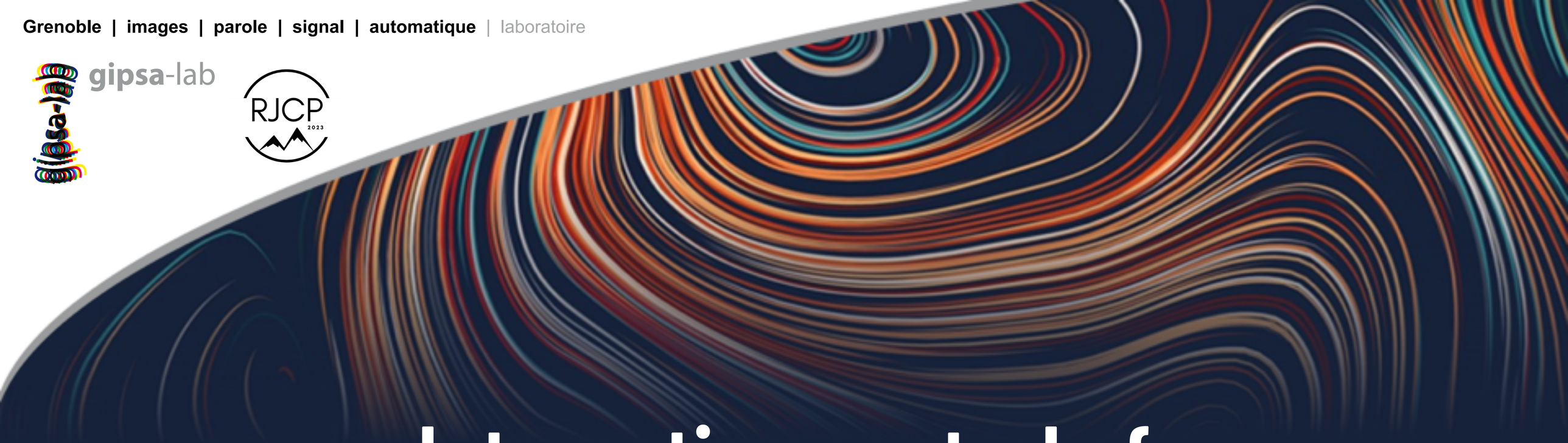
Grenoble | images | parole | signal | automatique | laboratoire

gipsa-lab



Interactive control of Expressive Speech Synthesis

Olivier PERROTIN

30.11.2023

Rencontre des Jeunes Chercheurs en Parole 2023

Parcours

Ecole d'ingénieur Grenoble INP - Phelma, *Spécialité traitement du signal (SICOM)* — TS

2012 Thèse au LIMSI, Université Paris-Sud (dir. C. d'Alessandro) — IHM
« Singing with hands : chironomic interfaces for digital musical instruments »

2015 Post-doctorat au LIMSI, CNRS — TS
« Vocal effort control in Singing Synthesis » - 16 months

2017 Post-doctorat à University of Kent, School of Computing, Royaume-Uni — TS
« Whisper-to-Speech conversion » - 16 months

2018 Chargé de recherche CNRS, GIPSA-lab — TS IHM
« Interactive control of expressive speech synthesis »

2023

Traitement du signal, Informatique

Interaction homme-machine

Contrôle gestuel
Contrôle temps-réel
Interfaces expressives

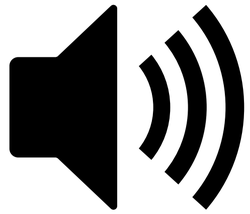
⟷

Synthèse vocale
Traitement de signaux de parole
Modèles de production de la parole

Interactive control of expressive speech synthesis

- An overview of **speech synthesis**
- Analysis-synthesis of **expressive speech**
- **Interactive control** of synthesis

Generating voices

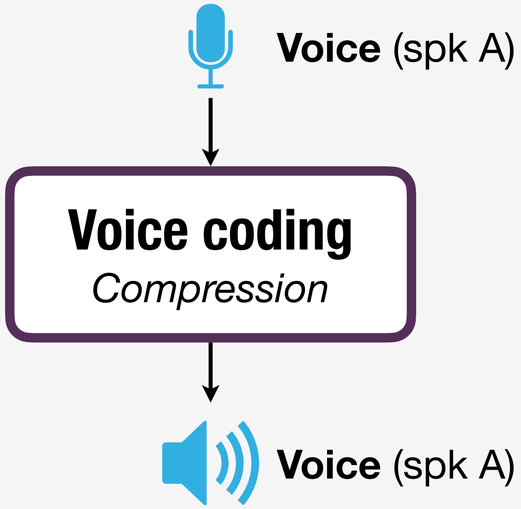


For whom?

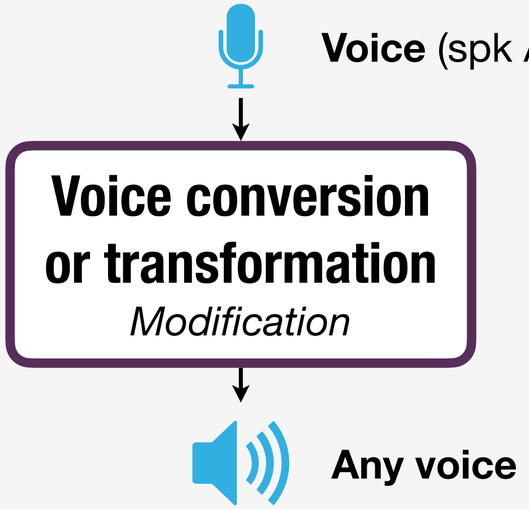
To say what?

Voice generation

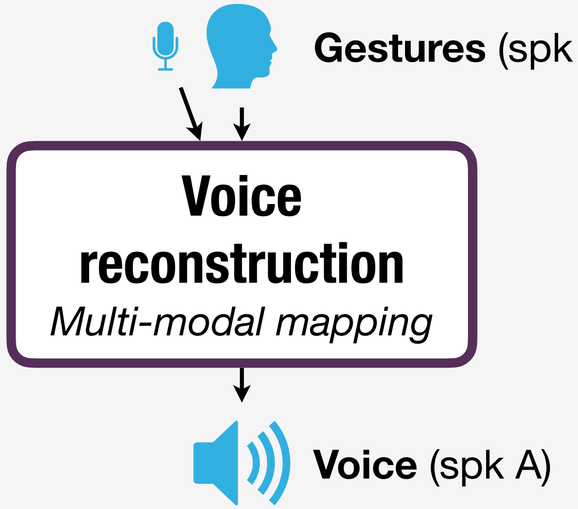
Applications



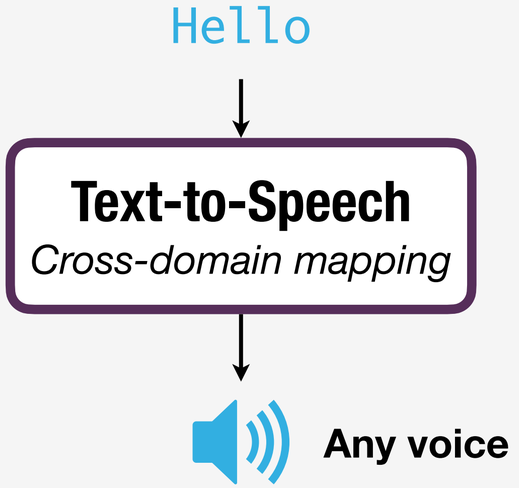
Communication
(phone, video calls, etc.)



Change speaker ID,
Change voice features (medical)

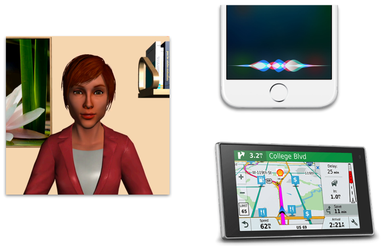


Enhance voice (medical)



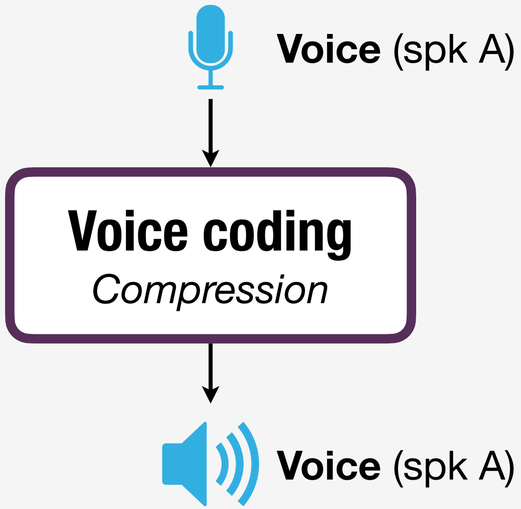
Read a text
(conversational agents)

zoom

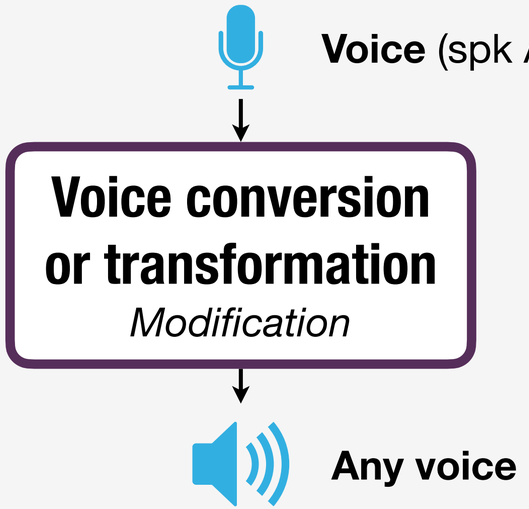


Voice generation

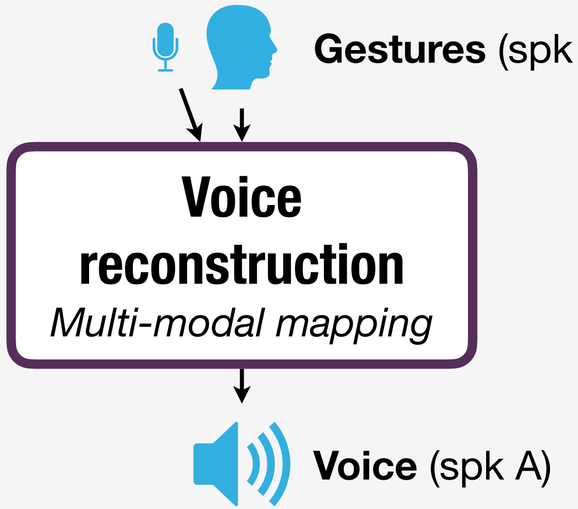
Applications



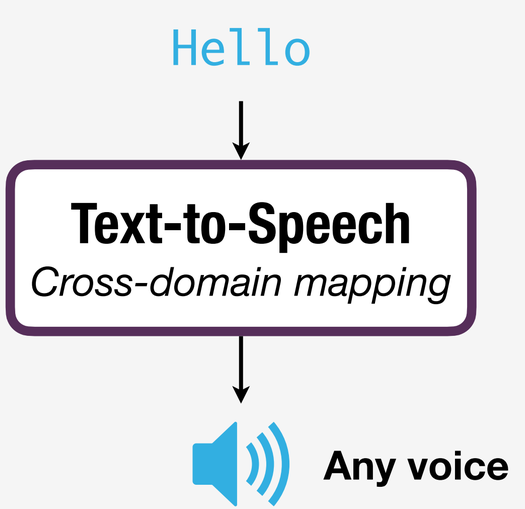
Communication
(phone, video calls, etc.)



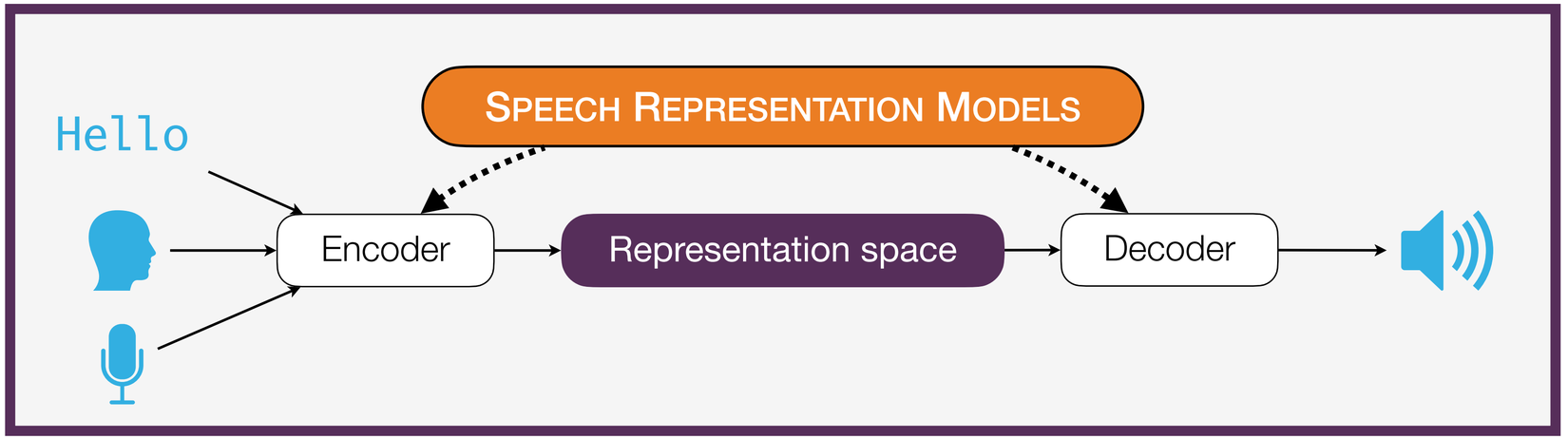
Change speaker ID,
Change voice features (medical)



Enhance voice (medical)

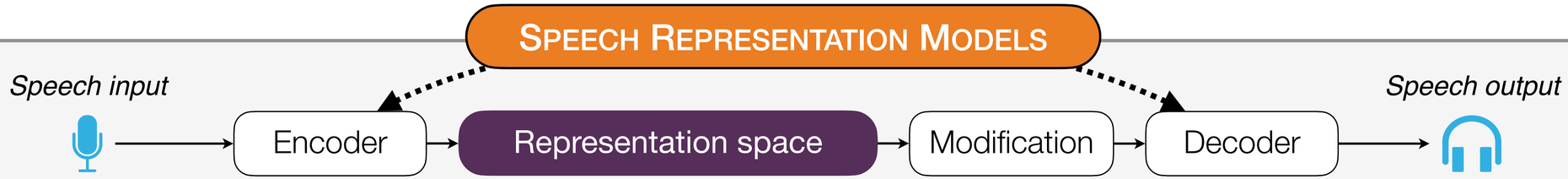


Read a text
(conversational agents)



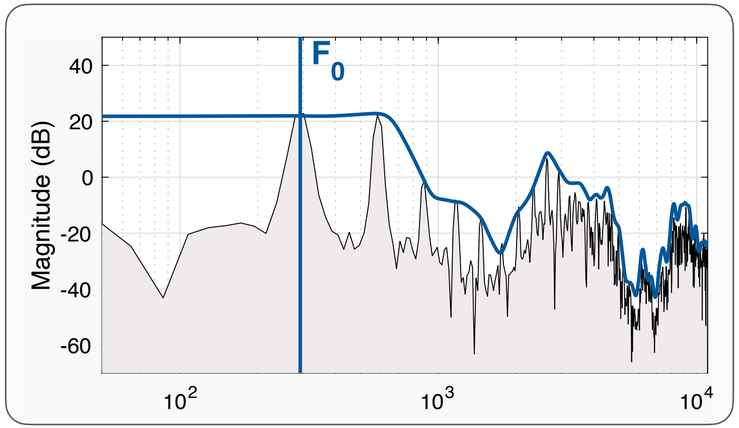
Voice generation

Speech to speech



- Speech input and output ➡ symmetric encoding/ decoding process

- Two kinds of encoding
  - Signal-based (F0, spectral envelope, etc.)
  - Source-filter based (Vocal tract and glottal flow parameters)

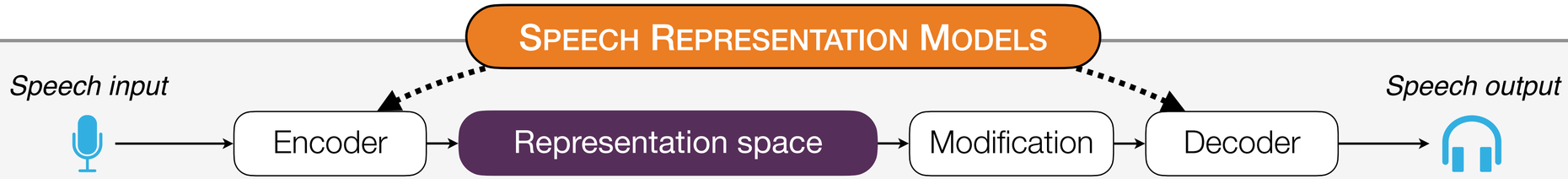


(References are non-exhaustive)

Compression	**LPC** Linear Predictive coding	*Makhoul J. (1975), Proc. of the IEEE, 63(4), pp. 561–580.*
	MLSA Mel log spectrum approximation	*Imai S. et al. (1983), Electronics and Communications in Japan, 66(2), pp. 10–18.*
	CELP Code-excited linear prediction	*Schroder M. et al. (1985), Proc. ICASSP, pp. 937—940.*
High-quality	**HNM** Harmonic plus Noise Model	*Stylianou Y. (1996),. PhD Thesis, Ecole Normale Supérieure des Télécommunications*
	STRAIGHT Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum	*Kawahara H. et al. (1999),. Speech Communication, vol. 27, no. 3-4, pp. 187–207.*
	WORLD	*Morise M. et al. (2016),. IEICE Trans. on Information and Systems, pp. 1877–1884.*
Glottal model	**IAIF** Iterative Adaptive Inverse Filtering	*Alku P. (1992), Speech Comm., 11(2–3), pp. 109–118.*
	SVLN Separation of the Vocal tract with the Liljencrants–fant model plus Noise	*Degottex G. et al. (2013), Speech Comm., 55(2), pp. 278—294.*

Voice generation

Speech to speech



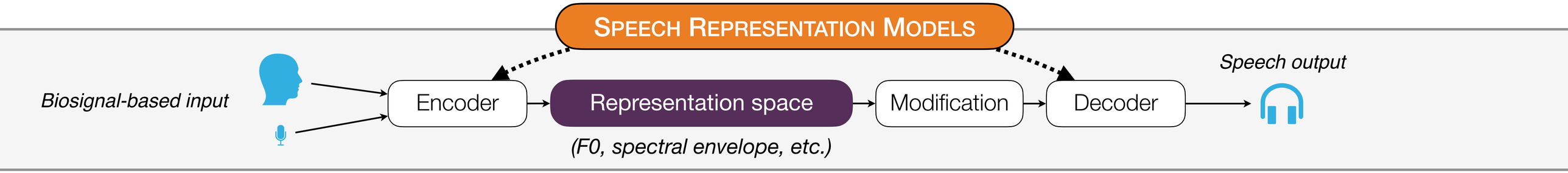
- Speech input and output ➡ symmetric encoding/ decoding process
- Two kinds of encoding
  - Signal-based (F0, spectral envelope, etc.)
  - Source-filter based (Vocal tract and glottal flow parameters)
- Reduced set of parameters ➡ Voice coding
- Modification of parameters
  - Change speaker ID ➡ Voice conversion
  - Change speaker voice ➡ Voice transformation

(References are non-exhaustive)

Coding	**LPC** Linear Predictive coding	*Makhoul J. (1975), Proc. of the IEEE, 63(4), pp. 561–580.*
Coding	**MLSA** Mel log spectrum approximation	*Imai S. et al. (1983), Electronics and Communications in Japan, 66(2), pp. 10–18.*
Coding	**CELP** Code-excited linear prediction	*Schroder M. et al. (1985), Proc. ICASSP, pp. 937—940.*
Conversion	**Statistical** GMM	*Toda T. (2007), Trans. IEEE TASLP, 15(8), pp. 2222–2235.*
Conversion	**VC Challenge** GMM / some DNN	*Toda T. et al. (2016), Proc. Interspeech, pp. 1632–1636.*
Transformation	**SVLN**	*Degottex G. et al. (2013), Speech Comm., 55(2), pp. 278—294.*
Transformation	**Whisper-to-Speech** GMM- / Rule- based	*Toda T. et al. (2012), Trans. IEEE TASLP, 20(9), pp. 2505–2517.* *Perrotin O. et al. (2020),. IEEE TASLP, 28, pp. 889–900.*

Voice generation

Speech to speech



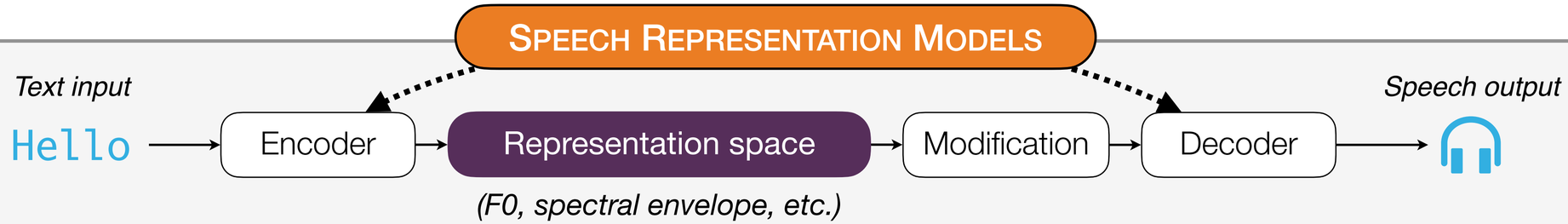
- Input specific encoding methods
  - Articulatory-to-acoustic mapping
  - Brain-to-acoustic mapping
  - Muscle activity-to-acoustic mapping

➡ Similar dynamics between input and output (causal effect)

Biosignal-based Litt. review	*Schultz T. et al. (2017),. IEEE TASLP, 25(12), pp. 2257–2271.*
Vocal tract lab Articulatory synthesis	*Birkholz P. (2013),. PLoS ONE, 8(4)*
Silent Speech HMM-based	*Hueber T. et al. (2016),. Comp. Speech Lang., 36, pp. 274—293.*

Voice generation

Speech to speech



- Input specific encoding methods
  - Articulatory-to-acoustic mapping
  - Brain-to-acoustic mapping
  - Muscle activity-to-acoustic mapping

  ➡ Similar dynamics between input and output (causal effect)

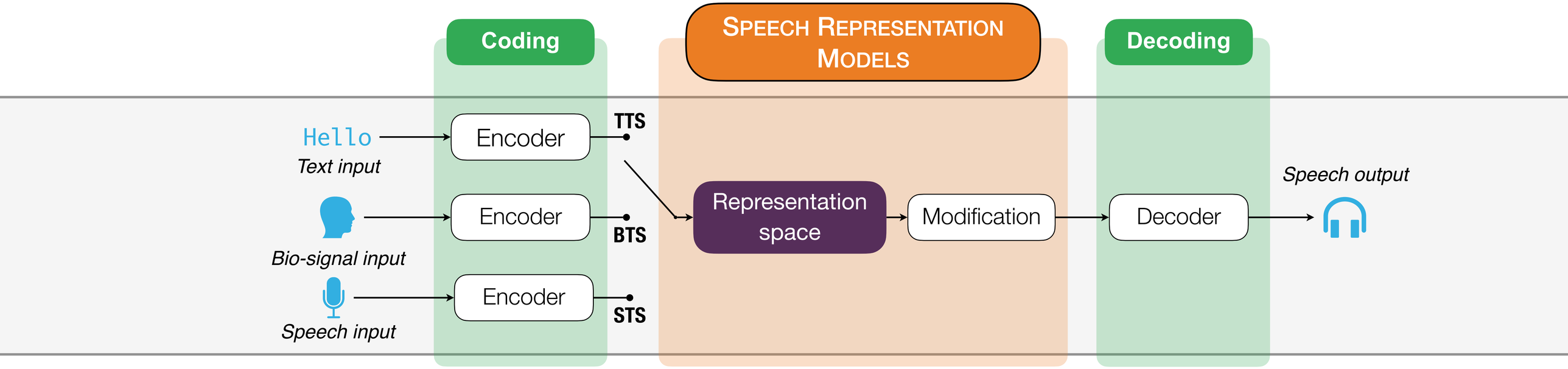
- Natural language processing methods
  - Grapheme to phoneme conversion
  - Part of speech tagging
  - Etc.

