids)

5. Summary and Practical Challenges

Summary and Practical Challenges

Topics	Categories
Sensor types	Off-body / On-body / In-body sensing
Functions	Multimodal learning / Silent speech generation
Sides	Speaker-side / Listener-side

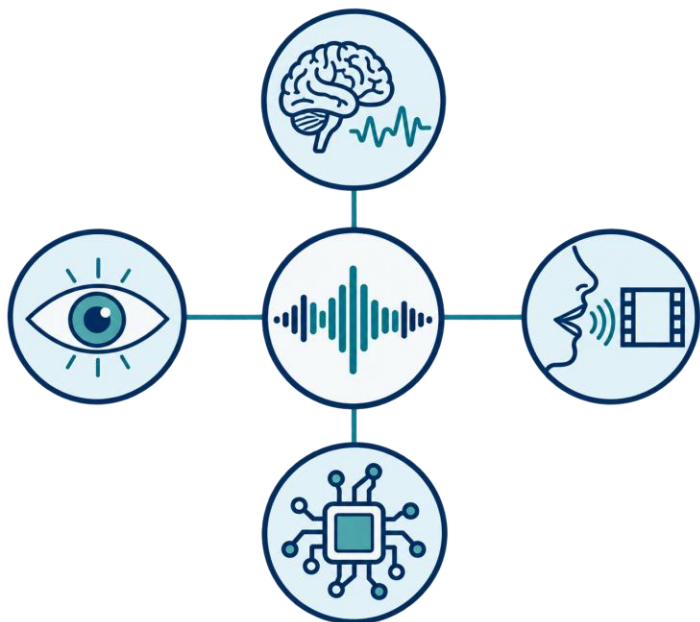
Issue	Why it matters	Examples
Sensor burden	Specialized hardware can limit real-world usage.	EMA, EPG, multichannel EEG
Speaker and session variability	Signals change with sensor placement and recording session.	Surface EMG, EMA, EEG
Cross-modal alignment	Modalities differ in sampling rate and physiological timing	Audio-visual and audio-physiological systems
Limited training data	Paired biological and acoustic datasets remain limited.	Articulatory, sEMG, and EEG datasets
Evaluation	Signal metrics do not fully describe listener benefit.	Quality, intelligibility, and listening effort



Despite **promising progress**, further research is needed to make biological signals **easier to use** and **fully realize their potential** in acoustic modeling.

Related Topics Not Covered in This Talk

Selected directions for further exploration



01 EEG- and eye-gaze-guided processing

Speech enhancement and target-speaker extraction guided by neural activity or visual attention [\[51–53, 64–66\]](#)

02 Speech generation from neural and visual signals

Speech synthesis from brain activity [\[67–70\]](#) or silent oral video [\[75\]](#), with related work using ultrasound and MRI observations [\[71–73\]](#)

03 Advanced architectures and learning objectives

Diffusion and flow matching, transformers, Mamba, and large language models [\[29, 74, 75, 77, 78\]](#)

04 Synthesizing articulatory data from acoustic or text

Model size, computational cost, latency, and implementation constraints [\[79–85\]](#)

 Selected references provide entry points for further reading.

References

- [1] J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in Proc. ICASSP, 2018.
- [2] Y. Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in Proc. ICLR, 2021.
- [3] S. Chen et al., “Neural codec language models are zero-shot text to speech synthesizers,” IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 705–718, 2025.
- [4] B. Sisman et al., “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 132–157, 2021.
- [5] C.-C. Hsu et al., “Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks,” in Proc. Interspeech, 2017.
- [6] K. Qian et al., “AutoVC: Zero-shot voice style transfer with only autoencoder loss,” in Proc. ICML, 2019.
- [7] M.-C. Yen et al., “Mandarin electrolaryngeal speech voice conversion with speech encoder loss learning and seq2seq modeling,” IEEE Internet of Things Magazine, vol. 8, no. 4, pp. 22–28, 2025.
- [8] S.-W. Fu et al., “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in Proc. ICML, 2019.
- [9] R. Chao et al., “An investigation of incorporating Mamba for speech enhancement,” in Proc. SLT, 2024.
- [10] Y. Luo et al., “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266, 2019.

