AI-generated voices
Applications and implications

Spyros Raptis
Innoetics / Samsung Electronics Greece

Outlook

- **Beyond supervision**
  The shifting paradigm in voice generation
- **Problem areas**
  Selected areas in voice generation
- **Applications**
  How voice generation is already affecting the creative industries and our everyday lives
- **Ethics reshaping**
  Sings of a missing framework, emergence of novel questions and regulation being born

Beyond supervision

- Unsupervised learning
- Self-supervised representations
- Audio as language

- Multi-modal embeddings
- Huge datasets for a truly global coverage of spoken language

Beyond supervision

- Unsupervised learning
- Self-supervised representations
- Audio as language

- Multi-modal embeddings

- Huge datasets with universal coverage

Problem areas

- Word-level style control
- Phoneme-level prosody control
- Speaker generation

Style control

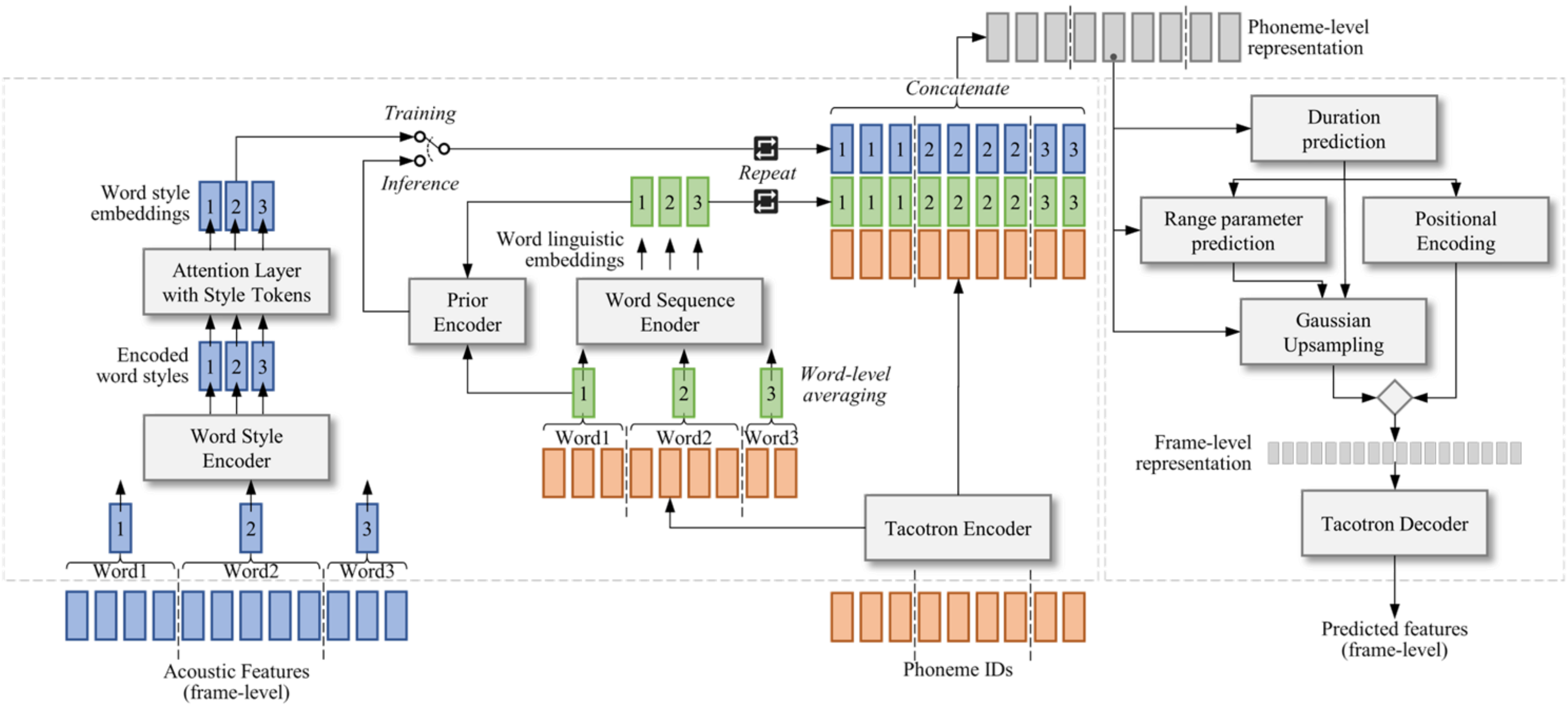
- “**Style**”:
  different qualities of the voice, depending on the use case