  ➡ Different dynamics between input and output (need to predict duration)

Biosignal-based Litt. review	*Schultz T. et al. (2017),. IEEE TASLP, 25(12), pp. 2257–2271.*
Vocal tract lab Articulatory synthesis	*Birkholz P. (2013),. PLoS ONE, 8(4)*
Silent Speech HMM-based	*Hueber T. et al. (2016),. Comp. Speech Lang., 36, pp. 274—293.*

Rule-based Formant synthesis	*Klatt D. et al. (1980),. J. Acoust. Soc. Am., 67(3), pp. 971—995.*
Concatenative Speech samples	*Hunt A. et al. (1996), Proc. ICASSP, pp. 373–376.*
Statistical Vocoder	*Tokuda K. et al. (2013),. Proc. of the IEEE, 101(5), pp. 1234–1252.*
Blizzard Challenge 2005 / 2013	*Black A. et al. (2005),. Proc. of Interspeech, pp. 77–80.* *King S. et al. (2013),. Evaluation of TTS systems.*

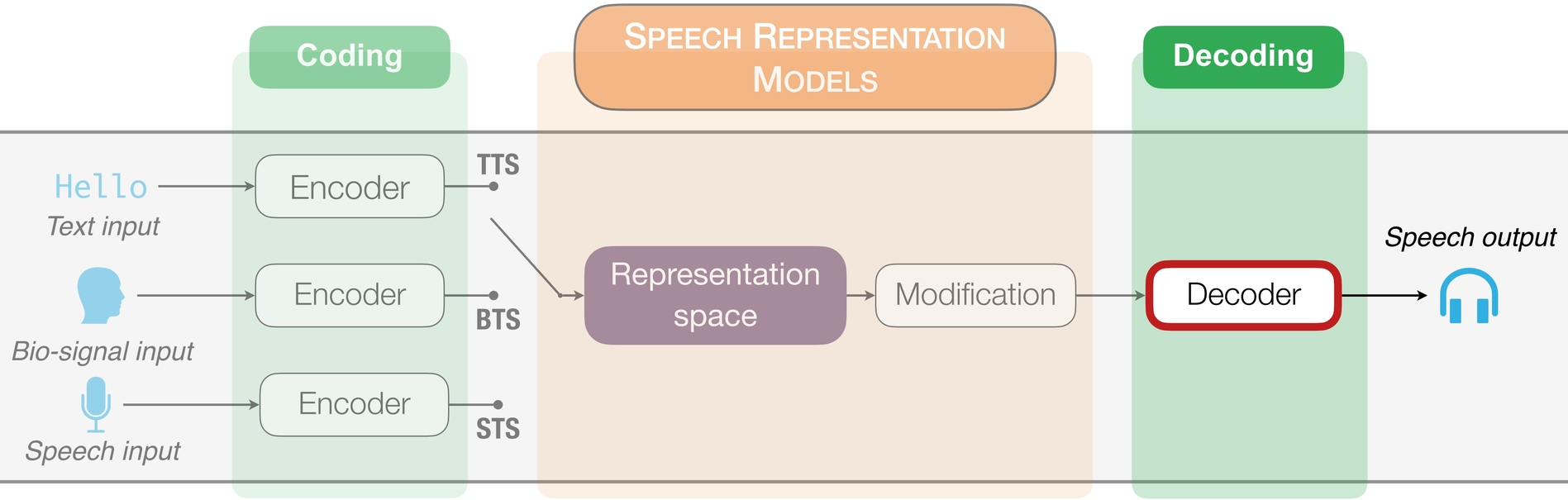
Voice generation



- All literature presented so far was < 2017
- Now everything is 'neural'
- How does this change the voice generation process?

Voice generation

Neural vocoders



➡ **Neural decoders**
(wrongly called Neural Vocoders)

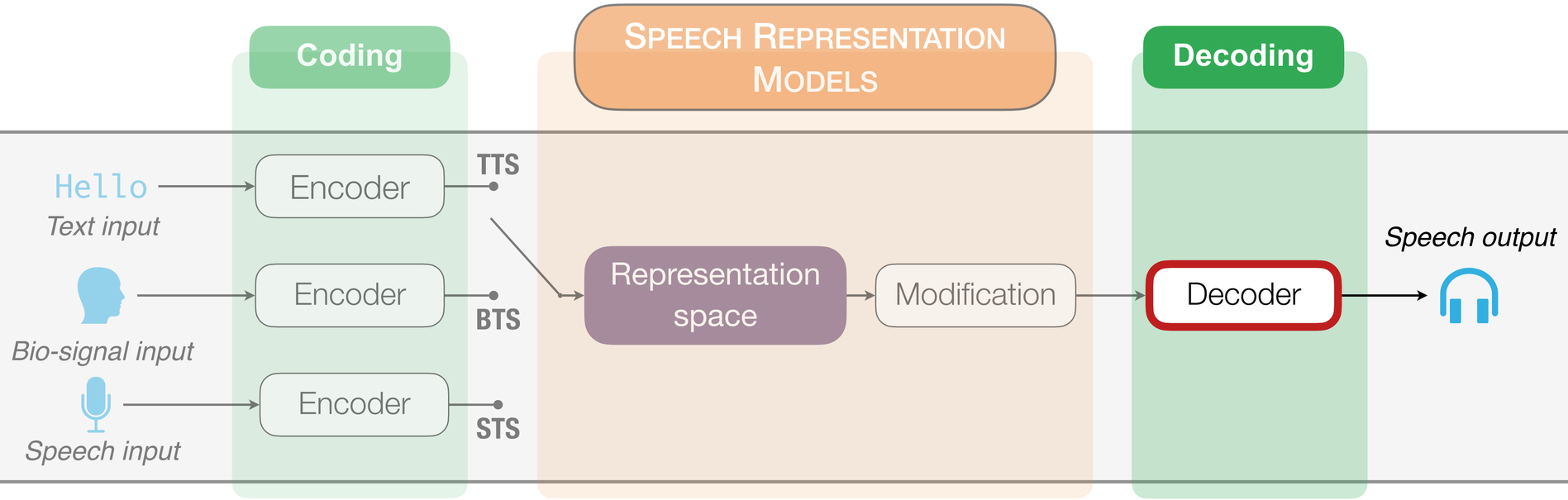
- From a speech representation to a waveform

(References are non-exhaustive)

Text-to-Speech (TTS)		**WaveNet**	*van den Oord A. et al. (2016), arXiv, vol. abs/1609.03499*
Speech-to-Speech (STS)	*Conversion*	**Voice Conversion Challenge 2018** Neural generation	*Lorenzo-Trueba L. et al. (2018), Proc. Odyssey, pp. 195–202.*
	Coding	**Lyra** Neural vocoder (WaveNet)	*Kleijn B. et al. (2018), Proc. ICASSP, pp. 676—680* *Kleijn B. et al. (2021), Proc. ICASSP, pp. 6478—6482*

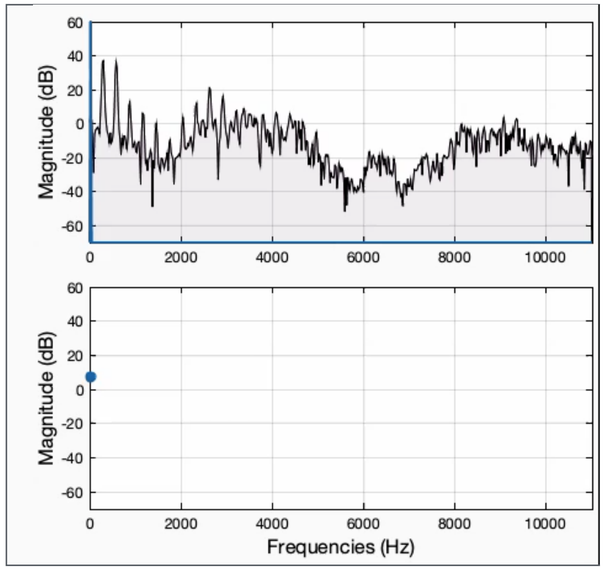
Voice generation

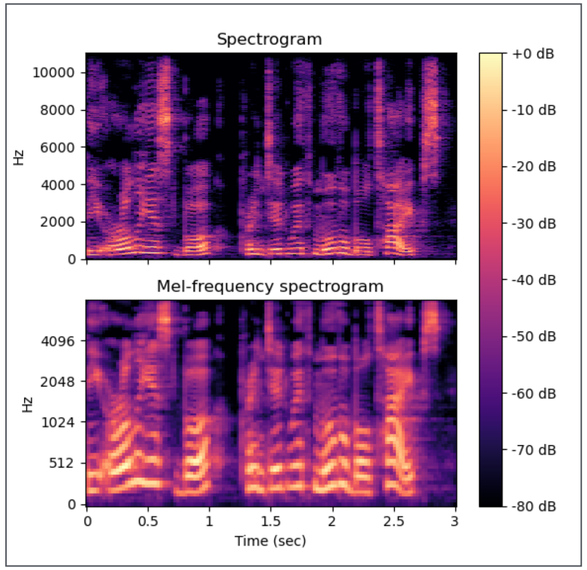
Neural vocoders



➡ **Neural decoders**
(wrongly called Neural Vocoders)

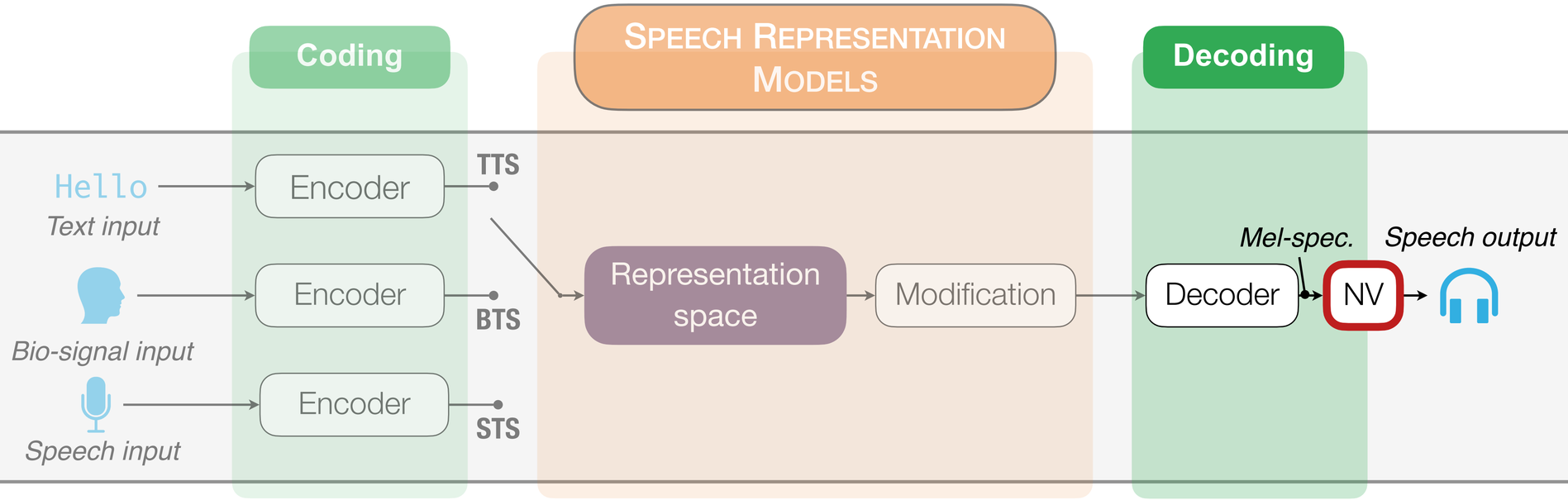
- Fom mel-spectrogram to waveform





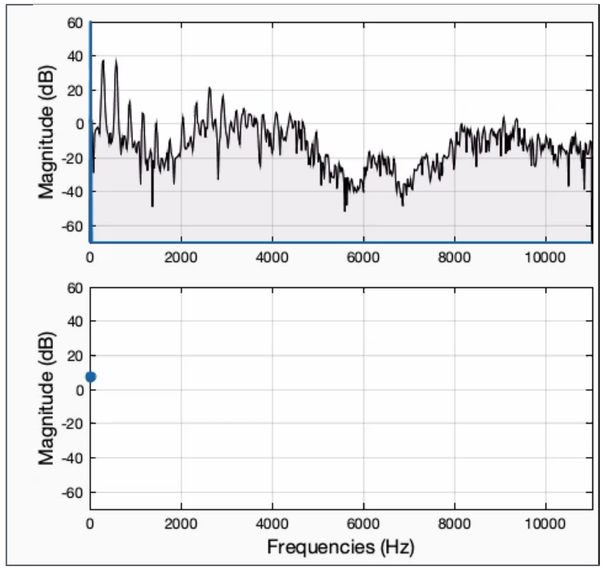
Voice generation

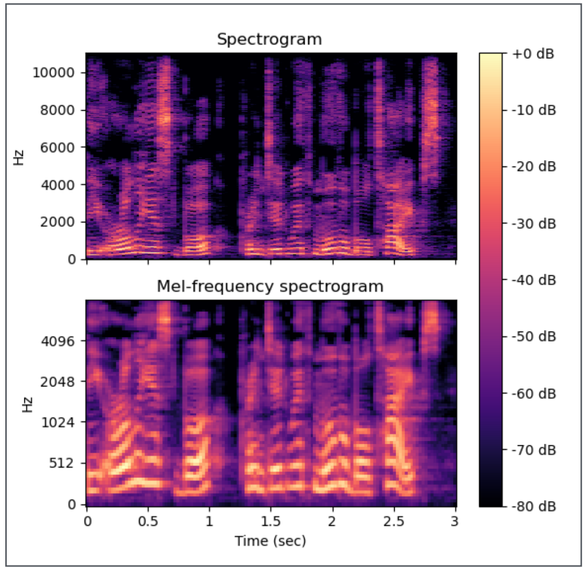
Neural vocoders



➡ **Neural decoders**
(wrongly called Neural Vocoders)

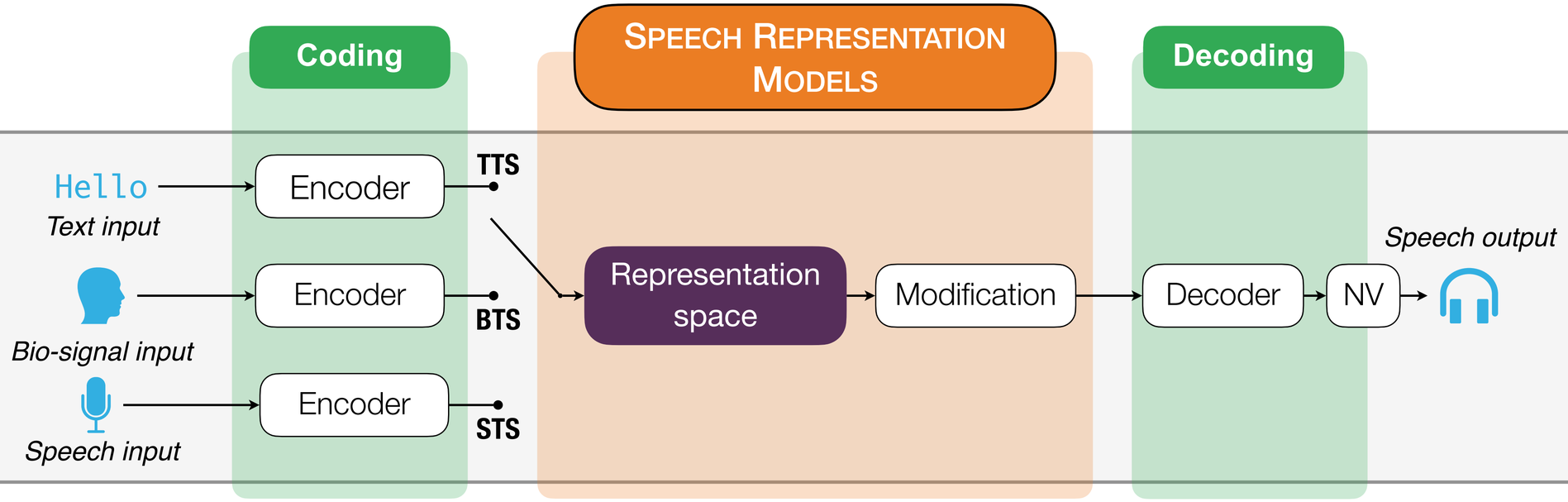
- Fom mel-spectrogram to waveform





Voice generation

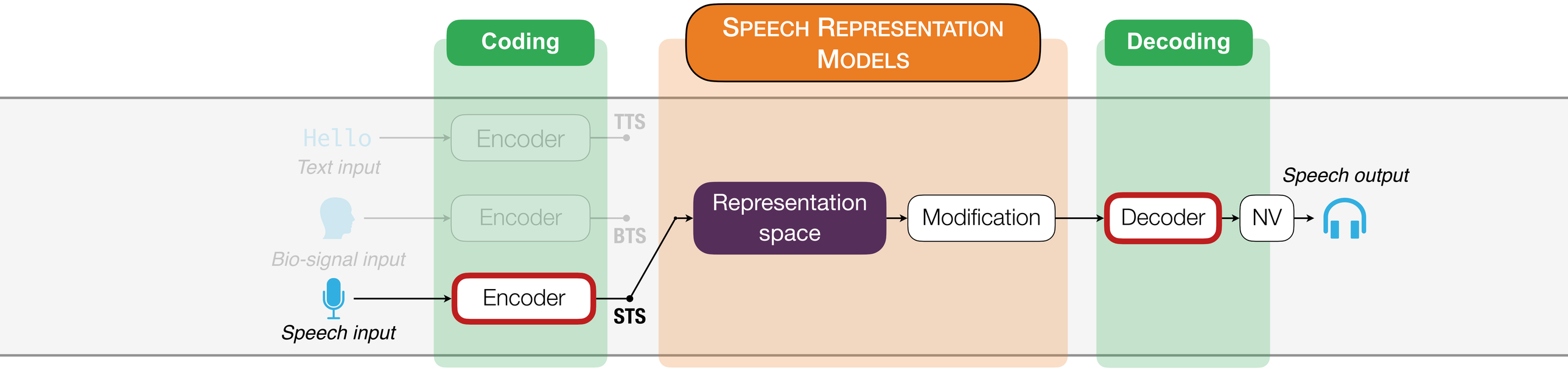
Neural Speech to speech



- Neural encoder-decoder
  - A neural network for the encoder and the decoder
  - Few constrains on the representation space, often considered as opaque

Voice generation

Neural Speech to speech



- Auto-encoders (unsupervised)
  - Low-dimensional representation space
  - Goal: Reconstruct input

(References are non-exhaustive)

Voice Conversion
From statistical to neural

Mohammadi S. H. et al. (2017), Speech Comm., 88(C), pp. 65–82.

Sisman B. et al. (2021), Trans. IEEE TASLP, 29, pp. 132–157.