- [11] D. Wang et al., “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [12] J. Chen et al., “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” in *Proc. Interspeech*, 2020.
- [13] T. Nakatani et al., “Speech dereverberation based on variance-normalized delayed linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.
- [14] B. Wu et al., “An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1289–1300, 2017.
- [15] A. W. Rix et al., “Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP*, 2001.
- [16] C. H. Taal et al., “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [17] C.-C. Lo et al., “MOSNet: Deep learning-based objective assessment for voice conversion,” in *Proc. Interspeech*, 2019.
- [18] E. Cooper et al., “A review on subjective and objective evaluation of synthetic speech,” *Acoustical Science and Technology*, vol. 45, no. 4, pp. 161–183, 2024.
- [19] T. Saeki et al., “UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022,” in *Proc. Interspeech*, 2022.
- [20] C. K. A. Reddy et al., “DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proc. ICASSP*, 2021.

- [21] C.-C. Sung et al., “EffortNet: A deep learning framework for objective assessment of speech enhancement technologies using EEG-based alpha oscillations,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 34, pp. 2795–2807, 2026.
- [22] J.-C. Hou et al., “Audio-visual speech enhancement using multimodal deep convolutional neural networks,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.
- [23] M. Sadeghi et al., “Audio-visual speech enhancement using conditional variational auto-encoders,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.
- [24] S.-Y. Chuang et al., “Improved lite audio-visual speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1345–1359, 2022.
- [25] V. A. Kalkhorani et al., “AV-CrossNet: An audiovisual complex spectral mapping network for speech separation by leveraging narrow- and cross-band modeling,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, no. 4, pp. 685–694, 2025.
- [26] H.-P. Liu et al., “Bone-conducted speech enhancement using deep denoising autoencoder,” *Speech Communication*, vol. 104, pp. 106–112, 2018.
- [27] C. Yu et al., “Time-domain multi-modal bone/air conducted speech enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.
- [28] H. Wang et al., “Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain,” in *Proc. ICASSP*, 2022.
- [29] Y. Sui et al., “TRAMBA: A hybrid transformer and Mamba architecture for practical audio and bone conduction speech super resolution and enhancement on mobile and wearable platforms,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, Art. no. 205, 2024.
- [30] Y.-W. Chen et al., “EMA2S: An end-to-end multimodal articulatory-to-speech system,” in *Proc. ISCAS*, 2021.

- [31] Y.-W. Chen et al., “A study of incorporating articulatory movement information in speech enhancement,” in Proc. EUSIPCO, 2021.
- [32] L.-C. Chen et al., “EPG2S: Speech generation and speech enhancement based on electropalatography and audio signals using multimodal learning,” IEEE Signal Processing Letters, vol. 29, pp. 2582–2586, 2022.
- [33] N. Kimura et al., “SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography,” in Proc. CHI, 2022.
- [34] T. Arias-Vergara et al., “Contrastive learning approach for assessment of phonological precision in patients with tongue cancer using MRI data,” in Proc. Interspeech, 2024.
- [35] B. Denby et al., “Silent speech interfaces,” Speech Communication, vol. 52, no. 4, pp. 270–287, 2010.
- [36] G. K. Anumanchipalli et al., “Speech synthesis from neural decoding of spoken sentences,” Nature, vol. 568, no. 7753, pp. 493–498, 2019.
- [37] N. Mesgarani et al., “Selective cortical representation of attended speaker in multi-talker speech perception,” Nature, vol. 485, no. 7397, pp. 233–236, 2012.
- [38] Y. Yang et al., “UTI-LLM: A personalized articulatory-speech therapy assistance system based on multimodal large language model,” in Proc. ICASSP, 2026.
- [39] G. Sanguedolce et al., “Physiologically-informed feature analysis of acquired speech disorders for stroke assessment,” in Proc. Interspeech, 2025.
- [40] P. Wu et al., “Deep speech synthesis from MRI-based articulatory representations,” in Proc. Interspeech, 2023.