- **Style control**:
  control mechanisms baked into a model VS retro-fitted ex post

- **Granularity**:
  different levels relevant to different applications

- **Style palette**:
  let the model discover the styles inherent in the data VS imposing an externally defined taxonomy

Word-level style control



Word-Level Style Control for Expressive, Non-attentive Speech Synthesis

K. Klapsas, N. Ellinas, J. S. Sung, H. Park, S. Raptis

SPECOM 2021: International Conference on Speech and Computer

Style controllability

Style controllability is achieved by:

- manipulating the weights of the word-level style tokens

Unified and robust **control of token weights**:

- estimate the distribution of each token's weights in the training corpus;
- z-normalize it;
- apply changes to the token weights that are multiples of their standard deviation.

Level of control:

- Single word, multiple words, or the entire utterance

Experiments

Training dataset: Subset of the Blizzard Challenge 2013 audiobook dataset (single speaker, rich prosody)

Style token dims: 15 tokens x 128 each

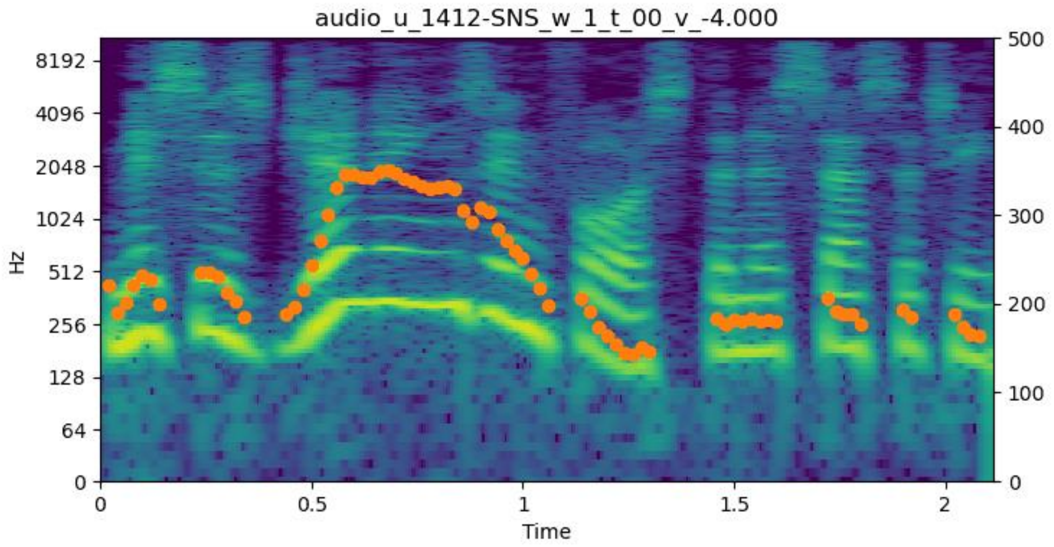
Vocoder: LPCNet

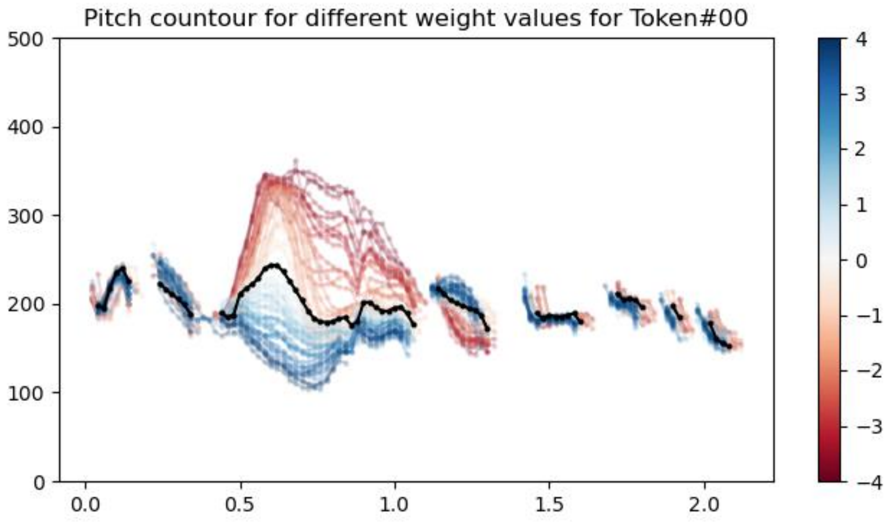
Observations:

- The model tended to generate **richer pitch patterns** than the plain NAT model
- Some of the tokens had simple intuitive **interpretations**:
  - some tokens were directly related to the pitch and some to the speaking rate;
  - For those tokens, decreasing their weight had the perceptually opposite effect of increasing it

Direct style manipulation at word-level

Word-level
Token 0 → Affects pitch

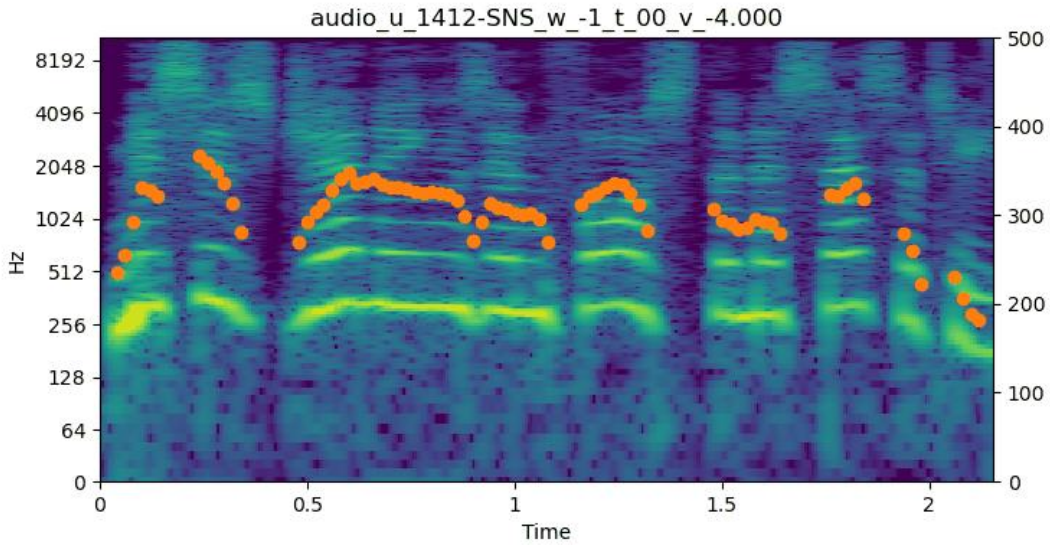


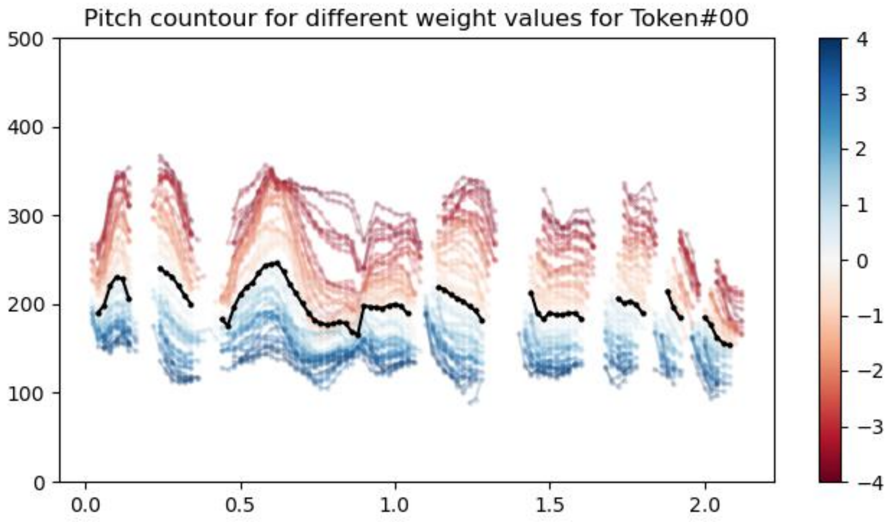


***1412-SNS** Mrs. Jennings enforced the necessity.*

Direct style manipulation at utterance-level

Utterance-level
Token #0 → Affects pitch

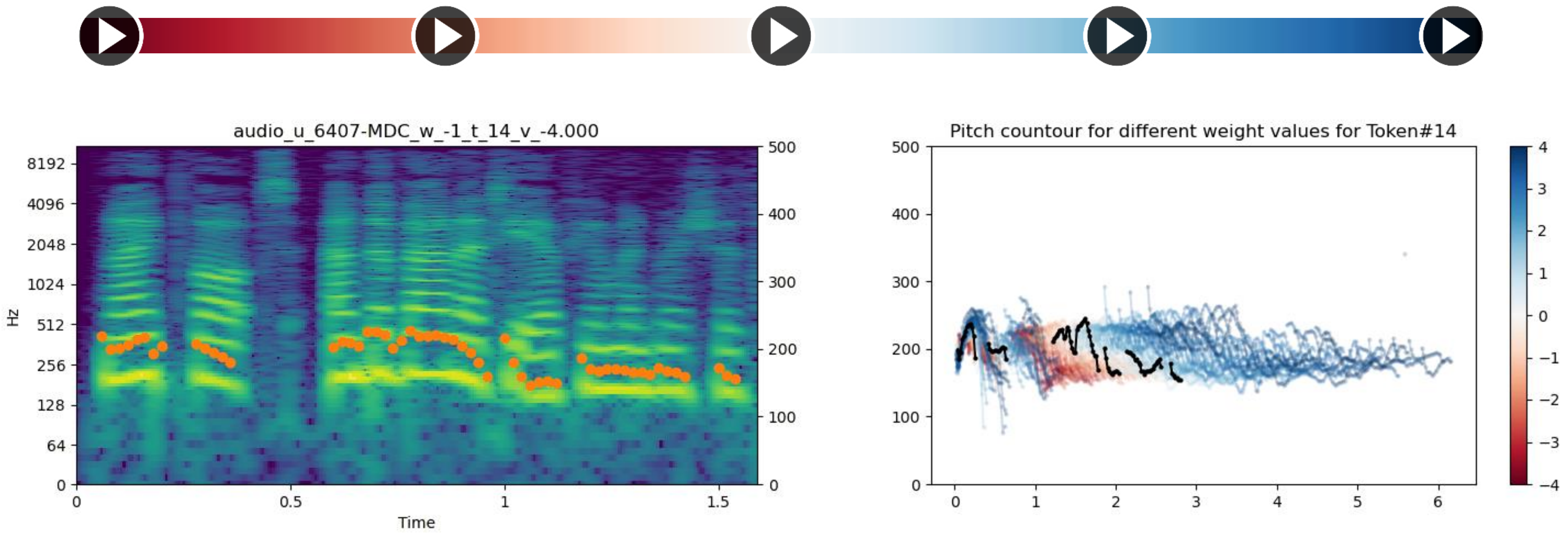




1412-SNS *Mrs. Jennings enforced the necessity.*

Direct style manipulation at utterance-level

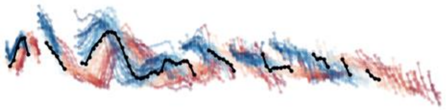
Utterance-level
Token #14 → Affects speaking rate



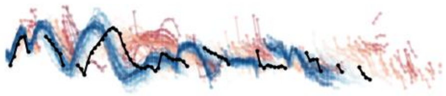
6407-MDC *Night Thoughts, and the Vanity of Human Wishes.*

Direct style manipulation at utterance-level

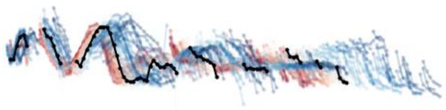
Token #01 | Acoustic conditions?



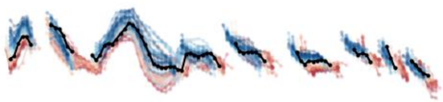
Token #03 | Effort vs relaxed?



Token #05 | Articulation?



Token #13 | Formal vs friendly?



Phoneme-level prosody control

- Unsupervised latent representations can capture speech variability, but:
    - the different qualities of speech are entangled and not amenable to our direct control

- However, in some use cases:
    - we do need to **control**;
    - at a **fine-grained** level;
    - with **discrete labels**.

- Approach:
    - condition at training time on features we care to control at inference time;
    - data augmentation;
    - within-speaker F0 normalization and speaker-independent F0 clustering;
    - balanced clustering for duration.

Dataset

- **Multi-speaker** dataset:
  - internal dataset (3 female + 2 male voices) → ~160h
  - the 2013 Blizzard Challenge voice (Cathy) → ~60h

- **Forced-alignment** to calculate phoneme boundaries
- Pitch- and duration-**augmentations** → improved robustness and value ranges
  - Pitch shifting (+- 3 semitones)
  - Time stretching (+- 30% of speaking rate)
- Extract phoneme-level values for pitch (average) and duration

Controllable speech synthesis by learning discrete phoneme-level prosodic representations
N. Ellinas, M. Christidou, A. Vioni, J. S. Sung, A. Chalamandaris, P. Tsiakoulis, P. Mastorocostas
Speech Communication, Vol. 146, pp. 22-31 (2023)