Voice transformation
VAE / VQ-VAE

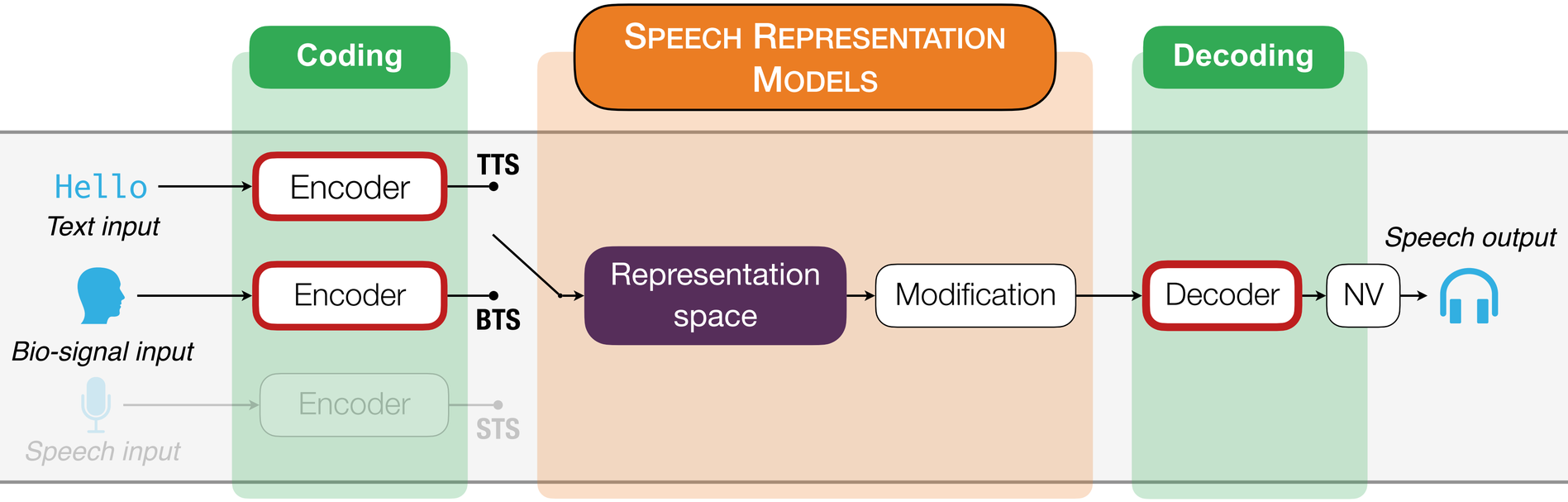
Hsu W.-N.. et al. (2019), ICLR.

Williams J. et al. (2021),. ISCA SSW, pp. 124—129.

Sadock S. et al. (2023), Speech Comm., 148, pp. 53—65.

Voice generation

Neural Speech to speech

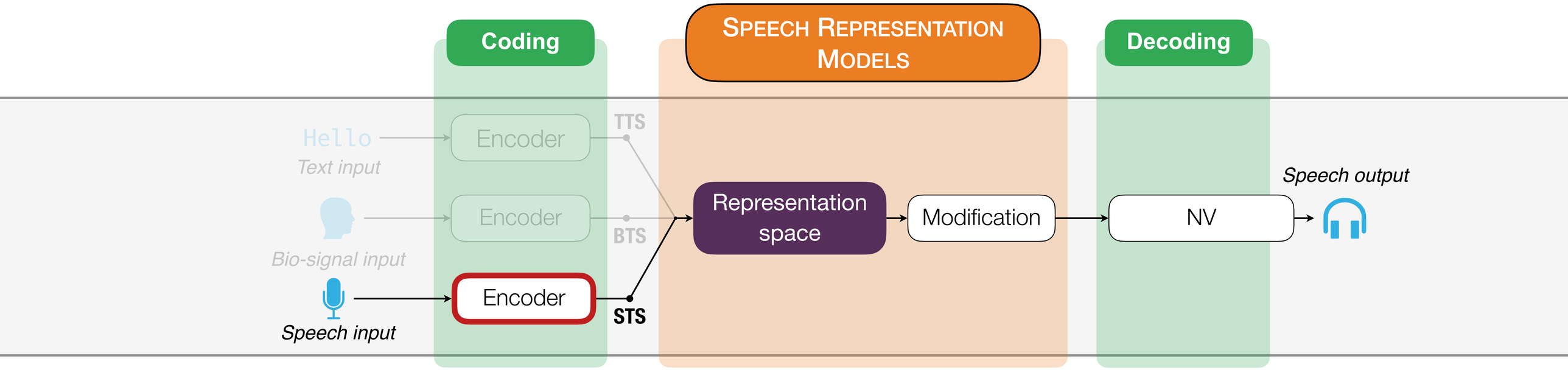


- Sequence-to-sequence (supervised)
  - Need for parallel data
  - Goal: predict audio from input

(References are non-exhaustive)

Autoregressive / Parallel Tacotron / FastSpeech	*Shen J. et al. (2018), Proc. ICASSP, pp. 4779—4783.* *Ren Y. et al. (2021), ICLR.*
Style modelling GST / VAE	*Wang Y. et al. (2018), Proc. ICML, 80, pp. 5180—5189.* *Zhang Y.-J.. et al. (2019),. Proc. ICASSP, pp. 6945—6949*
Blizzard Challenge 2023	*Perrotin O. et al. (2023),. Proc. of Blizzard Challenge, pp. 1–27.*
Ultrasound-to-speech Sequence-to-sequence	*Zhang J.-X.. et al. (2021),. AAAI Conf. on Artificial Intelligence, 35(16), pp. 14402—14410.*

Voice generation



- Self-supervised encoders
  - Inspired from large language models (par ex. GPT)
  - Learn audio representations (masking)
  - Need for a separate decoder

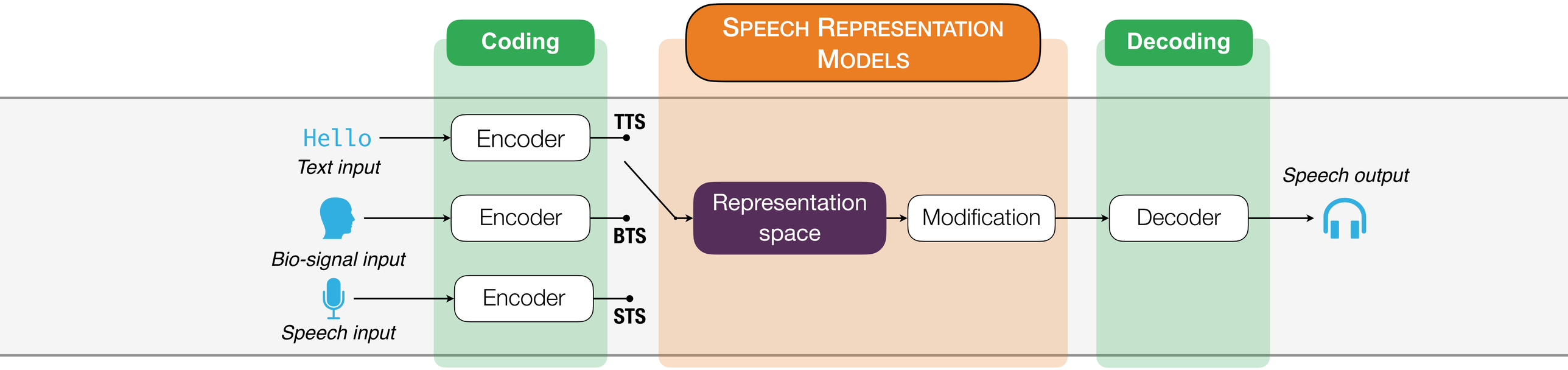
➡ back to neural vocoders with any input

(References are non-exhaustive)

Audio SSL CPC / wav2vec / HuBERT	*van den Oord A. et al. (2018), arXiv, vol. abs/1807.03748.* *Baevski A. et al. (2020), NeurIPS, 33, pp. 12449—12460.* *Hsu W.-H.. et al. (2021), IEEE TASLP, 29, pp. 3451—3460.*
GSLM (HubERT encoder, HiFiGAN decoder)	*Lakhotia K. et al. (2021), Trans. Association for Comp. Linguistics, 9, pp. 1336—1354* *Polyak A. et al. (2021), Proc. ICASSP, pp. 3615—3619*

Voice generation

Analysis-synthesis process



Signal-based

- Small model (10-100 parameters)
  - Explicit model (acoustically / signal / physiologically informed)
  - Generally fast to compute
- Limited modelling power
  - Light modelling of covariations
  - Not so high-quality synthesis

Neural-based

- Large model (M parameters)
  - Implicit model
  - Generally heavy to compute
- High modelling power
  - Modelling of covariations
  - High-quality synthesis

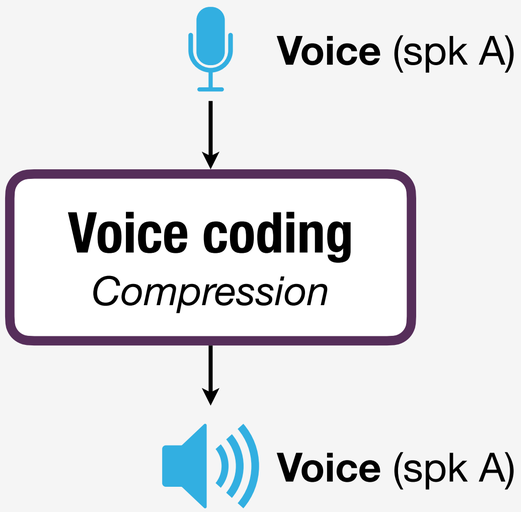
Do these voices sound ok?

- **Who** seems to speak? *Naturalness*
- **What** is being said? *Intelligibility*
- **How** is it said? *Expressivity*

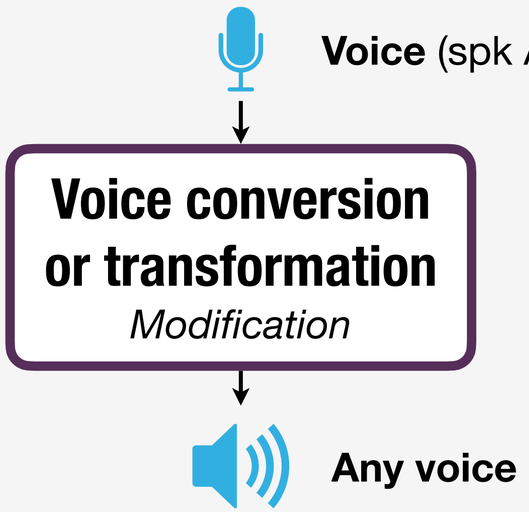
Blizzard Challenge 2023 *Perrotin O. et al. (2023),. Proc. of Blizzard Challenge, pp. 1–27.*

Voice generation

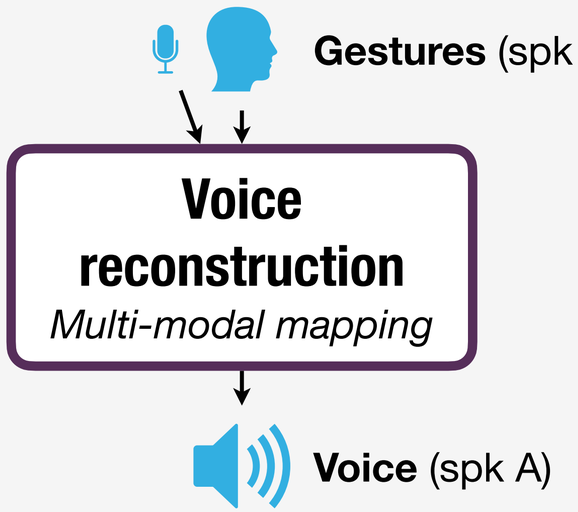
Expressive speech synthesis



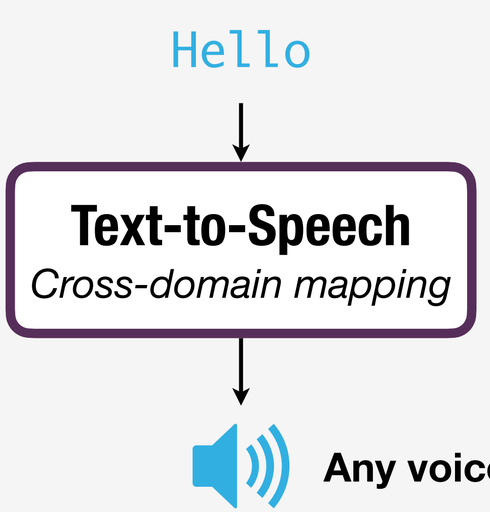
Communication (phone, video calls, etc.)



Change speaker ID, Change voice features (medical)



Enhance voice (medical)



Read a text (conversational agents)

- Naturalness

✓	✓	On its way	✓

- Intelligibilty

✓	✓	On its way	✓

- Expressivity

In the input	On its way	On its way	On its way

Interactive control of expressive speech synthesis

How to **generate variations of expressivity** within the representation space?

- Axis 1: Analysis-synthesis of **expressive speech**
  - Signal-based models : encoding / decoding and applications ➡ **Voice reconstruction**
  - Neural-based models : study of representation spaces ➡ **Voice transformation / Text-to-speech**

How to **control** the synthesiser to generate correct expressivity at the right time?

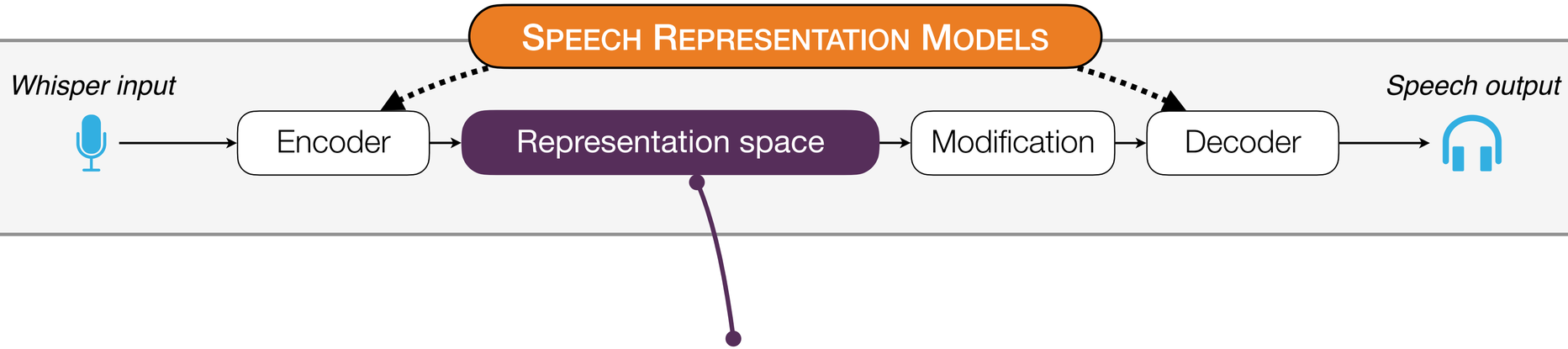
- Axis 2: **Interactive control** of synthesis ➡ **Voice transformation**
  - Explicit control of F0 in singing
  - From implicit to explicit control of F0 in speaking
  - Towards a co-adaptation between human and machine learning of control mapping

Analysis-synthesis of expressive speech

- Signal-based models : encoding / decoding and applications
- Neural-based models : study of representation spaces