- [41] N. Shah et al., “MRI2Speech: Speech synthesis from articulatory movements recorded by real-time MRI,” in Proc. ICASSP, 2025.
- [42] Y. Otani et al., “Speech synthesis from articulatory movements recorded by real-time MRI,” in Proc. Interspeech, 2023.
- [43] L. Tóth et al., “Adaptation of tongue ultrasound-based silent speech interfaces using spatial transformer networks,” in Proc. Interspeech, 2023.
- [44] S. Hu et al., “Exploiting cross-domain and cross-lingual ultrasound tongue imaging features for elderly and dysarthric speech recognition,” in Proc. Interspeech, 2023.
- [45] K.-C. Wang et al., “EMGSE: Acoustic/EMG fusion for multimodal speech enhancement,” in Proc. ICASSP, 2022.
- [46] K. Scheck et al., “DiffMV-ETS: Diffusion-based multi-voice electromyography-to-speech conversion using speaker-independent speech training targets,” in Proc. Interspeech, 2025.
- [47] J. Lee et al., “Articulatory feature prediction from surface EMG during speech production,” arXiv preprint arXiv:2505.13814, 2025.
- [48] D. Gaddy et al., “Digital voicing of silent speech,” in Proc. EMNLP, 2020.
- [49] K. Daoudi et al., “Electroglottography for the assessment of dysphonia in Parkinson’s disease and multiple system atrophy,” in Proc. Interspeech, 2024.
- [50] Y. Geng et al., “EEG-based voice conversion: Hearing the voice of your brain,” in Proc. Interspeech, 2025.

- [51] C. Jehn et al., “Attention decoding at the cocktail party: Preserved in hearing aid users, reduced in cochlear implant users,” *NeuroImage*, vol. 328, Art. no. 121771, 2026.
- [52] J. A. O’Sullivan et al., “Attentional selection in a cocktail party environment can be decoded from single-trial EEG,” *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.
- [53] V. Choudhari et al., “Real-time brain-controlled selective hearing enhances speech perception in multi-talker environments,” *Nature Neuroscience*, 2026.
- [54] K. T. Littlejohn et al., “A streaming brain-to-voice neuroprosthesis to restore naturalistic communication,” *Nature Neuroscience*, vol. 28, no. 4, pp. 902–912, 2025.
- [55] G. Keidser et al., “The NAL-NL2 prescription procedure,” *Audiology Research*, vol. 1, no. 1, Art. no. e24, 2011.
- [56] S. Ahmed et al., “NeuroAMP: A novel end-to-end general purpose deep neural amplifier for personalized hearing aids,” *IEEE Transactions on Artificial Intelligence*, vol. 7, no. 3, pp. 1610–1625, 2026.
- [57] C. Tang et al., “Sensing technologies for silent speech interfaces,” *Nature Sensors*, vol. 1, no. 1, pp. 16–26, 2026.
- [58] H. R. Bosker, “Putting Laurel and Yanny in context,” *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. EL503–EL508, 2018.
- [59] S. Chen et al., “Context-aware audio-visual speech enhancement based on neuro-fuzzy modeling and user preference learning,” *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 10, pp. 5400–5412, 2024.
- [60] H. Fletcher et al., “The frequency—sensitivity of normal ears,” *Physical Review*, vol. 19, no. 6, pp. 553–565, 1922.