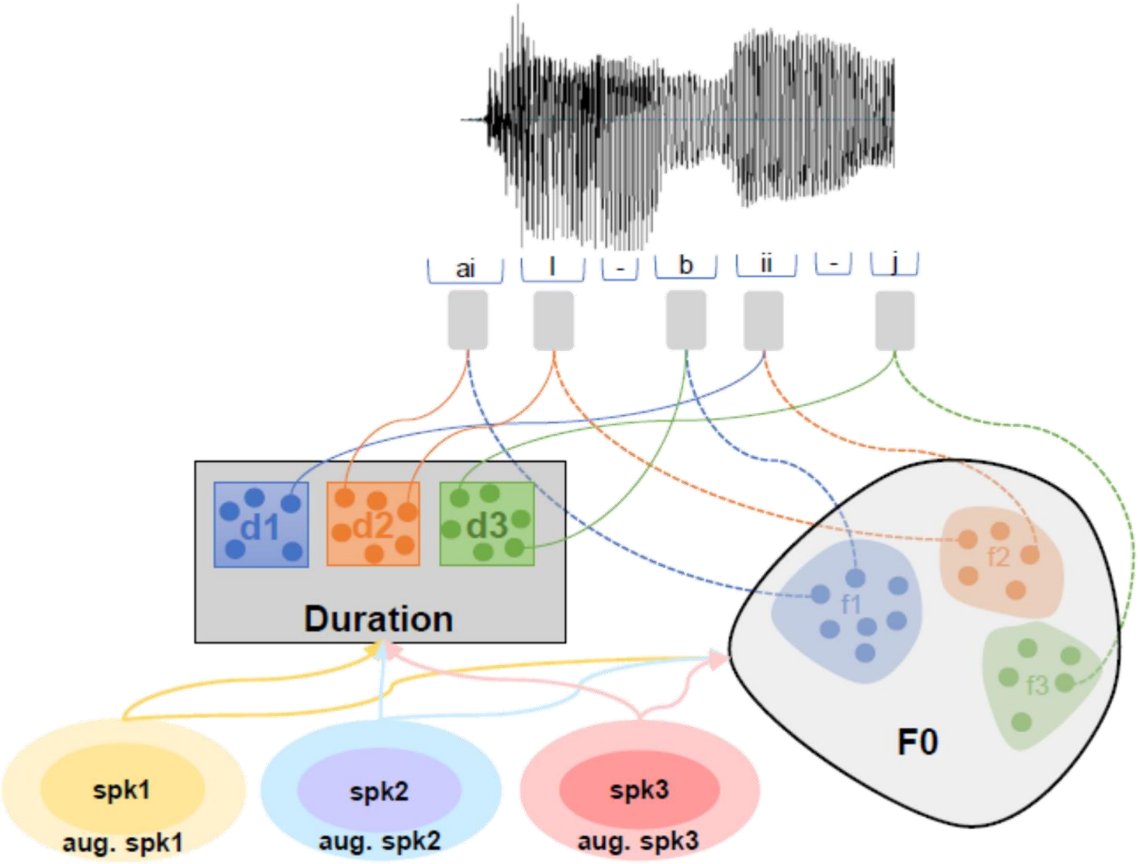
Pitch- and duration clustering

- **Pitch clustering** per speaker:
  - Average pitch values at phoneme-level
  - z-normalize with the speaker's mean and std
  - K-means clustering

  Helps deal with pitch range variations across genders/speakers, and facilitates the adaptation to new speakers