Glottal model for analysis and synthesis

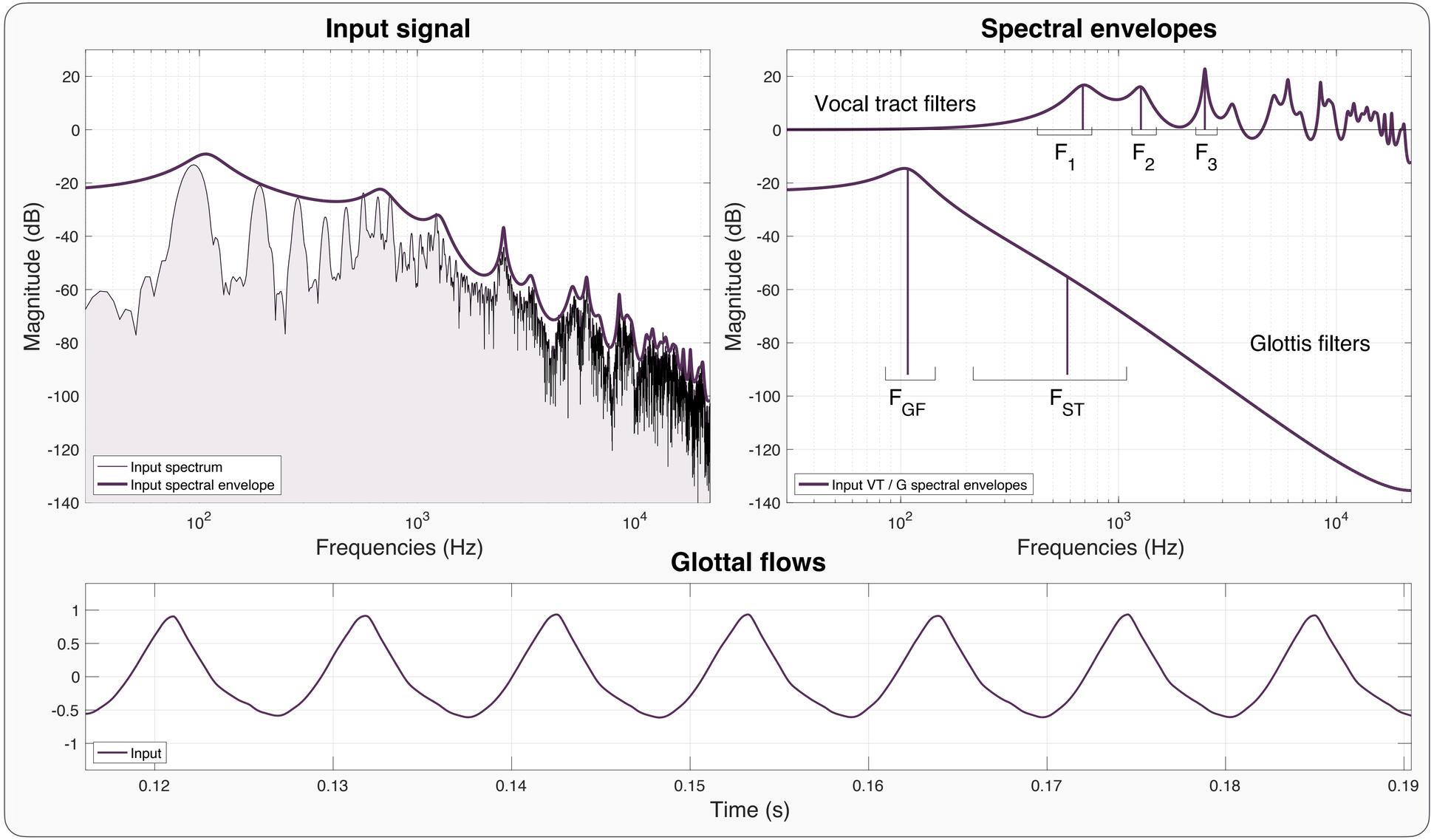
GFM-Vocoder



Source-filter model

Glottal model for analysis and synthesis

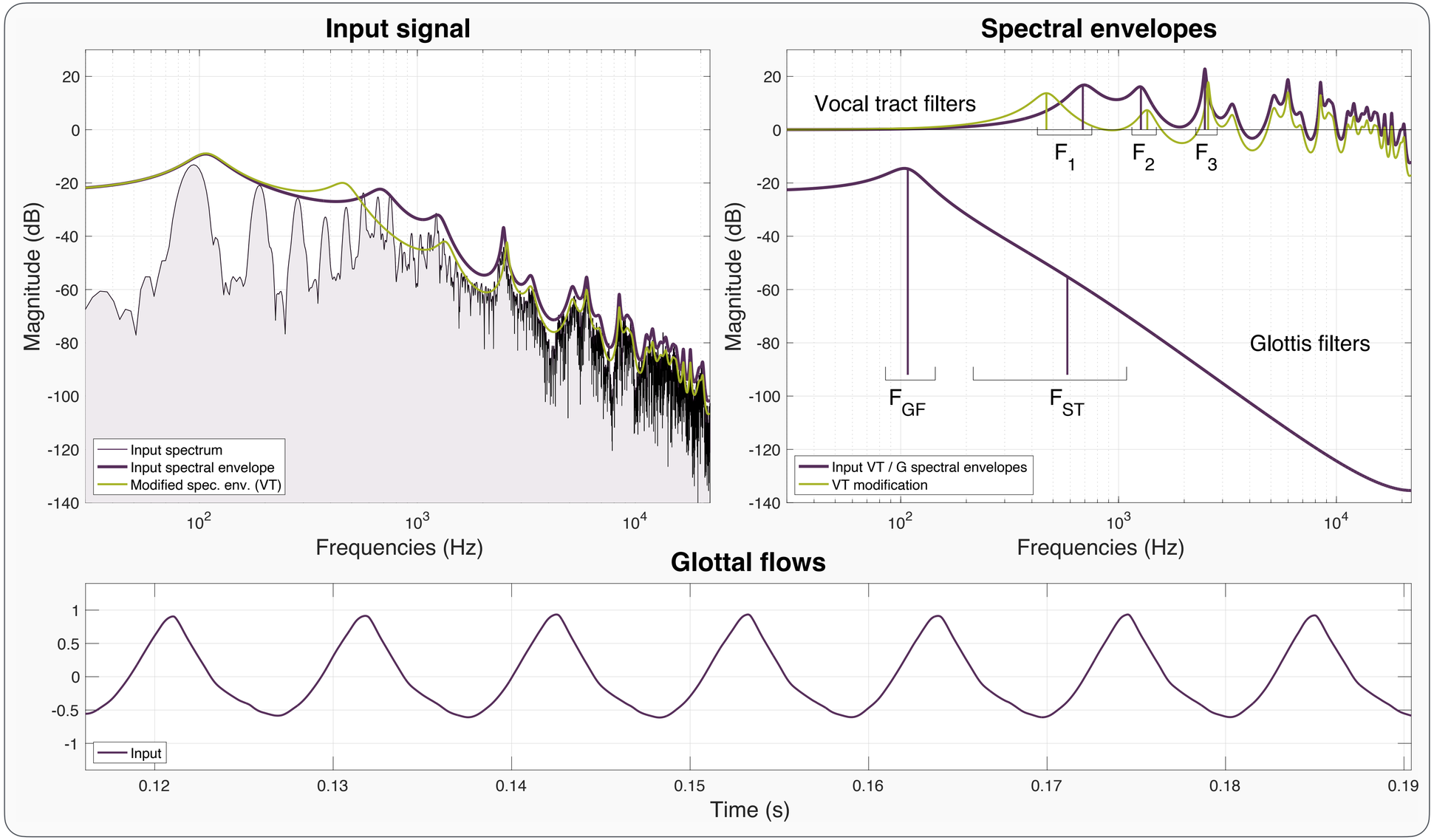
Modification of filter parameters



Perrotin, O. *and McLoughlin, I. V. (2019), Proc. Interspeech, pp. 3685—3686*

Glottal model for analysis and synthesis

Modification of filter parameters



Acoustic correlates

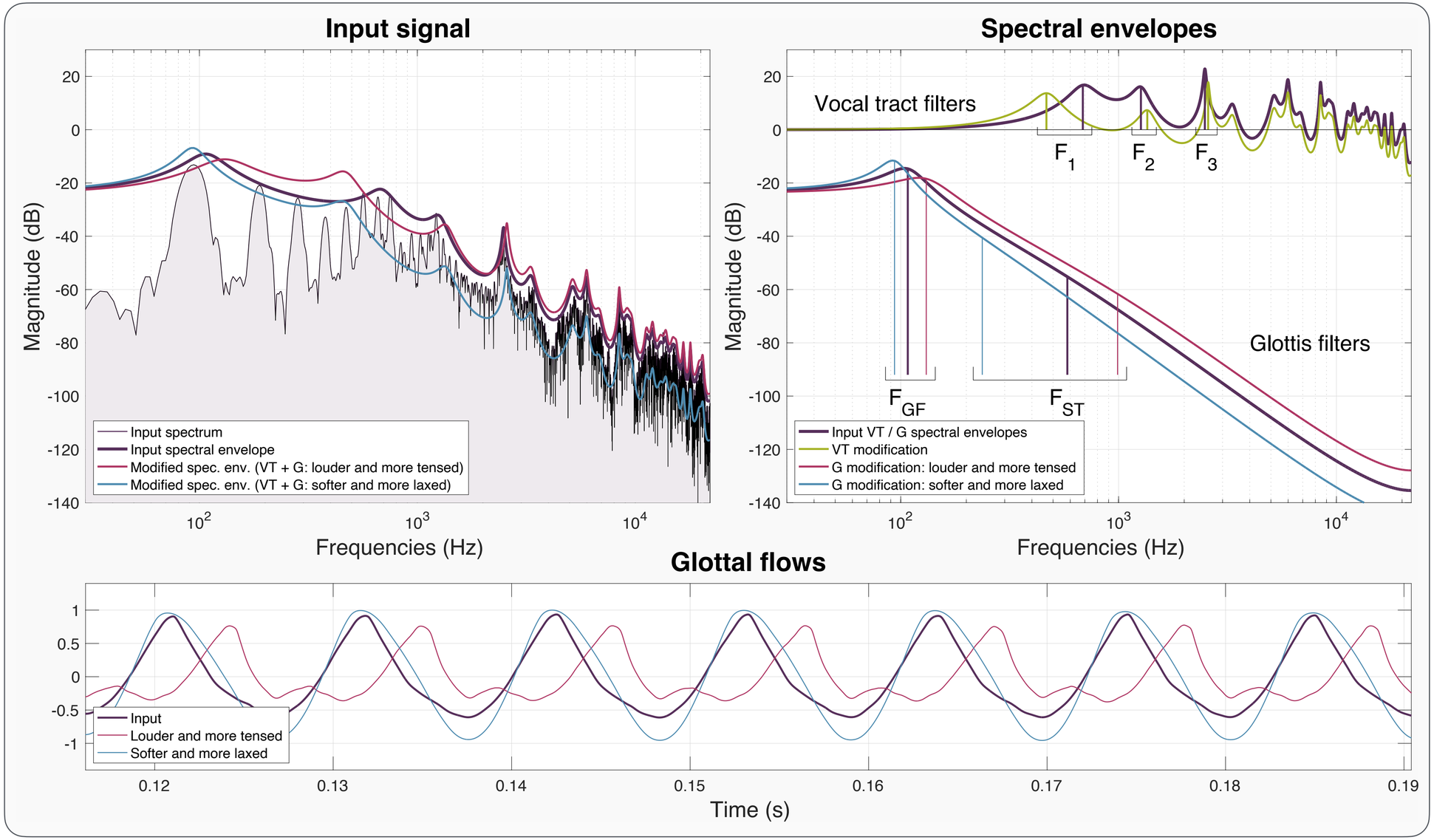
Vocal Tract

Jaw	Closing ↔	Opening
F_1	Lower	Higher
Tongue	Backwards ↔	Forwards
F_2	Lower	Higher
Lip	Rounding ↔	Stretching
F_3	Lower	Higher

Perrotin, O. *and McLoughlin, I. V. (2019), Proc. Interspeech, pp. 3685—3686*

Glottal model for analysis and synthesis

Modification of filter parameters



Acoustic correlates

Vocal Tract

Jaw	Closing ↔	Opening
F_1	Lower	Higher
Tongue	Backwards ↔	Forwards
F_2	Lower	Higher
Lip	Rounding ↔	Stretching
F_3	Lower	Higher

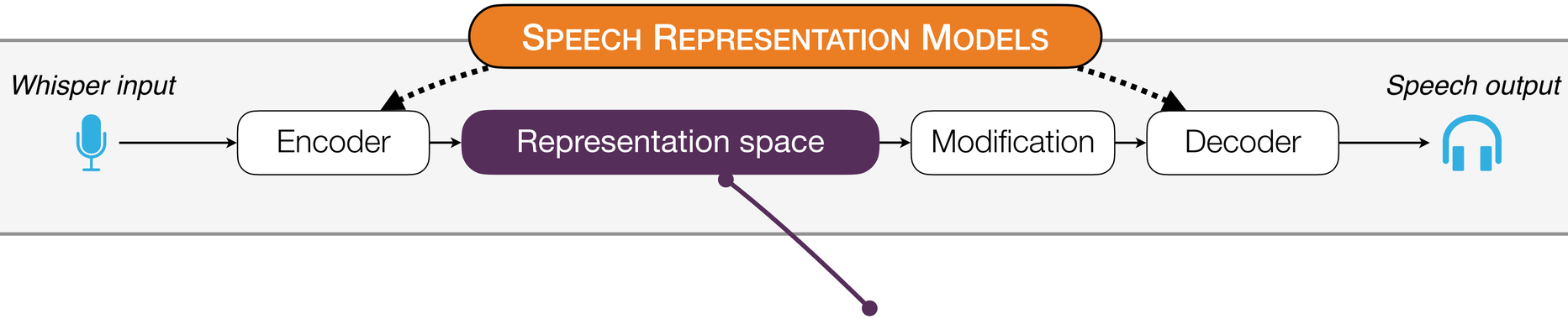
Glottis

Effort	Softer ↔	Louder
F_{GF}	Lower	Higher
F_{ST}	Lower	Higher
Tenseness	More laxed ↔	More tensed
F_{GF}	Lower	Higher
Q_{GF}	Narrower	Wider

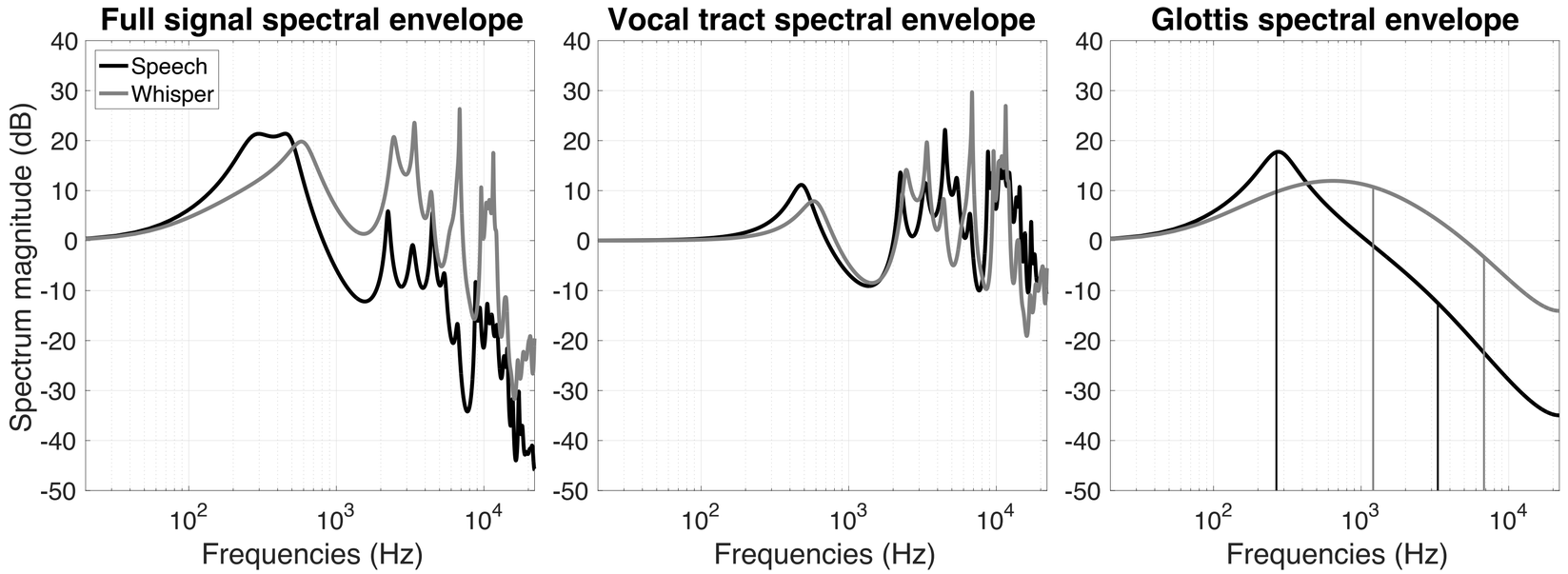
Perrotin, O. *and McLoughlin, I. V. (2019), Proc. Interspeech, pp. 3685—3686*

Glottal model for analysis and synthesis

Whisper-to-Speech conversion



- Change noise to harmonic excitation (pulse train)
- Modify glottal parameters
- Leave vocal tract parameters (for now)

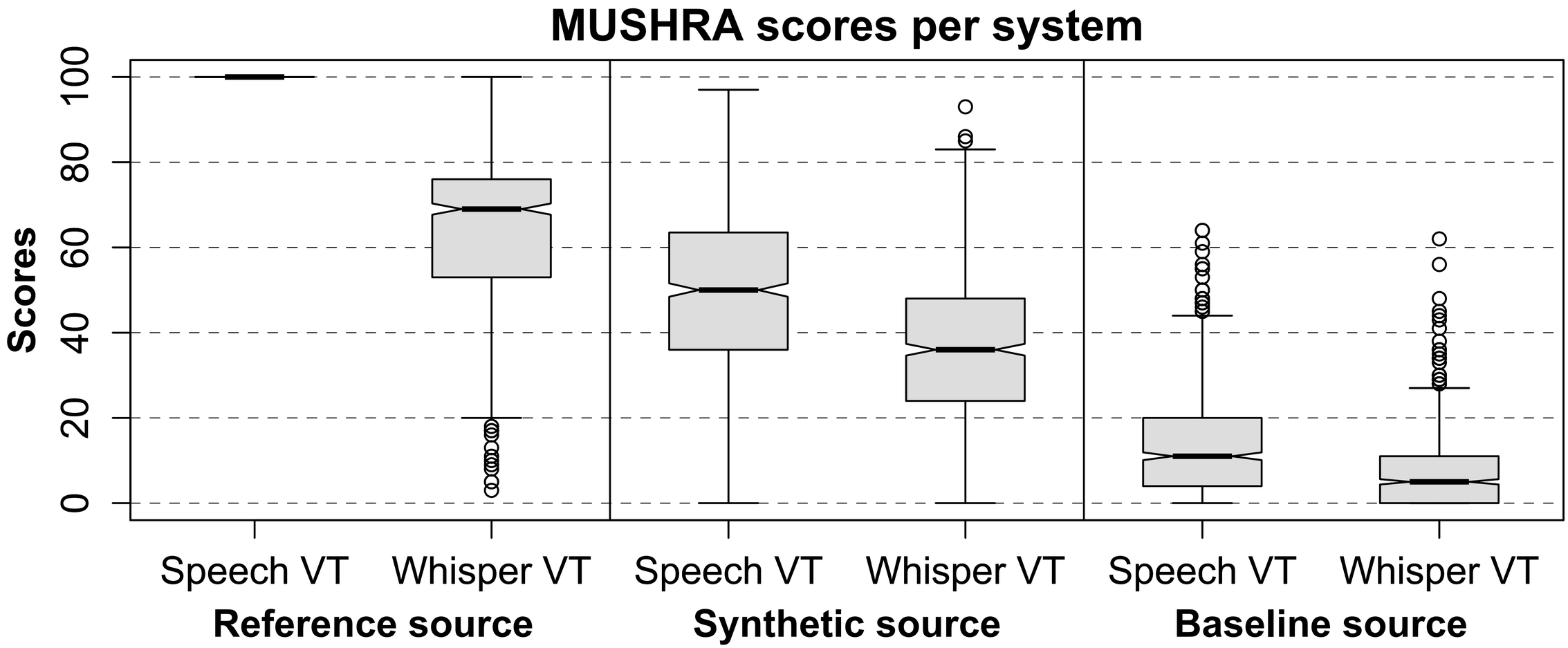


Extracted from vowel /i/ in 'is'

Perrotin, O. *and McLoughlin, I. V. (2020), IEEE TASLP, 28, pp. 889—900*

Glottal model for analysis and synthesis

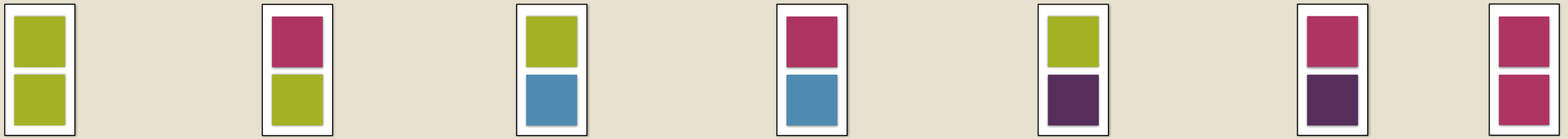
Whisper-to-speech conversion



Whisper
Speech
Synth
Baseline

Vocal Tract

Glottis



Glottal model for analysis and synthesis

Summary

- **GFM-Voc:** Analysis-Synthesis method
  - Source-filter decomposition method based on a Glottal Flow Model
  - ➡ Allow to analyse / modify glottis and vocal tract parameters independently
- Various applications since
  - Real-time voice transformation
    - Vocalic formants and voice quality modification
    - Whisper to speech conversion **in real-time** — Luc Ardaillon
  - Voice analysis
    - Assessment of glottal dysfunction

- Small model (10-100 parameters)
  - Explicit model (acoustically / signal / physiologically informed)
  - Generally fast to compute
- Limited modelling power
  - Light modelling of covariations
  - Not so high-quality synthesis

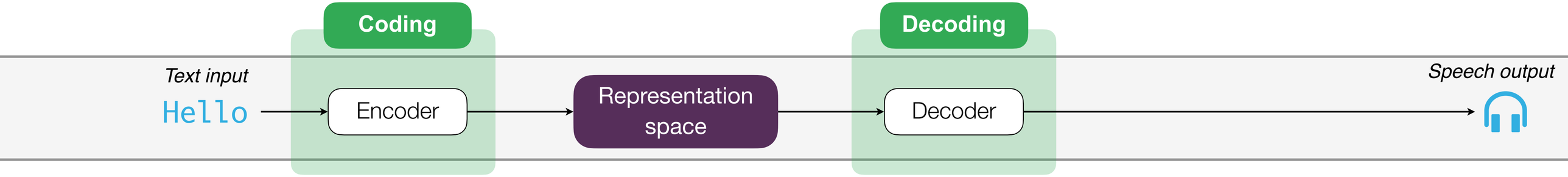
Analysis-synthesis of expressive speech

- Signal-based models : encoding / decoding and applications
- Neural-based models : study of representation spaces

Exploration of representation spaces

Martin Lenglet

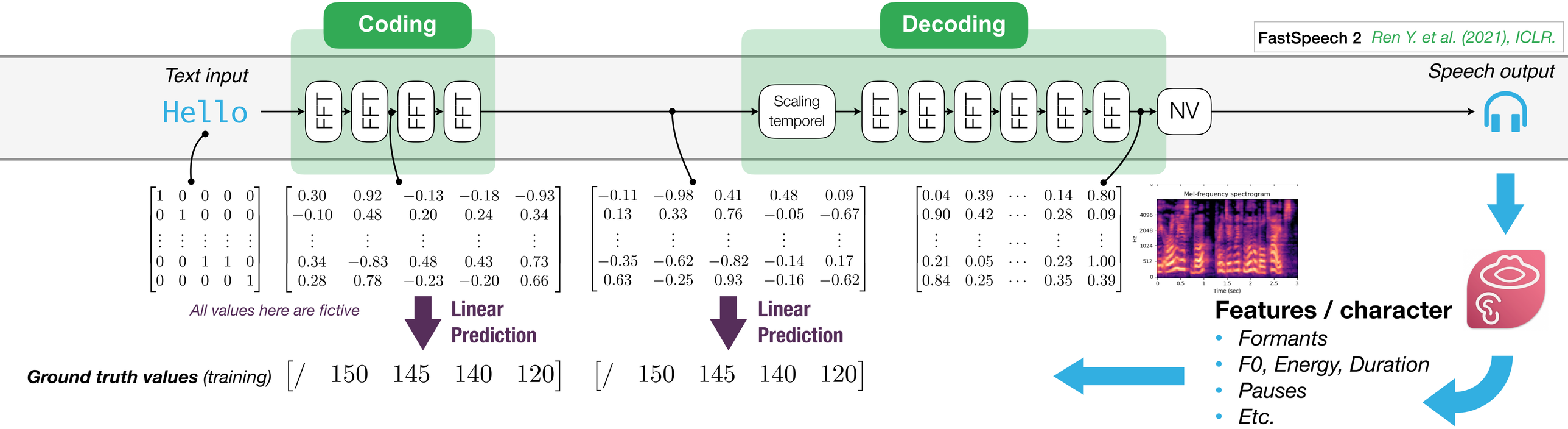
Linear probing on TTS



Exploration of representation spaces

Martin Lenglet

Linear probing on TTS

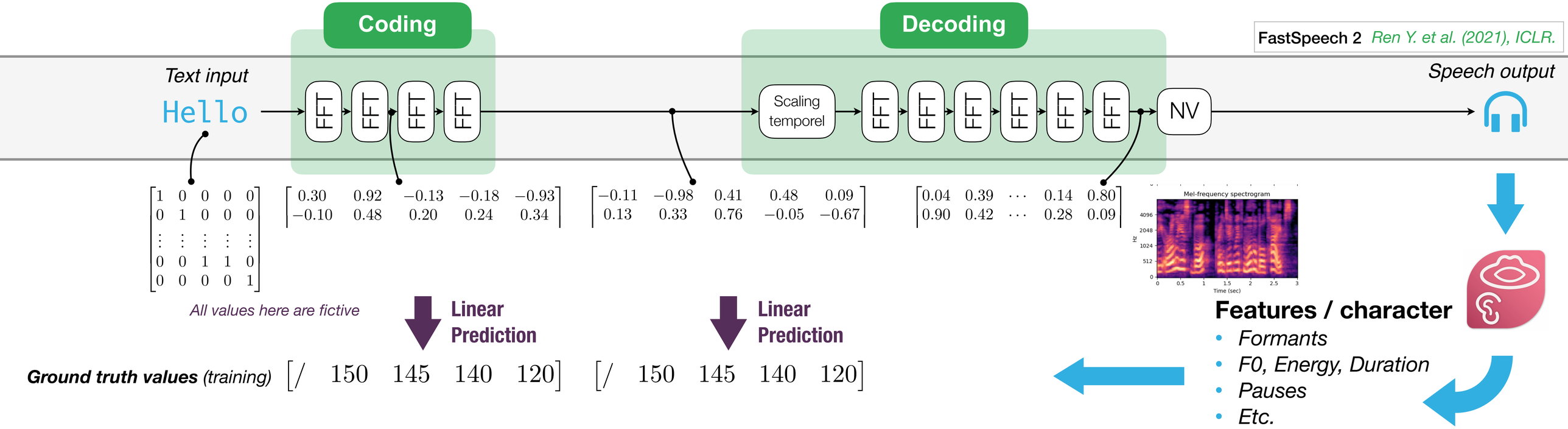


- Train the model on a dataset (33h French female speaker)
- Put ~2000 utterances in the model, save all layers, and acoustic analysis
- Learn a linear prediction of acoustic parameters at each layer
  - Linear regression for continuous parameters
  - Linear discriminant analysis (LDA) for discrete parameters

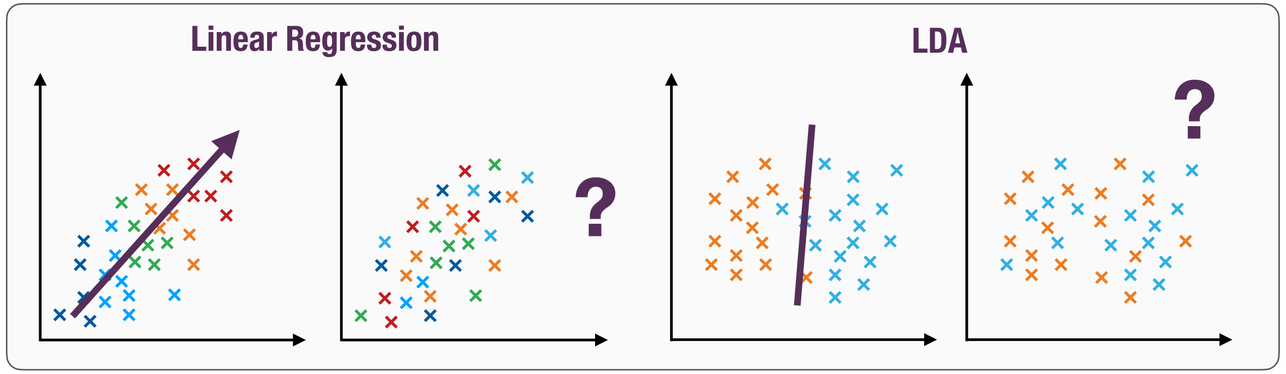
Exploration of representation spaces

Martin Lenglet

Linear probing on TTS



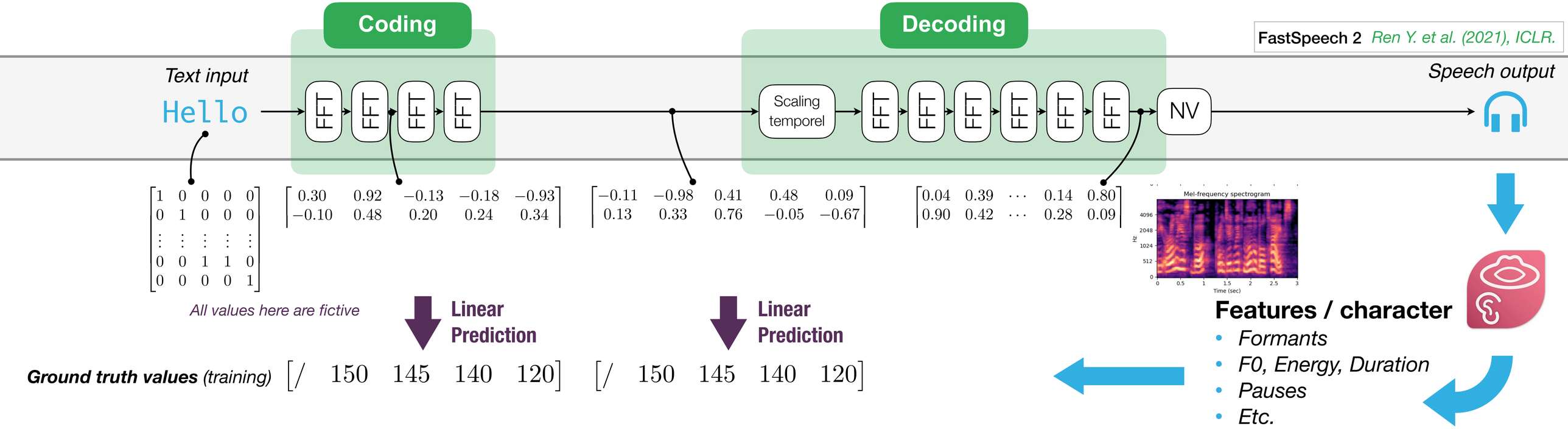
- Train the model on a dataset (33h French female speaker)
- Put ~2000 utterances in the model, save all layers, and acoustic analysis
- Learn a linear prediction of acoustic parameters at each layer
  - Linear regression for continuous parameters
  - Linear discriminant analysis (LDA) for discrete parameters