- [61] J. M. Kates et al., “The hearing-aid speech perception index (HASPI),” *Speech Communication*, vol. 65, pp. 75–93, 2014.
- [62] J. Kates et al., “The hearing-aid speech quality index (HASQI) version 2,” *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.
- [63] H.-T. Chiang et al., “Multi-objective non-intrusive hearing-aid speech assessment model,” *The Journal of the Acoustical Society of America*, vol. 156, no. 5, pp. 3574–3587, 2024.
- [64] Z. Pan et al., “NeuroHeed: Neuro-steered speaker extraction using EEG signals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4456–4470, 2024.
- [65] X. Wang et al., “EEG and eye-tracking driven dynamic target speaker extraction with spontaneous attention switching,” in *Proc. ICASSP*, 2026.
- [66] H.-C. Yang et al., “Tracking listener attention: Gaze-guided audio-visual speech enhancement framework,” in *Proc. ICASSP*, 2026.
- [67] S. S. Chandra et al., “Brain2Speech-Net: Intelligible, real-time brain-to-speech synthesis without text decoding,” *arXiv preprint arXiv:2609.04455*, 2026.
- [68] M. Angrick et al., “Speech synthesis from ECoG using densely connected 3D convolutional neural networks,” *Journal of Neural Engineering*, vol. 16, no. 3, Art. no. 036019, 2019.
- [69] M. Wairagkar et al., “An instantaneous voice-synthesis neuroprosthesis,” *Nature*, vol. 644, no. 8075, pp. 145–152, 2025.
- [70] J. Kwon et al., “Direct speech synthesis from non-invasive, neuromagnetic signals,” in *Proc. Interspeech*, 2024.

- [71] R. Trencsényi et al., “Ultrasound- and MRI-based speech synthesis applying neural networks,” in Proc. ICCV, 2024.
- [72] X. Liu et al., “Tagged-MRI sequence to audio synthesis via self residual attention guided heterogeneous translator,” in Proc. MICCAI, 2022.
- [73] Y. Yang et al., “An audio-ultrasound synchronized database of tongue movement for Mandarin speech,” Scientific Data, vol. 12, Art. no. 607, 2025.
- [74] W. Gao et al., “NeuroSonic: Conditional flow matching for EEG-to-speech reconstruction,” in Proc. MICCAI, 2026.
- [75] R.-C. Zheng et al., “Speech reconstruction from silent lip and tongue articulation by diffusion models and text-guided pseudo target generation,” in Proc. ACM MM, 2024.
- [76] M. Angrick et al., “Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity,” Communications Biology, vol. 4, no. 1, Art. no. 1055, 2021.
- [77] N. Sivasubramaniam, “From silent signals to natural language: A dual-stage transformer-LLM approach,” arXiv preprint arXiv:2509.04507, 2025.
- [78] Z. Ren et al., “Diff-ETS: Learning a diffusion probabilistic model for electromyography-to-speech conversion,” in Proc. EMBC, 2024.
- [79] A. S. Shahrehabaki et al., "Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 135-147.
- [80] A. S. Shahrehabaki et al., "Raw speech-to-articulatory inversion by temporal filtering and decimation." Proc. Interspeech 2021.

- [81] T. Toda et al., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication* 50.3 (2008): 215-227.
- [82] W.-J. Chung et al., "Speaker-independent acoustic-to-articulatory inversion through multi-channel attention discriminator." in *Proc. 2024*.
- [83] C. McGhee et al. "Training Articulatory Inversion Models for Interspeaker Consistency." in *Proc. 2025*.
- [84] H. K. Kim et al., "ArtBoost: Synthetic articulatory data augmentation for acoustic-to-articulatory inversion." in *Proc. Interspeech 2026*.
- [85] X. Yin et al., "STArK: towards synthesizing articulatory kinematics from text," in *Proc. Interspeech 2026*.

Thank You

Questions & Discussion

Bio-Informed Speech Processing, Modeling, and Generation

Yu Tsao

Research Fellow (Professor)

Academia Sinica, Taiwan

yu.tsao@citi.sinica.edu.tw

<https://www.citi.sinica.edu.tw/pages/yu.tsao/>

<http://bio-asplab.citi.sinica.edu.tw/>