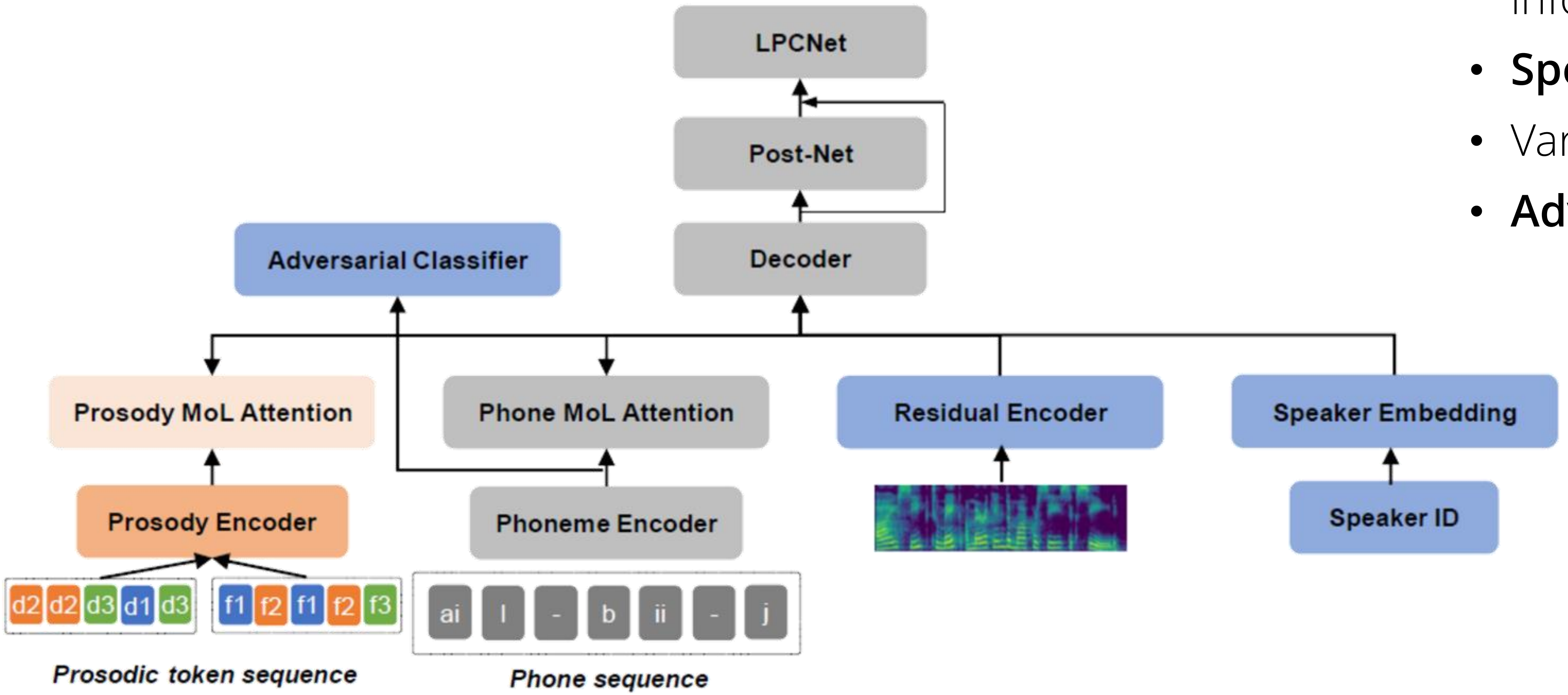
- **Duration clustering**:
  - Calculate the duration of each phoneme
  - Balanced clustering per phoneme class

  No duration normalization was necessary in this case.



Model training

Autoregressive attention-based model



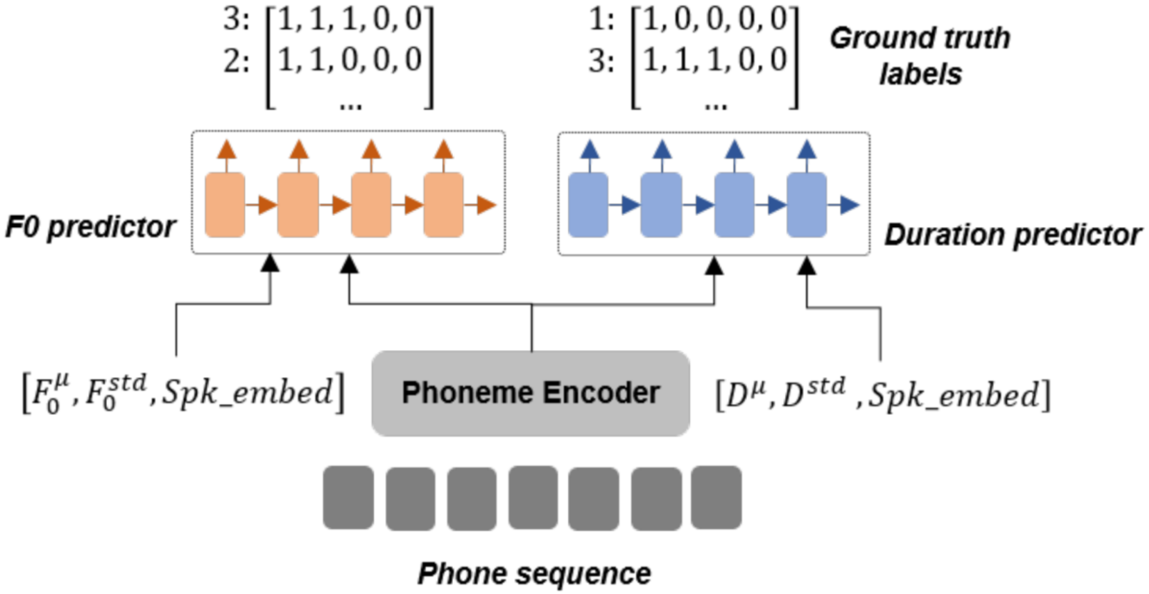
- Separate **MoL attentions** for phoneme and prosodic information
- **Speaker embedding** size: 64
- Variational **residual encoder**
- **Adversarial** speaker classifier