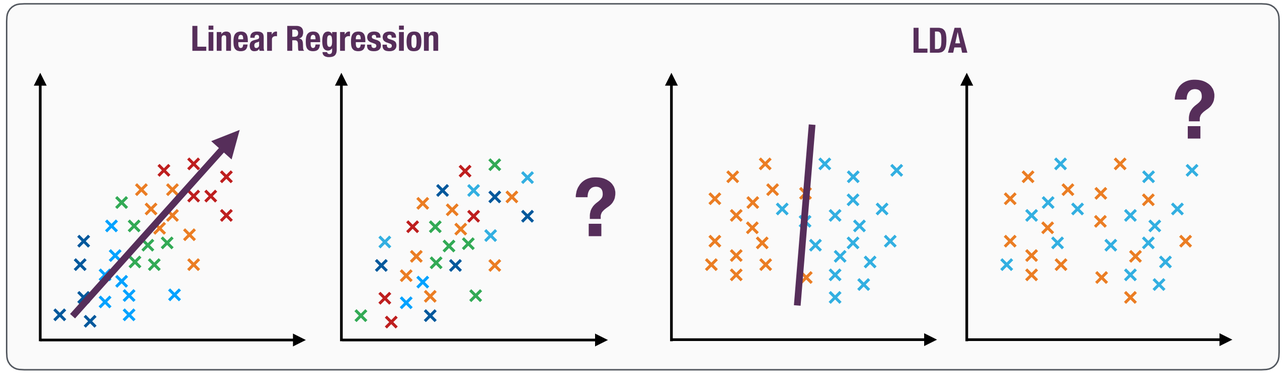
Exploration of representation spaces

Martin Lenglet

Linear probing on TTS



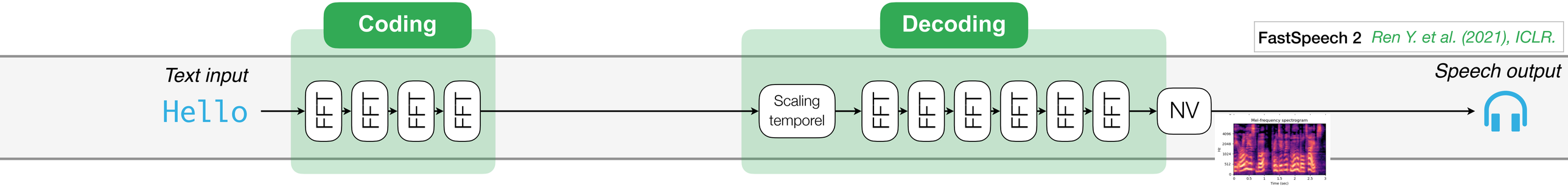
Where in the model are acoustic parameters linearly encoded?



Exploration of representation spaces

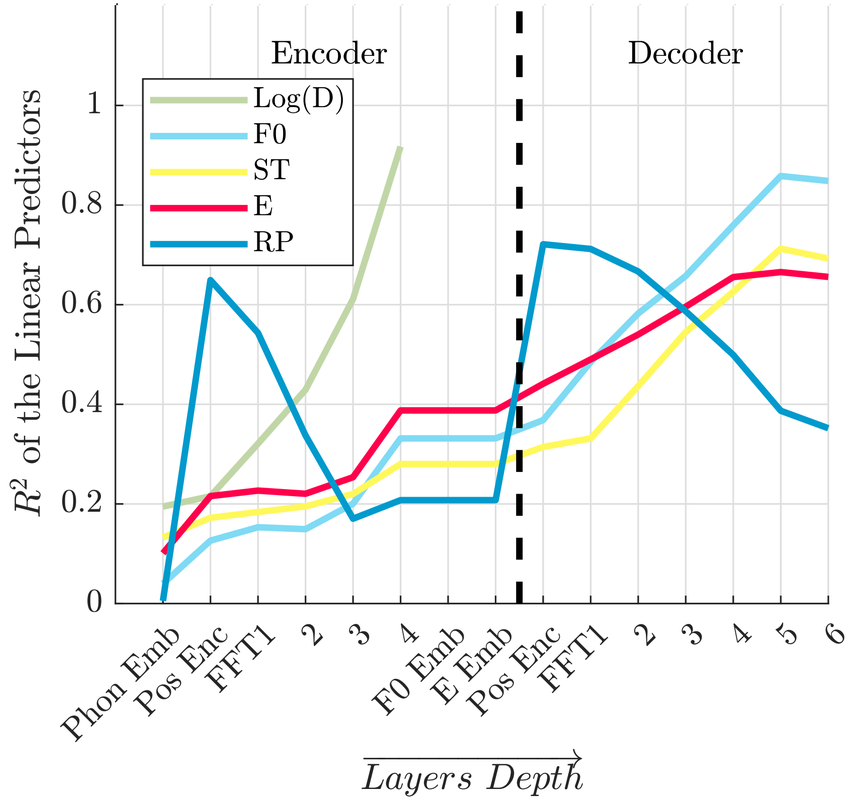
Martin Lenglet

Linear probing on TTS



FastSpeech 2 *Ren Y. et al. (2021), ICLR.*

Supra-segmental parameters

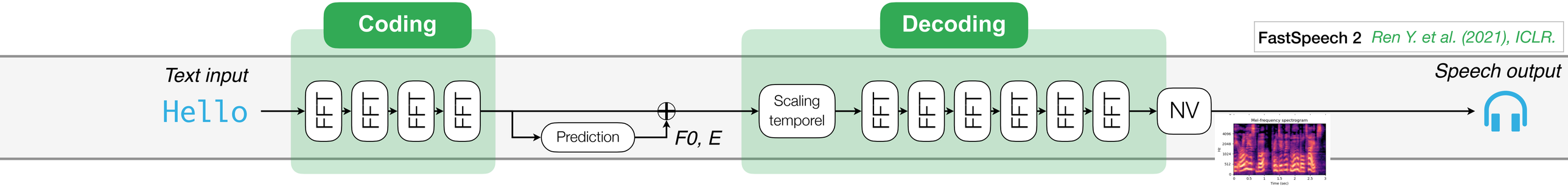


Lenglet Martin, (2023), PhD Thesis

Exploration of representation spaces

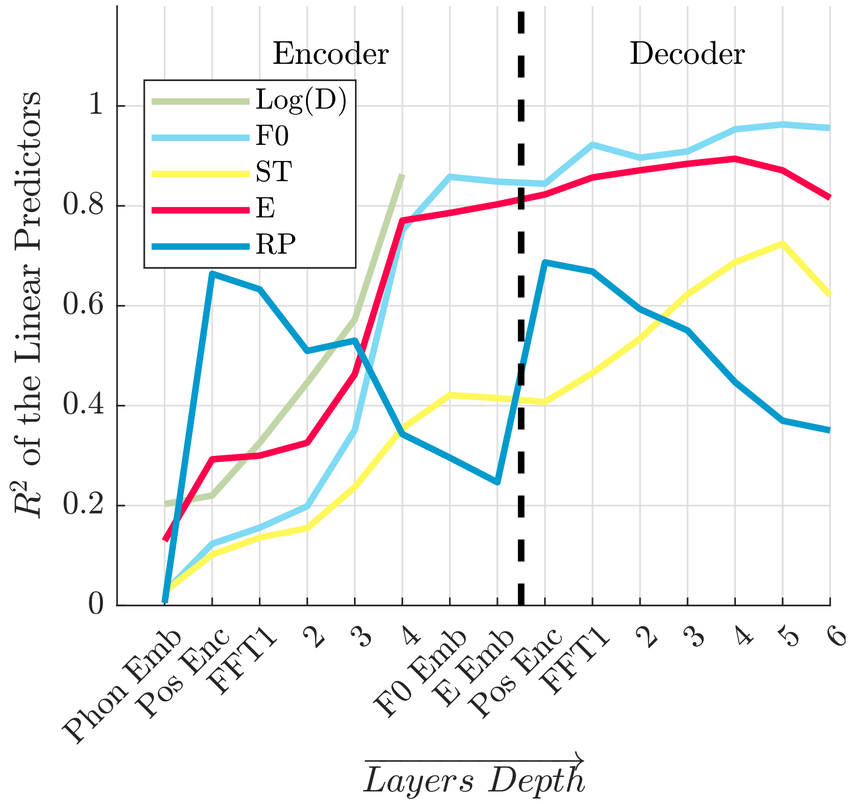
Martin Lenglet

Linear probing on TTS



FastSpeech 2 *Ren Y. et al. (2021), ICLR.*

Supra-segmental parameters

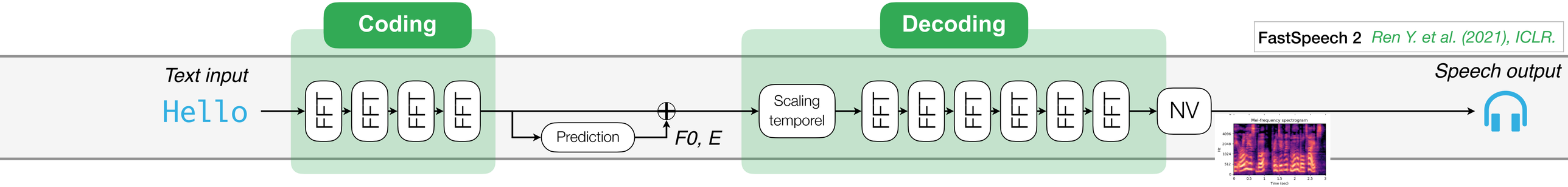


Lenglet Martin, (2023), PhD Thesis

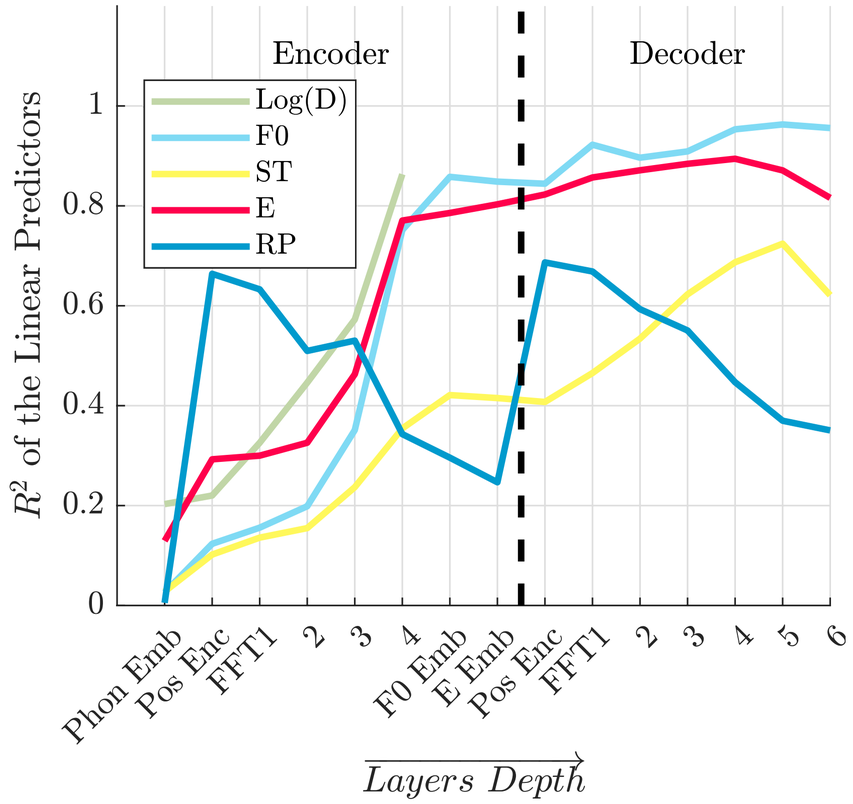
Exploration of representation spaces

Martin Lenglet

Linear probing on TTS



Supra-segmental parameters



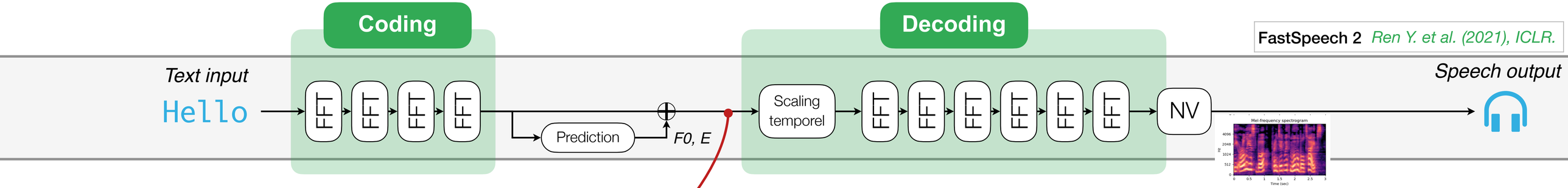
- Progressive linear coding of acoustic parameters
- Phonetic first, then prosodic
- Helped with forced predictions

Lenglet Martin, (2023), PhD Thesis

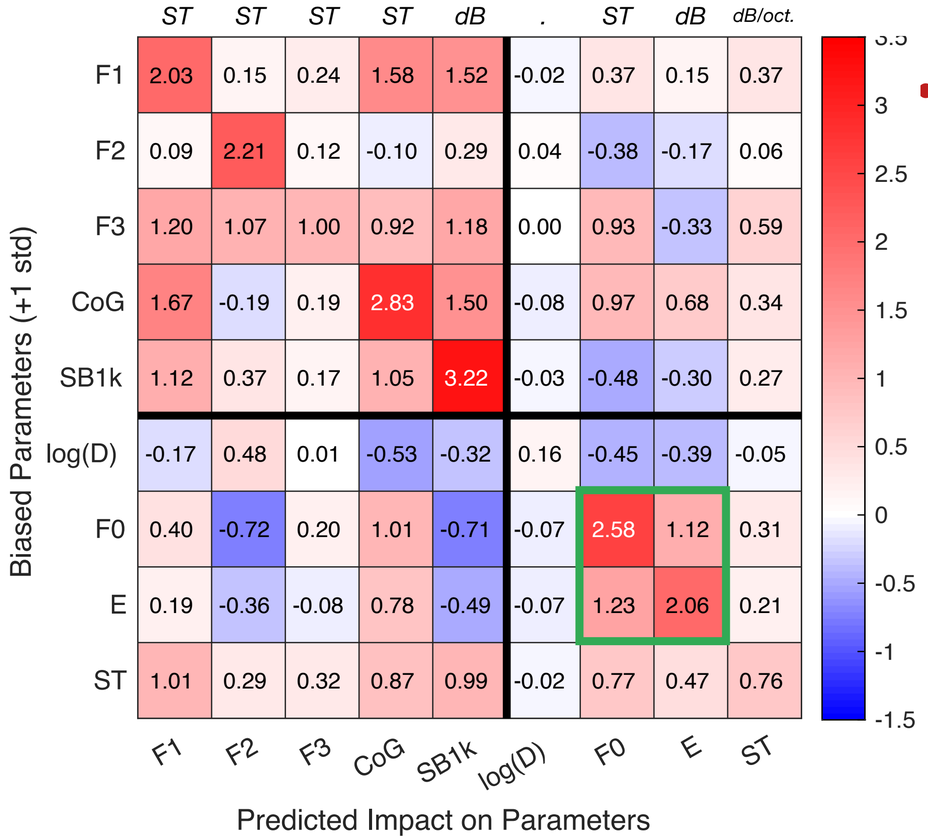
Exploration of representation spaces

Martin Lenglet

Control



FastSpeech 2 *Ren Y. et al. (2021), ICLR.*



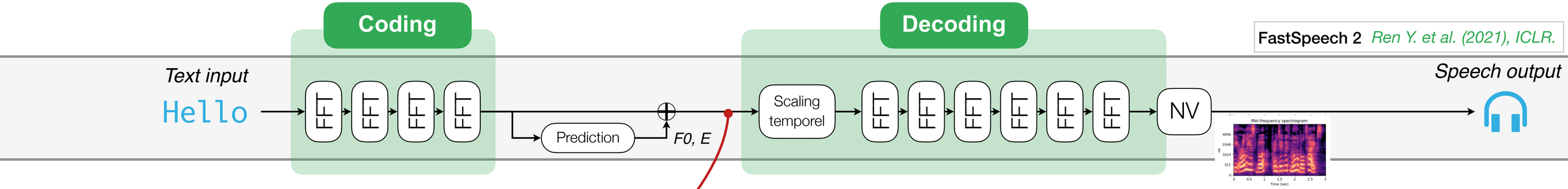
- Observation of covariations
- Consistent with literature
  - F0 and Energy

Lenglet Martin, (2023), PhD Thesis

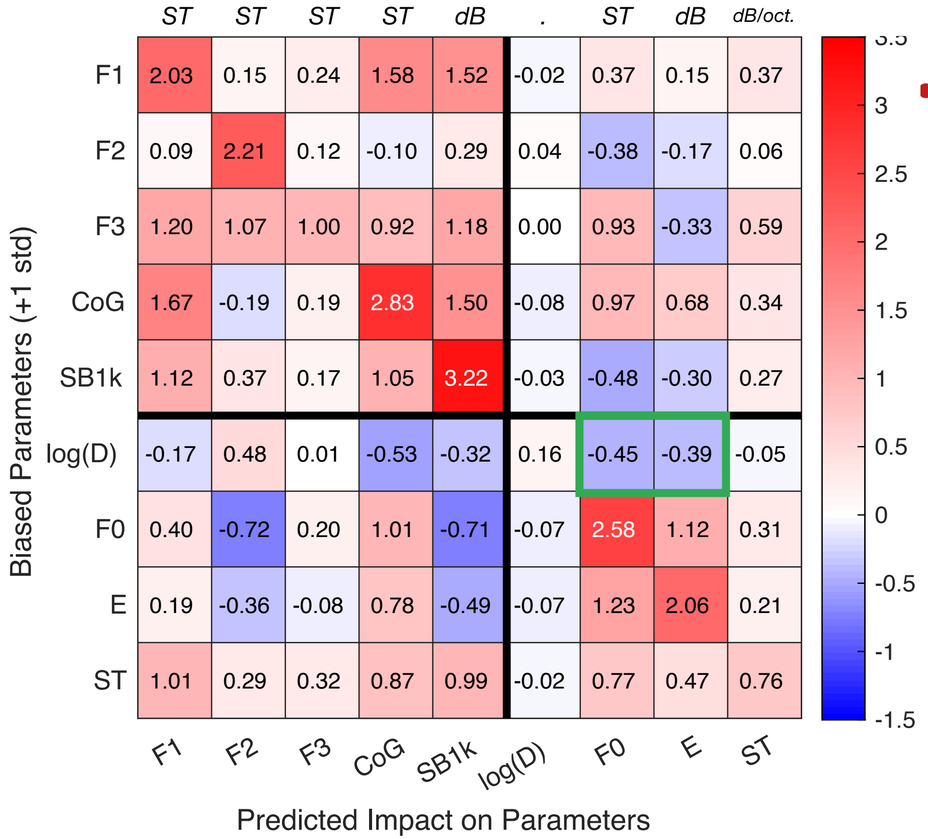
Exploration of representation spaces

Martin Lenglet

Control



FastSpeech 2 *Ren Y. et al. (2021), ICLR.*



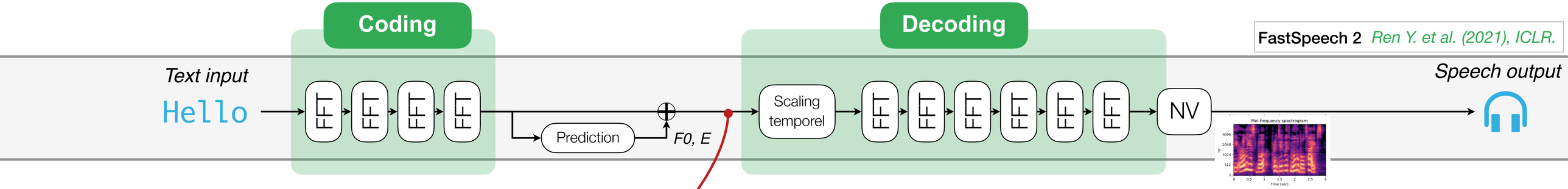
- Observation of covariations
- Consistent with literature
  - F0 and Energy
  - Duration, F0 and Energy

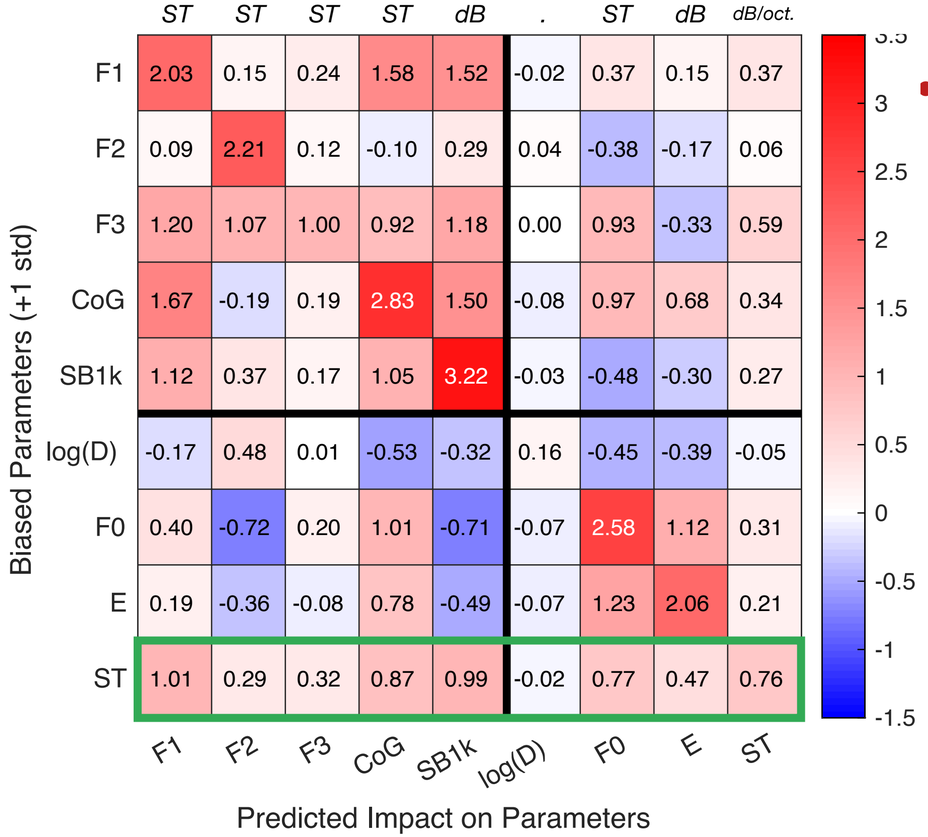
Lenglet Martin, (2023), PhD Thesis

Exploration of representation spaces

Martin Lenglet

Control





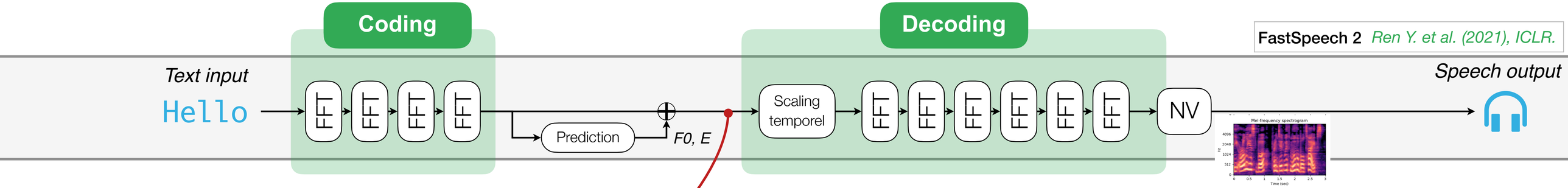
- Observation of covariations
- Consistent with literature
  - F0 and Energy
  - Duration, F0 and Energy
  - Vocal effort correlates (F0, Energy, F1, Spectral tilt)

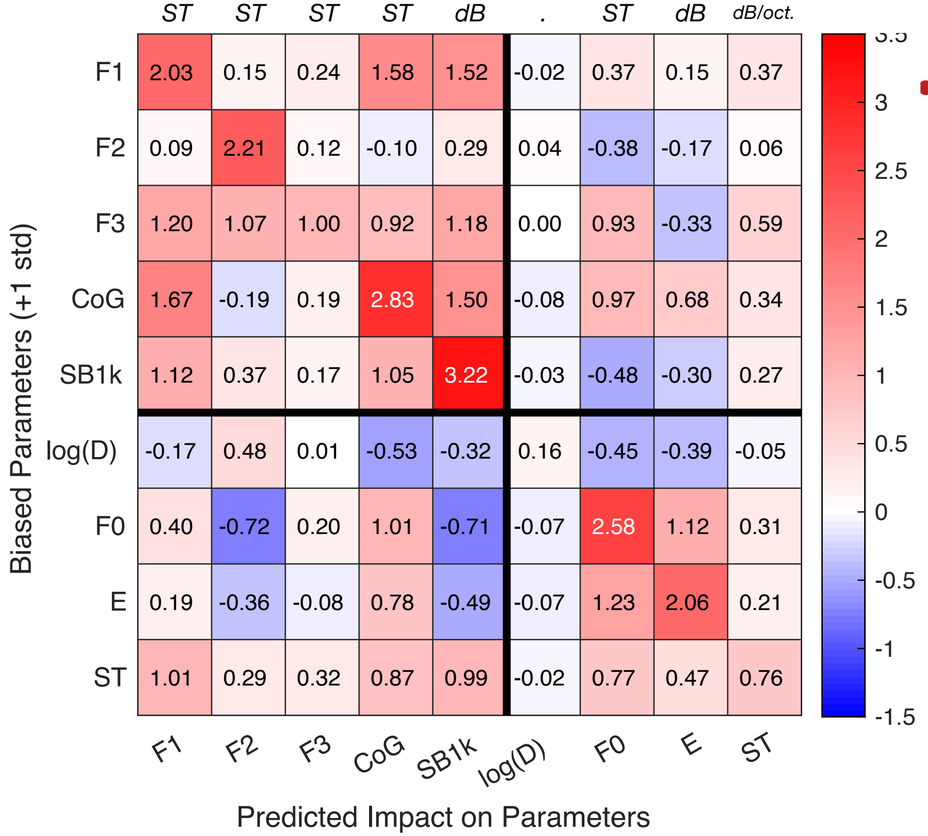
Lenglet Martin, (2023), PhD Thesis

Exploration of representation spaces

Martin Lenglet

Control





- Observation of covariations
- Consistent with literature
  - F0 and Energy
  - Duration, F0 and Energy
  - Vocal effort correlates (F0, Energy, F1, Spectral tilt)
- New ones to find?

➡ Powerful analysis tool learnt on massive data

Lenglet Martin, (2023), PhD Thesis

Exploration of representation spaces

Martin Lenglet

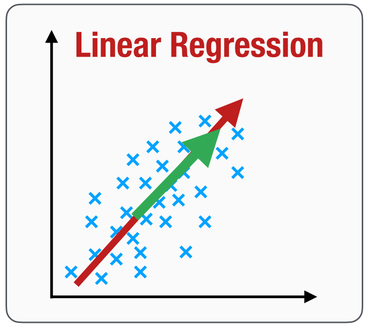
Control of duration and pauses

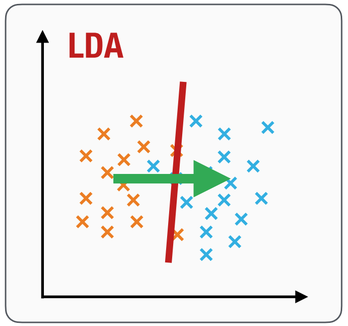
- Control of duration in an utterance
  - Single modification for all phones
  - Different behaviour depending on segment

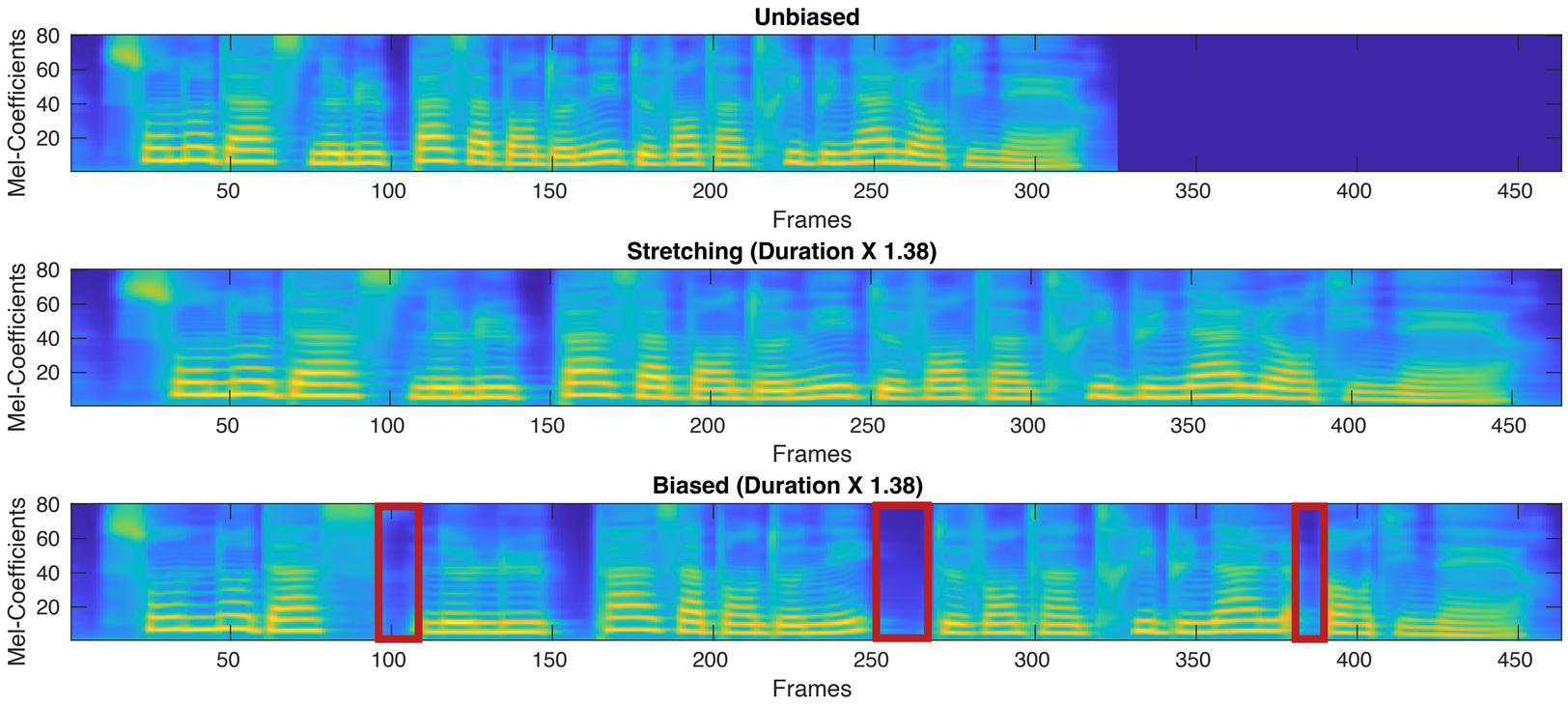
  ➡ Observe a saturation of elongation for final vowels and silences

- Control of percentage of pauses in an utterance
  - Data-grounded heuristic to link percentage of pause and duration
  - Position of pauses left to the model

- Evaluation of both
  - Significant preference (resp. worst score) when pauses are well (resp. wrongly) placed





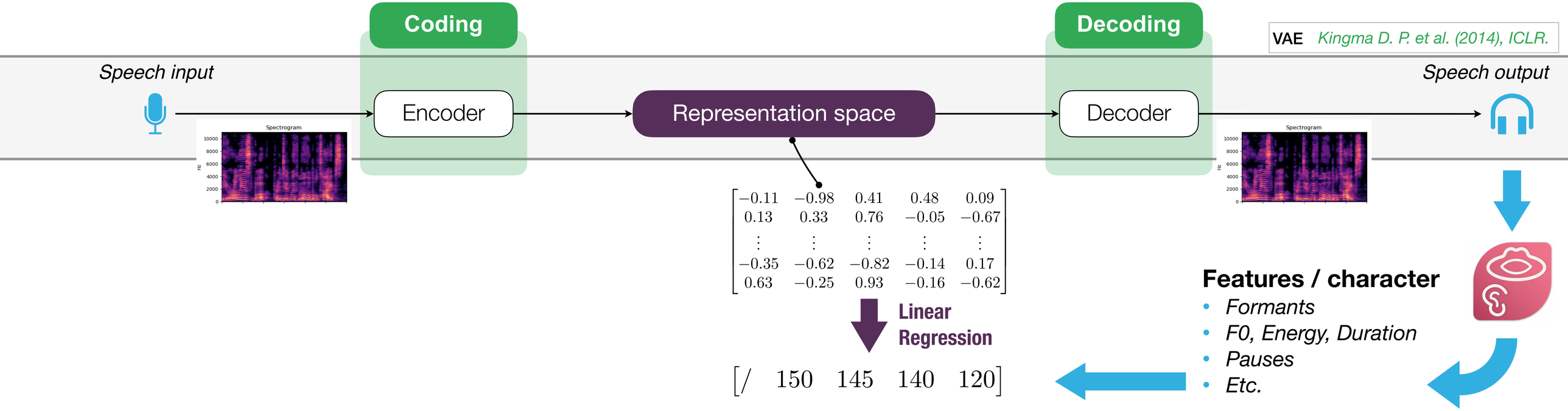


Lenglet Martin, (2023), PhD Thesis

Exploration of representation spaces

Maxime Jacquelin

Linear probing on STS



- **Multi-speaker** training (109 English Speakers M/F)
- Linear probing
  - Linear regression for continuous parameters: **consistent results with text-to-speech**
  - Linear discriminant analysis (LDA) **for speaker classes**

Jacquelin M., Garnier M., Girin L., Vincent R. ***Perrotin O.****, (2023), Proc. SSW, pp. 240—241*

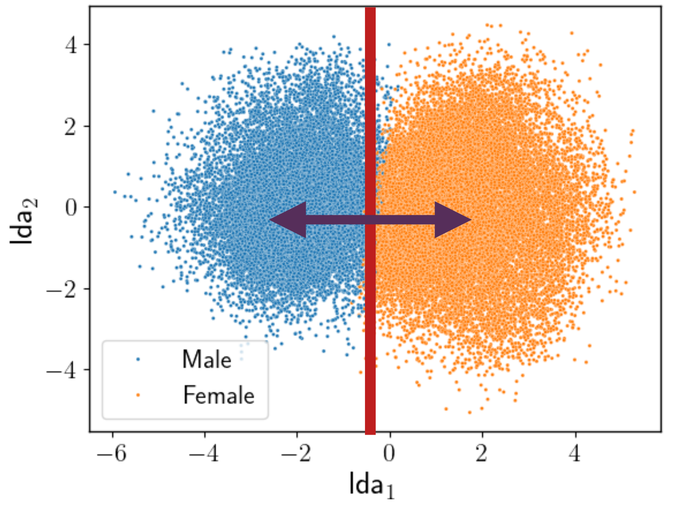
Exploration of representation spaces

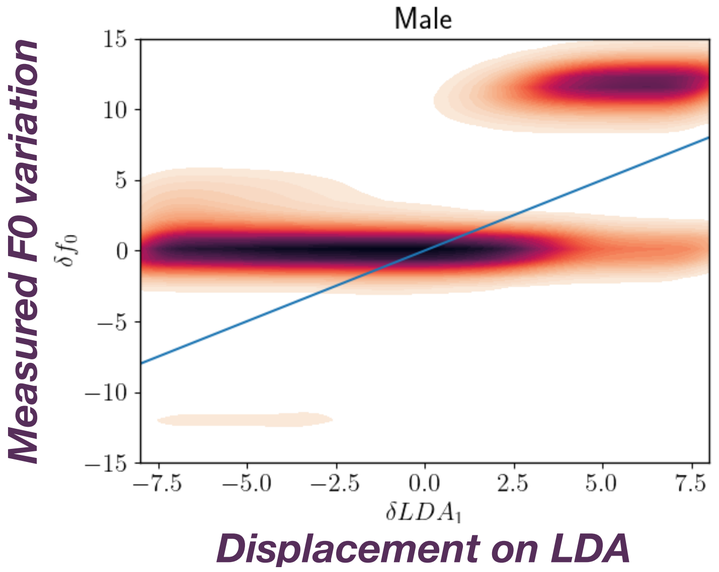
Maxime Jacquelin

Linear probing on STS

- First hyperplan of LDA discriminates gender
- Control across first LDA hyperplan
  - Discrete 1-octave jumps of F0

Each point is a spectrogram frame





Jacquelin M., Garnier M., Girin L., Vincent R. ***Perrotin O.****, (2023), Proc. SSW, pp. 240—241*

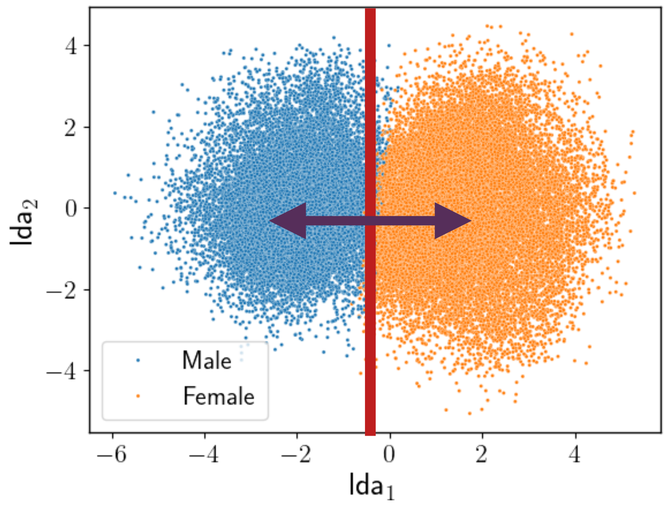
Exploration of representation spaces

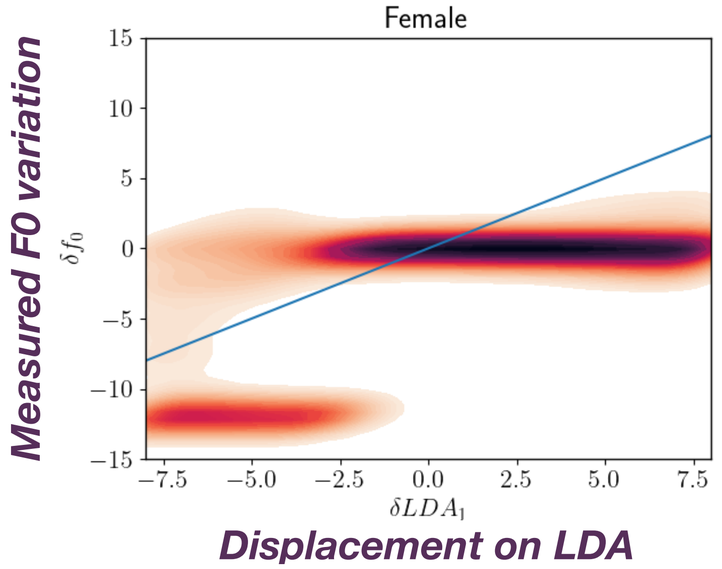
Maxime Jacquelin

Linear probing on STS

- First hyperplan of LDA discriminates gender
- Control across first LDA hyperplan
  - Discrete 1-octave jumps of F0

Each point is a spectrogram frame





Jacquelin M., Garnier M., Girin L., Vincent R. ***Perrotin O.****, (2023), Proc. SSW, pp. 240—241*

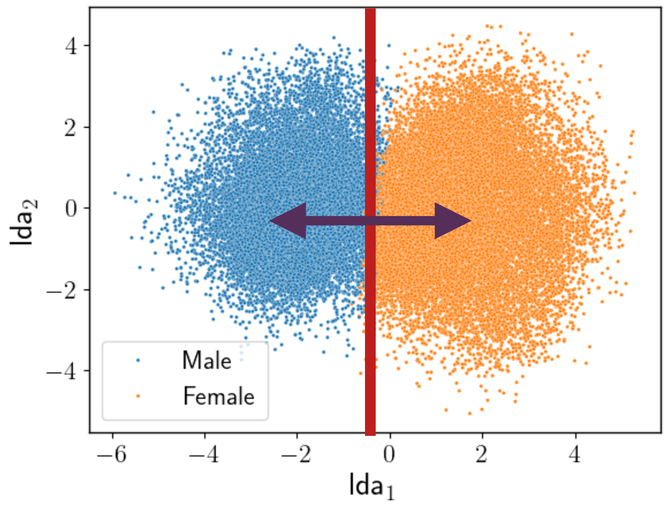
Exploration of representation spaces

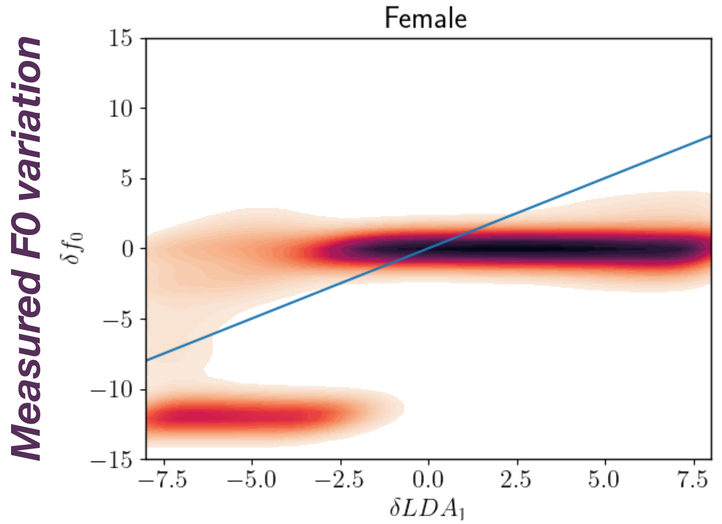
Maxime Jacquelin

Linear probing on STS

- First hyperplan of LDA discriminates gender
- Control across first LDA hyperplan
  - Discrete 1-octave jumps of F0
- Dual coding behaviour of acoustic parameters
  - Continuous (via regressions)
    Intra-class
  - Categorial (via LDA on speech production classes)
    Inter-class

Each point is a spectrogram frame





Jacquelin M., Garnier M., Girin L., Vincent R. ***Perrotin O.****, (2023), Proc. SSW, pp. 240—241*

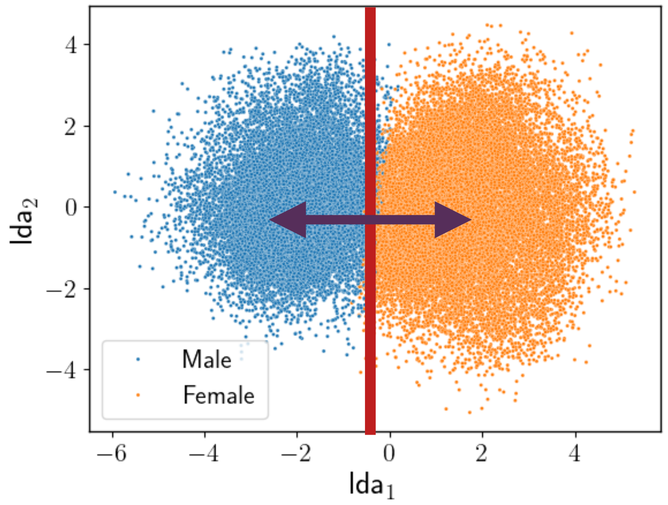
Exploration of representation spaces

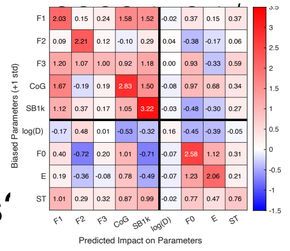
Maxime Jacquelin

Linear probing on STS

- First hyperplan of LDA discriminates gender
- Control across first LDA hyperplan
  - Discrete 1-octave jumps of F0
- Dual coding behaviour of acoustic parameters
  - Continuous (via regressions)
    Intra-class
  - Categorial (via LDA on speech production classes)
    Inter-class
- Open the door to rich analysis of complex production modes in a single representation space
  - Future experiment on classes of vocal force
  - Can we expect categorial coding of covariations?
  - Can we expect speaker-dependant coding of covariations'
  - Can we control intra- and inter-class variations?