Controllable speech synthesis by learning discrete phoneme-level prosodic representations
N. Ellinas, M. Christidou, A. Vioni, J. S. Sung, A. Chalamandaris, P. Tsiakoulis, P. Mastorocostas
Speech Communication, Vol. 146, pp. 22-31 (2023)

Inference

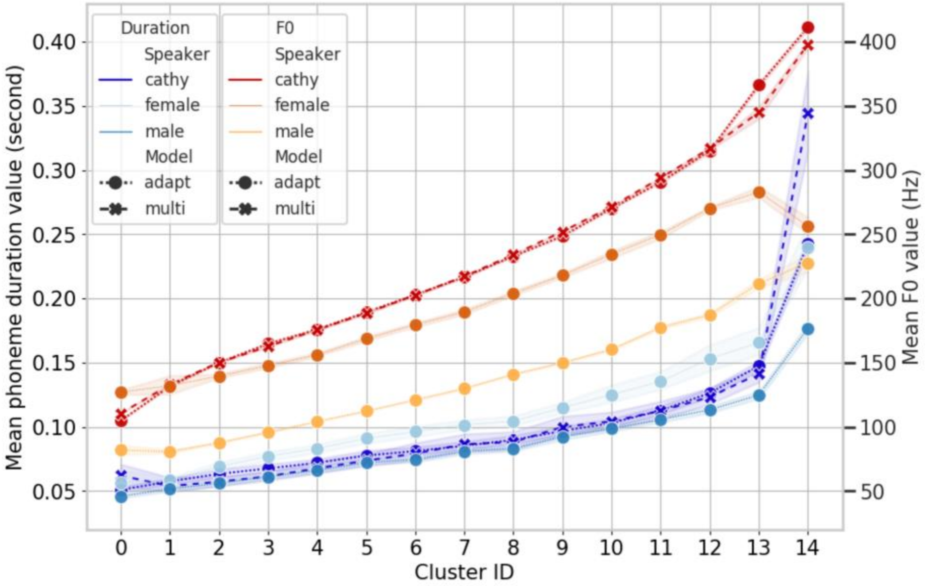
Prosody predictor

- trained separately ex post (phoneme encoder frozen)
- leverages the fact that the prosodic representations are:
  - discrete, and
  - ordinal