Each point is a spectrogram frame





Jacquelin M., Garnier M., Girin L., Vincent R. ***Perrotin O.****, (2023), Proc. SSW, pp. 240—241*

Exploration of representation spaces

Linear probing and control in neural models

- Method
  - Linear probing: show the encoding of acoustic parameters at each layer
  - Causal control: non-linear control of both continuous and discrete acoustic parameters

➡ Powerful analysis tool learnt on massive data

➡ Works on different architectures (VAE, RNN, Transformers)

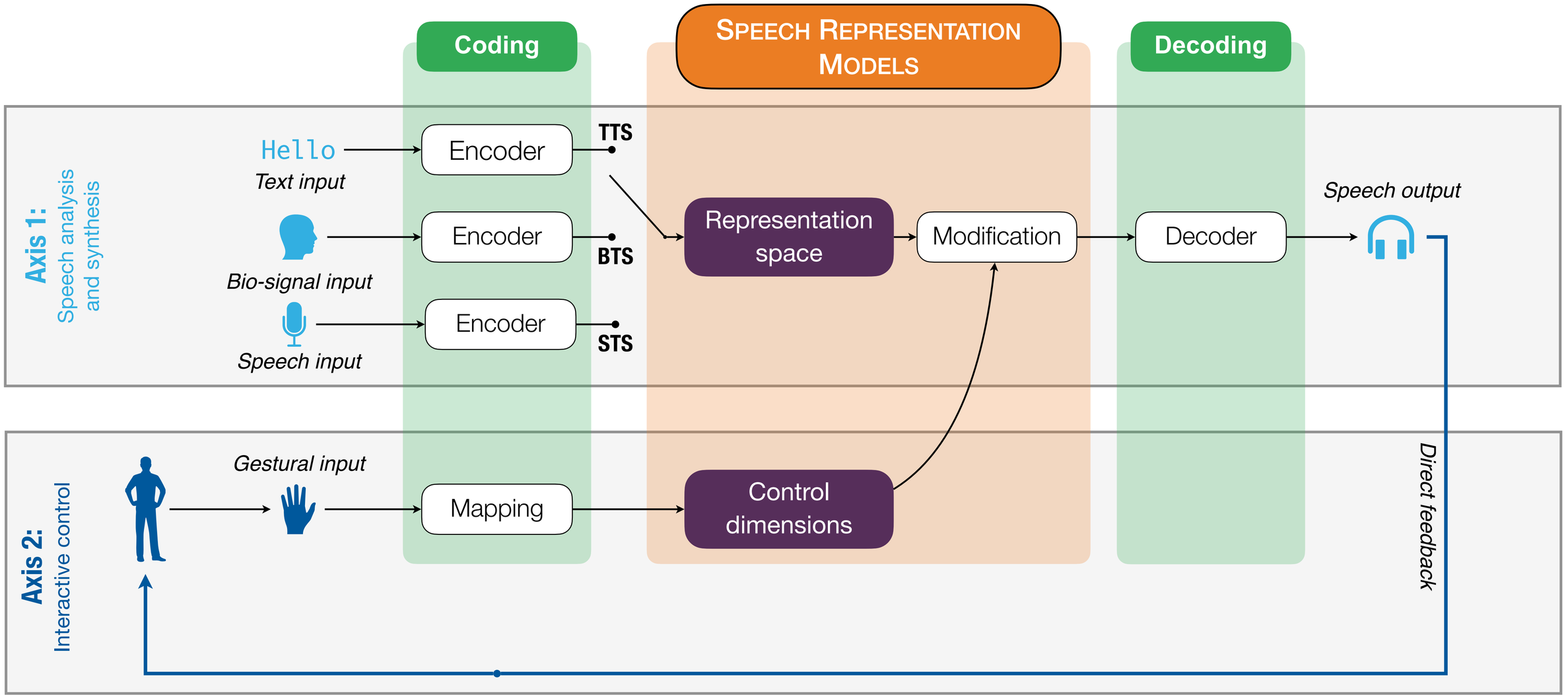
Exploration of representation spaces

	Signal-based	Neural-based
• Generate high naturalness speech signals	Limited	✓
• Identify speech representation models which display interpretable and controllable expressive dimensions	Ceiling	Very promising
• Perform a robust modification of these parameters without degrading the signal quality	Within the limits of the model quality	Very promising
• Implement everything in real-time	✓	To come

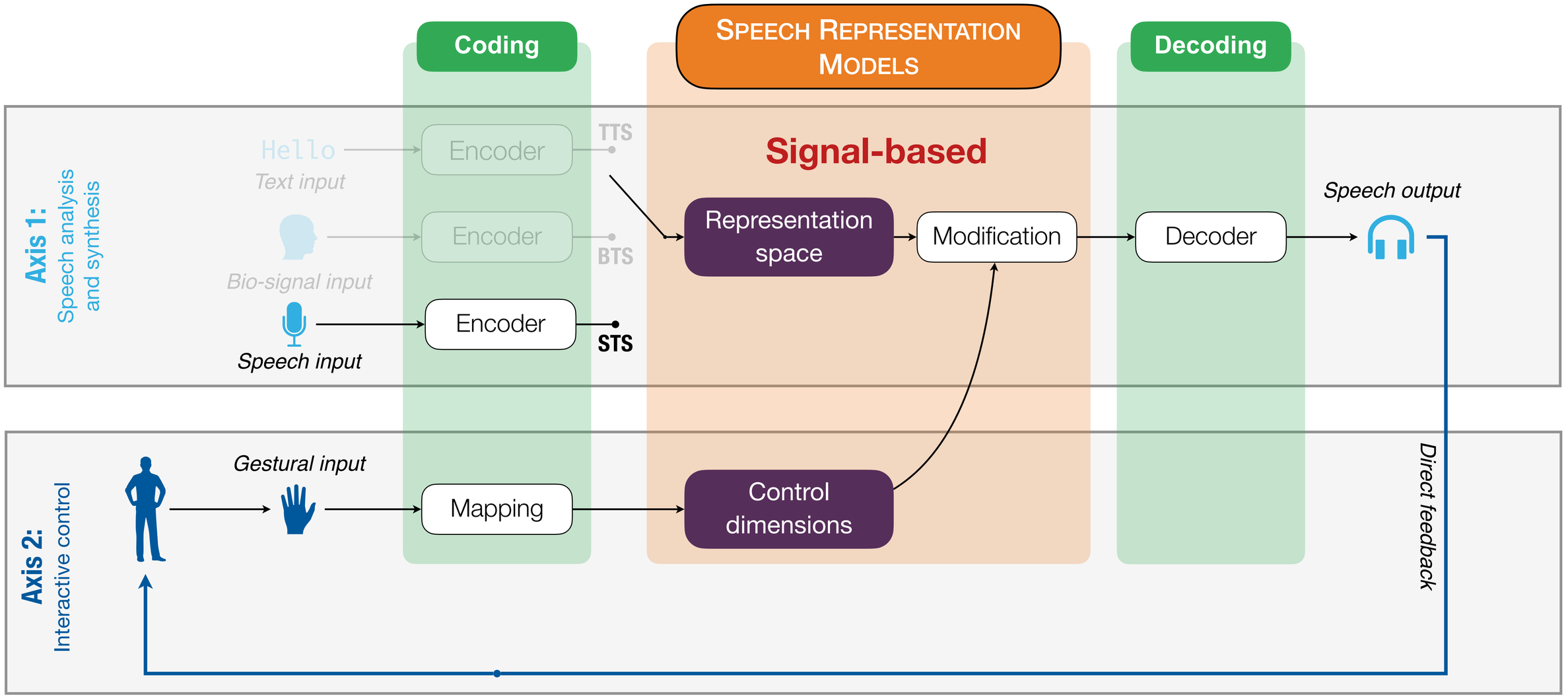
Interactive control of synthesis

- Explicit control of F0 in singing
- From implicit to explicit control of F0 in speaking
- Towards a co-adaptation between human and machine learning of control mapping

Interactive control of expressive speech synthesis



Interactive control of expressive speech synthesis



Interactive control of synthesis

Litt. Review

d'Alessandro C. et al. (2022), JEP, pp. 625–636.

Singing

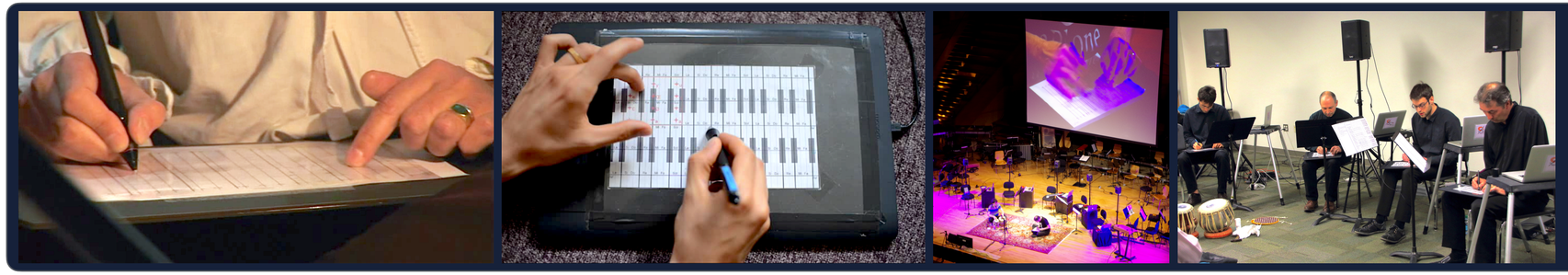
Speech Conductor
eNTERFACE - Belgique

Synthesis : formants / concatenative
Control : keyboard / data glove

d'Alessandro C. et al. (2005), eNTERFACE, pp. 52–61.

Cantor Digitalis
Le Beux / Feugère / Perrotin

Synthesis : formant
Control:
- Intonation: **graphic tablet**
- Vowels: **graphic tablet**



*Feugère L., d'Alessandro C., Doval B., **Perrotin O.** (2017), JASM, 2017(2).*

- **Evaluation of the interface** for the control of intonation
  - Evaluation of the **precision** allowed by the graphic tablet
  - Quantification of the **influence of visual and auditive modalities** on the control
  - Proposition of a series of **gestures** for expressive melodic control

*d'Alessandro C., Feugère L., Le Beux S., **Perrotin O.**, Rilliard A. (2014), JASA, pp. 3601–3612.*

***Perrotin O.**, d'Alessandro C. (2016), ACM Trans. Applied Perception, 14(2).*

***Perrotin O.** (2015), PhD Thesis.*

➡ Singing intonation is **explicit**, speech intonation is **implicit**

Interactive control of synthesis

Control of explicit F0

Litt. Review

d'Alessandro C. et al. (2022), JEP, pp. 625–636.

Speech Conductor
eNTERFACE - Belgique

Synthesis : formants / concatenative
Control : keyboard / data glove

d'Alessandro C. et al. (2005), eNTERFACE, pp. 52–61.

Singing

Cantor Digitalis
Le Beux / Feugère / Perrotin

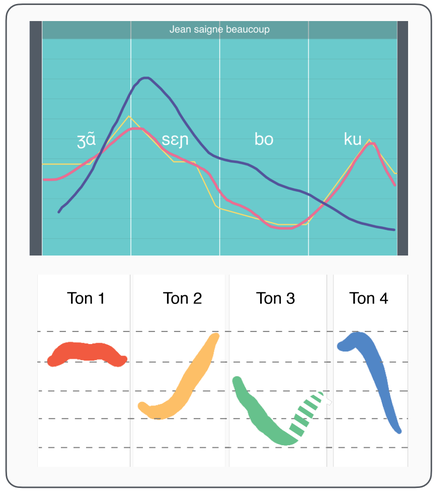
Synthesis : formant
Control:
- Intonation: **graphic tablet**
- Vowels: **graphic tablet**

Speech

Voks (Calliphony/ Vokinesis)
Le Beux / Delalez / Locqueville / Xiao

Synthesis : transformation of playback
Control:
- Intonation: **graphic tablet / theremin**
- Biphasic rhythm: **foot pedal, hand button**

Le Beux S. et al. (2007), ISCA SSW, pp. 345–350.
Delalez S. et al. (2017), Proc. NIME, pp. 198–203.
Delalez S. et al. (2017), Proc. Interspeech, pp. 864–868.
Locqueville G. et al. (2020), Speech Comm., 120, pp. 97–113.



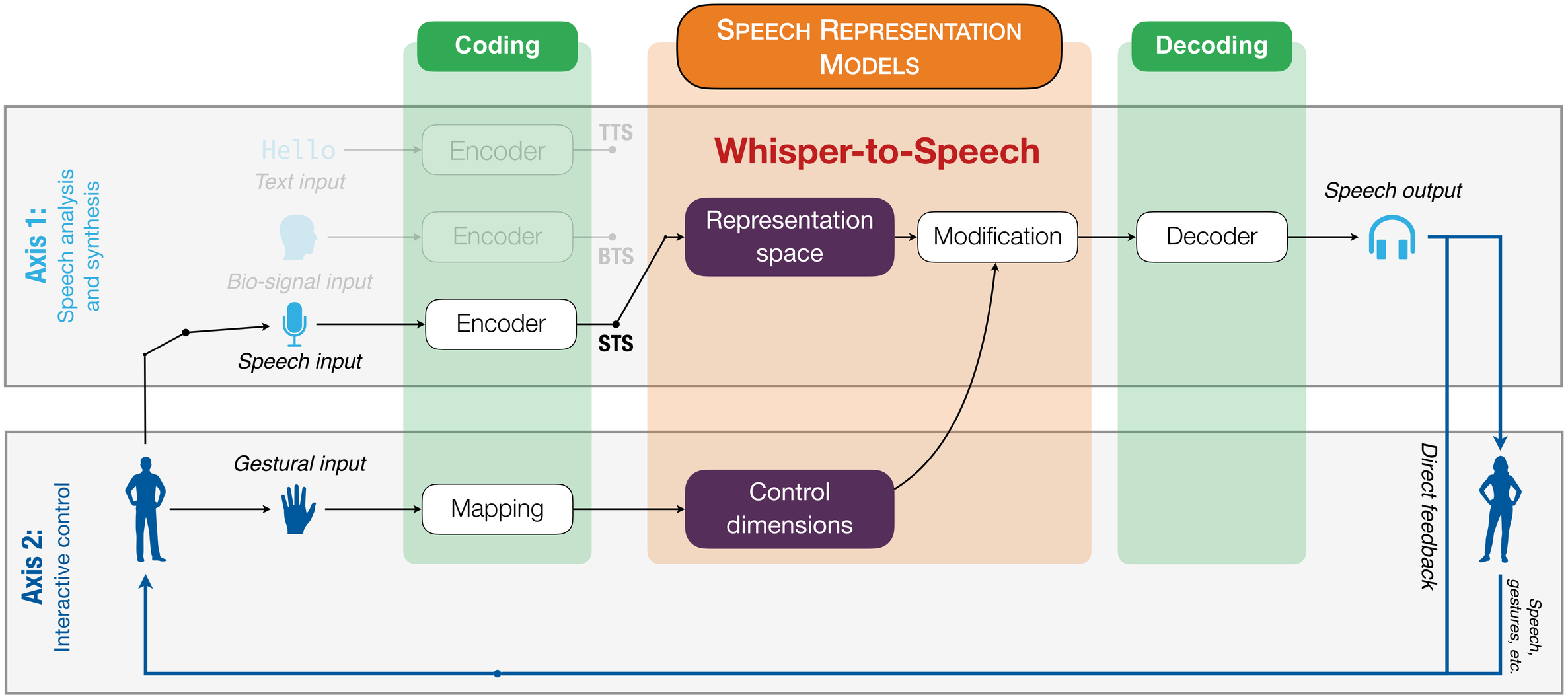
- Control of pre-recorded sound samples (intonation, rhythm)
  - Very accurate in **imitation** paradigms *d'Alessandro C. et al. (2011), JASA, pp. 1594–1604.*
  - Use multi-modal reinforcement for learning intonation of foreign languages (English, French, Mandarin) *Xiao X. et al. (2021), Proc. Interspeech, pp. 516–520.*

- What about **production** tasks with a **simultaneous control of articulation**?

Interactive control of synthesis

- Explicit control of F0 in singing
- From implicit to explicit control of F0 in speaking
- Towards a co-adaptation between human and machine learning of control mapping

Interactive control of expressive speech synthesis



Interactive control of expressive speech synthesis

Delphine Charuau

Control of implicit F0

*Implicit manual control of **contrastive focus** in a whisper-to-speech conversion set up*

- Dual control
  - Natural control of articulation (whispering)
  - Manual control of intonation with either a wrist rotation (accelerometer) or a finger pressure
  - Reference : natural voice

- Elicitation of contrastive focus
  - Simulated dialogue
  - No explicit instructions on fous

Hypothesis

Vous : - Lou du Mans a suivi le loup doux. ← No focus

L'expérimentatrice : - Lou du Mans a suivi le chat doux ?

Vous : - Lou du Mans a suivi le loup doux. ← Focus

- Corpora
  - 6 sentences of 9 syllables
  - 3 constituants : subject (3 syll)-verbal (3 syll) –object (3 syll)
  - Target syllabe cible [lu]

S1 : **Lou** du Mans a suivi le loup doux
O2 : Lou du Mans a suivi le **loup** doux
S2 : Le **loup** doux a suivi le beau loup
O3 : Le loup doux a suivi le beau **loup**
S3 : Le beau **loup** a suivi Lou du Mans
O1 : Le beau loup a suivi **Lou** du Mans

- 16 participants

Interactive control of expressive speech synthesis

Delphine Charuau

Control of implicit F0

Duration

- Measure
  - Duration of each syllable [lu] relatively to the utterance
- Hypothesis
  - Only Post-cible is focused
  - Effect on all controls
- Observations
  - Post-cible longer than all other syllables
  - Valid for all controls

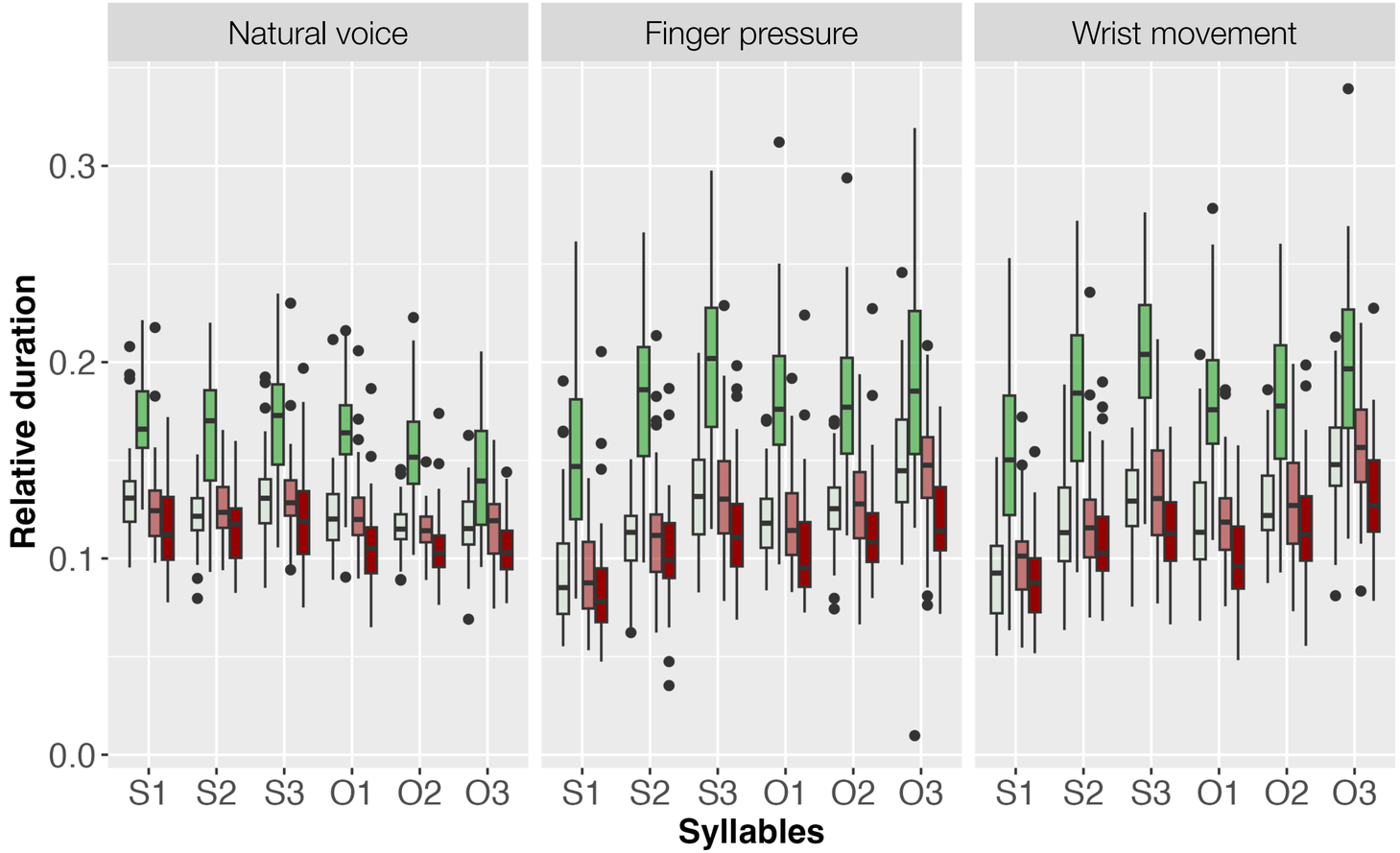
➡ Focus is produced with articulation

Vous : - Lou du Mans a suivi le loup doux.
Pre - non-cible — Pre - cible

L'expérimentatrice : - Lou du Mans a suivi le chat doux ?

Vous : - Lou du Mans a suivi le loup doux.
Post - non-cible — Post - cible

Condition
- Pre : Target
- Post : Target
- Pre : Non-Target
- Post : Non-Target



Interactive control of expressive speech synthesis

Delphine Charuau

Control of implicit F0

Intonation

- Measure
  - F0 centred per participant and interface
  - Peak F0 on all [lu] in semi-tones
- Hypothesis
  - Only Post-cible is focused
  - Effect on all controls
- Observations
  - Post-cible > all other syllables
  - Valid for all controls
  - Little variability on manual control

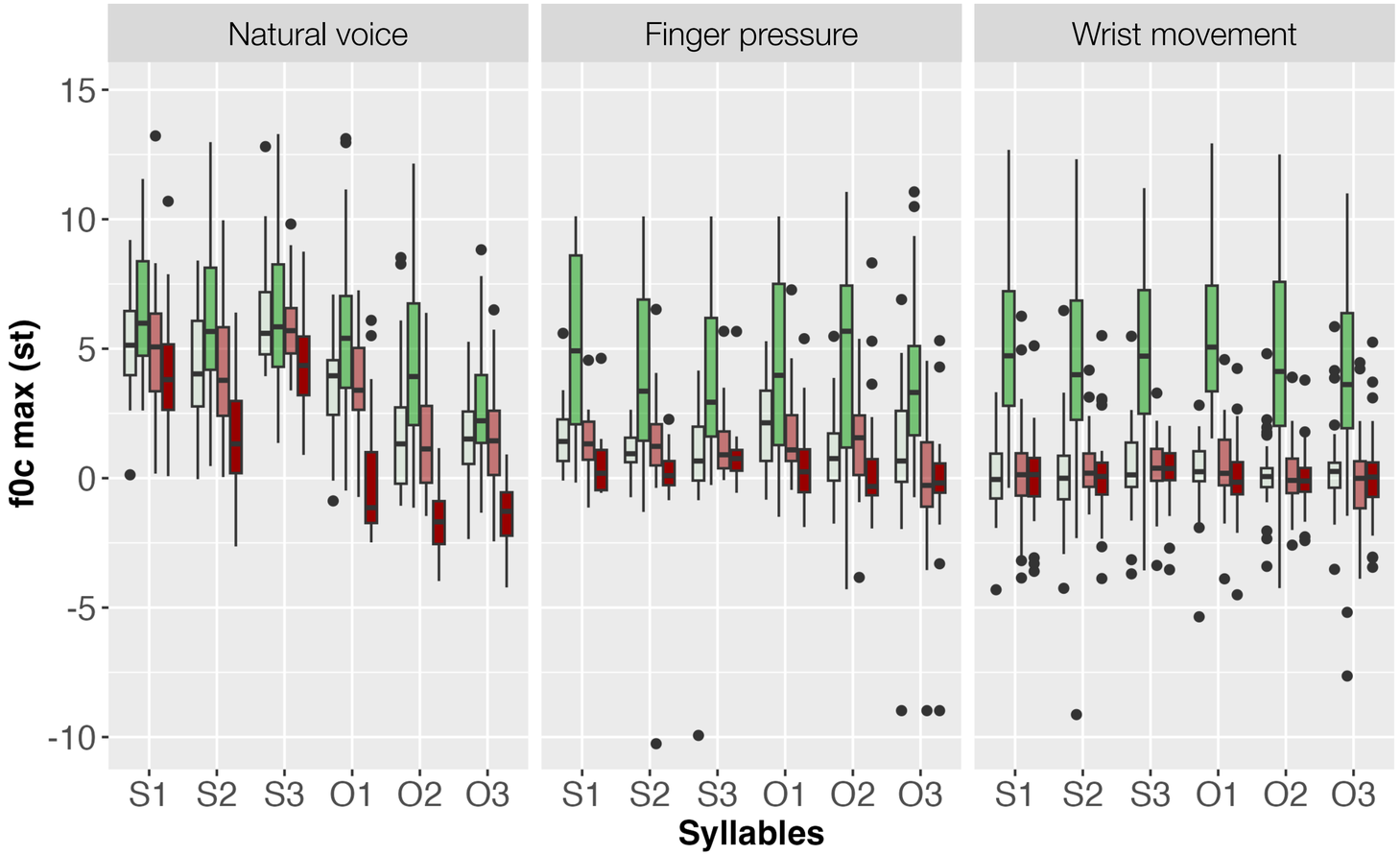
➡ Focus is also produced with the hand

Vous : - Lou du Mans a suivi le loup doux.
Pre - non-cible — Pre - cible

L'expérimentatrice : - Lou du Mans a suivi le chat doux ?

Vous : - Lou du Mans a suivi le loup doux.
Post - non-cible — Post - cible

Condition
- Pre : Target
- Post : Target
- Pre : Non-Target
- Post : Non-Target



Conclusions

- Control of focus
  - Contrastive focus was elicited (no explicit instructions to realise a focus)
  - All but one participants produced a focus at the right place (syllable elongation, raise of F0)

➡ Successful explicitation of F0 contour by participants

- Understood they had to produce some focus
- Knew that a focus is done by a local raise of F0
- Planned and produced the raise correctly (in terms of dynamics)

➡ Very encouraging for the explicit control of F0 paradigm

- Participants have the potential to learn such a control
- Learning curve quite long (many functions of F0 to learn)

Interactive control of synthesis

- Explicit control of F0 in singing
- From implicit to explicit control of F0 in speaking
- Towards a co-adaptation between human and machine learning of control mapping

Future work

Instead of having the human to fully learn the F0 control,
Can the system learn to predict F0?

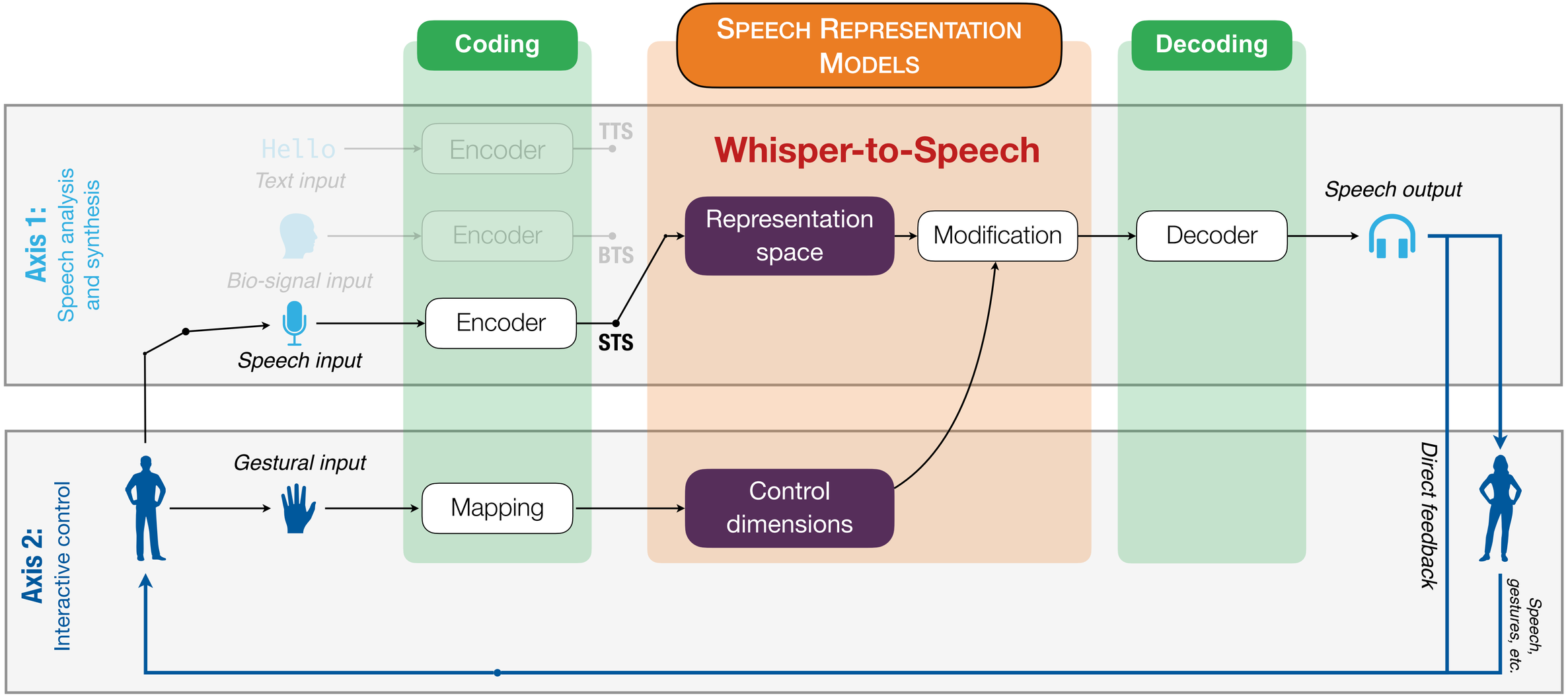
SilentPitch Project 2024-2027

- Strong correlations between intonation and gestures

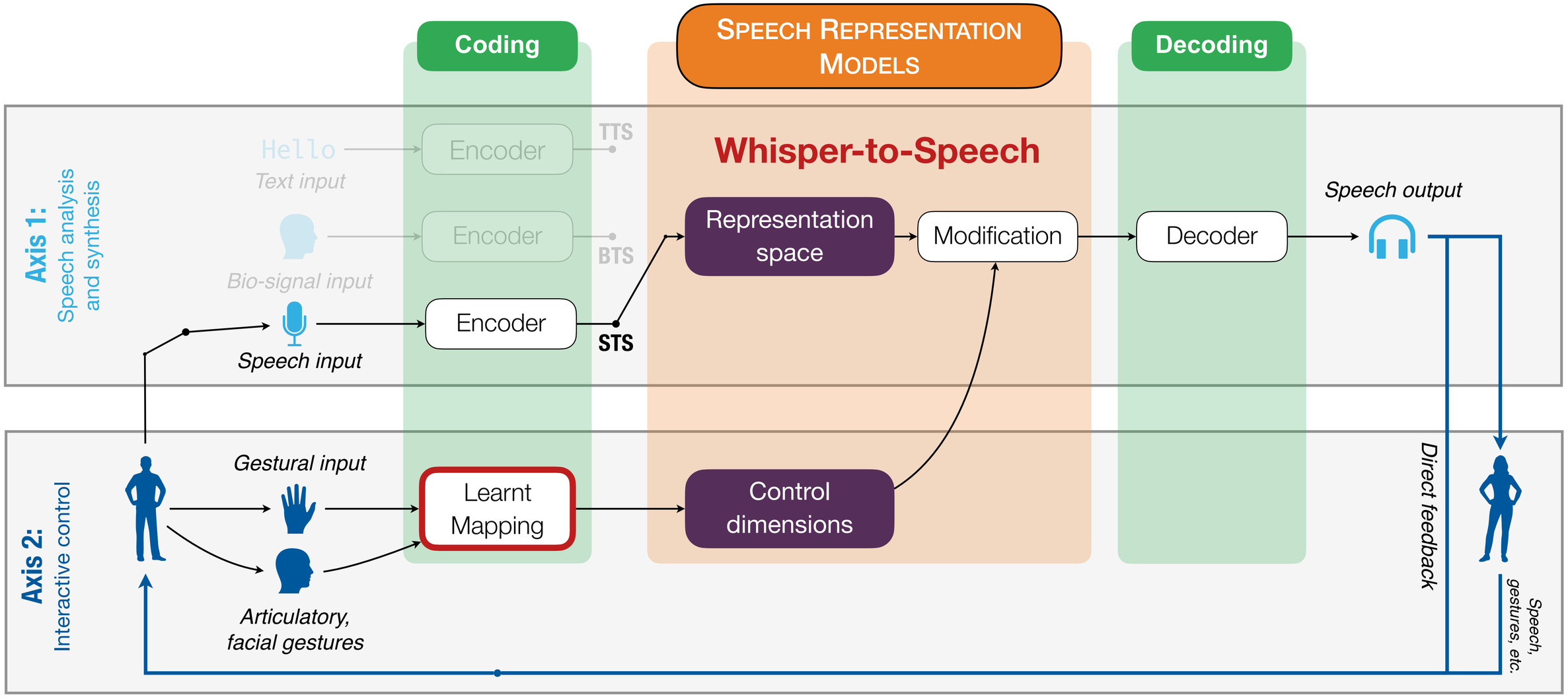
Fonction of prosody (not exhaustive)	Co-occuring speech gesture (not exhaustive)
Delimitative cues of prosodic phrasing (Fougeron & Keating, 1997)	Movements of: - the lips (Dohen et al., 2004), - the tongue (Krivokapić et al., 2017), - the eyebrows (Cave et al., 1996), - the head (Wagner et al., 2014), - the hands (Leonard & Cummins, 2011; Roustan & Dohen, 2010a, 2010b)
Focus (Jun & Fougeron, 2000)	
Social attitudes (Fónagy et al., 1983; Ward, 2019)	Intonation contour control (Evrard et al., 2015; Perrotin, 2015)
Emotions (Goudbeek & Scherer, 2010)	

➡ Can we use this correlations to automatically predict intonation patterns from those gestures?

Interactive control of expressive speech synthesis

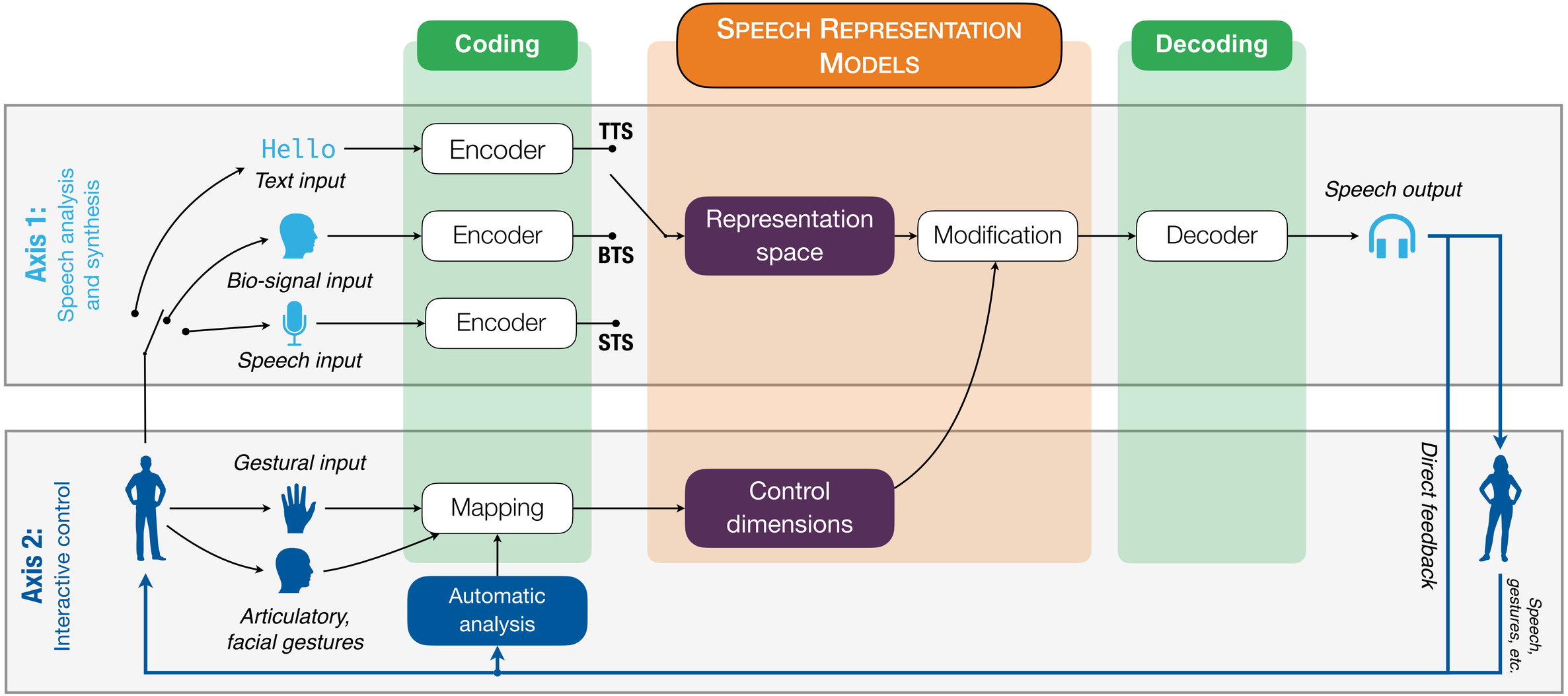


Interactive control of expressive speech synthesis



➡ Study the co-adaptation between human learning and machine learning in interaction

Interactive control of expressive speech synthesis



Interactive control of expressive speech synthesis

- Axis 1: Analysis-synthesis of **expressive speech**
  - Many applications (speech coding, voice transformation, text-to-speech)
  - Signal-based models: global understanding of acoustic variations linked to expressivity
  - Neural-based models: very powerful modelling tools that are often considered as opaque

➡ Using signal, acoustical, physiological knowledge to probe neural representation spaces open a highway to:

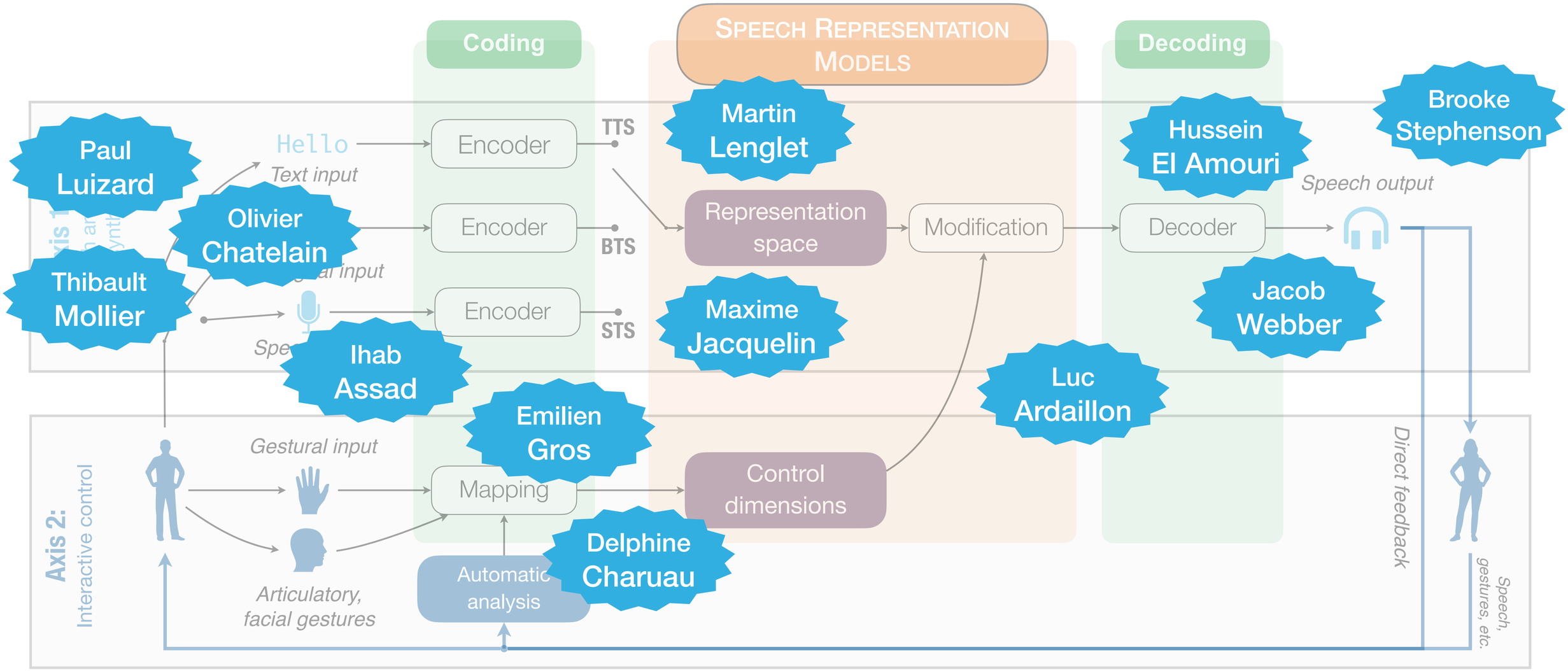
  - Complex and very fine-grained analysis of voice production on large datasets
  - High-quality synthesis in terms of expressivity rendering and control

- Axis 2: **Interactive control** of synthesis
  - Successful explicitation of F0 contour by participants in production tasks
  - Full intonation contour may have a long learning curve

➡ Exploitation of speaker and interlocutor data for a semi-automatic control of F0

  - Speech-to-speech: speech co-occurring gestures (co-adaptation between human and machine)
  - Text-to-speech: interlocutor behaviour

Merci !



Christophe d'Alessandro, Ian McLoughlin, Nathalie Henrich Bernardoni, Laurent Girin, Maëva Garnier, Rémy Vincent, Thomas Hueber, Gérard Bailly, *Et tous les autres*