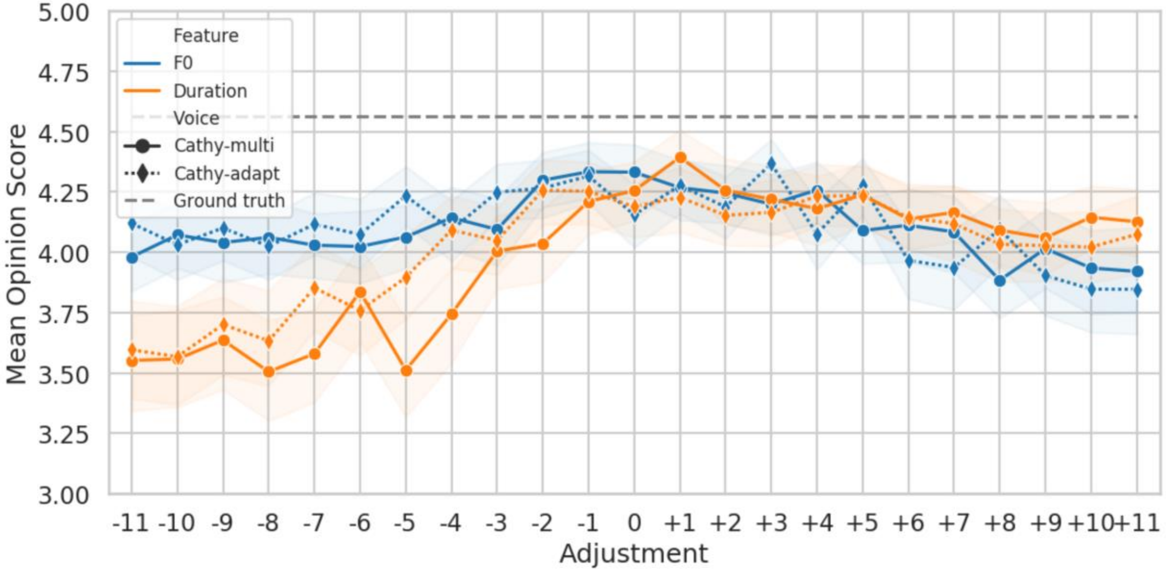
Controllability

Objective measures



- **x-axis**: prosodic category specified in input (for pitch or duration)
- **y-axis**: actual mean value of pitch (right) and duration (left) measured in the respective synthetic utterances generated

Subjective evaluation

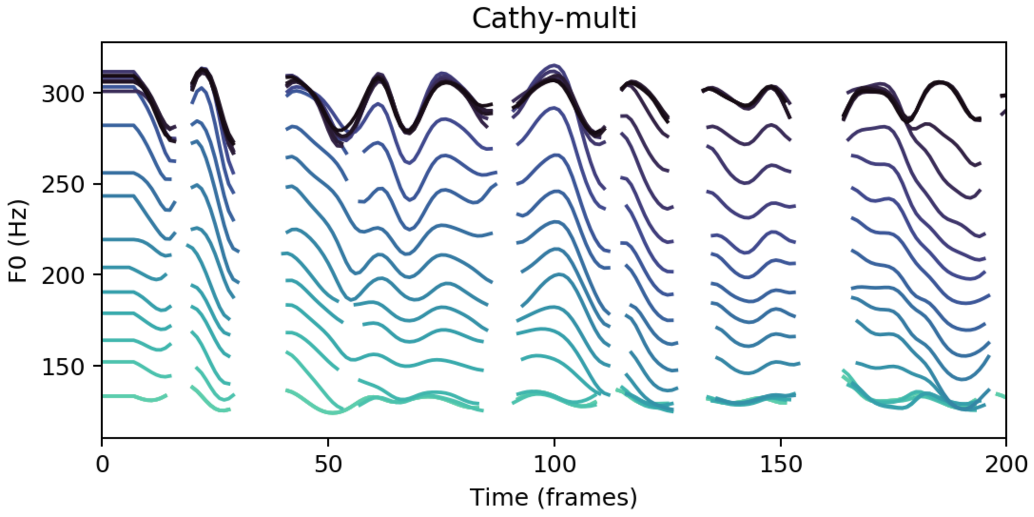


- **x-axis**: modification offset for pitch (blue) or duration (orange)
- **y-axis**: MOS score

Modifications: utterance-level

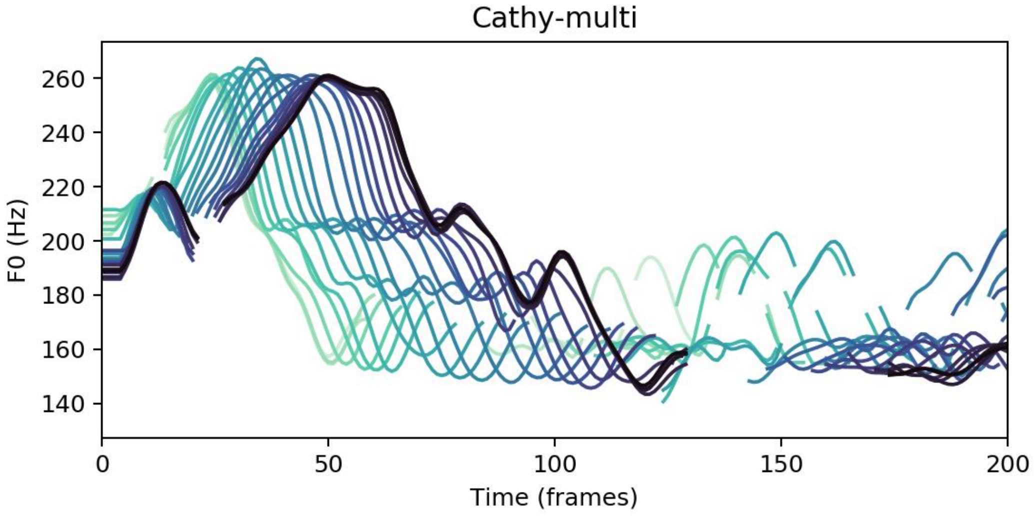
F0 modification
based on offset from ground-truth labels

He could see every object in his cottage and his gold was not there.



Duration modification
based on offset from ground-truth labels

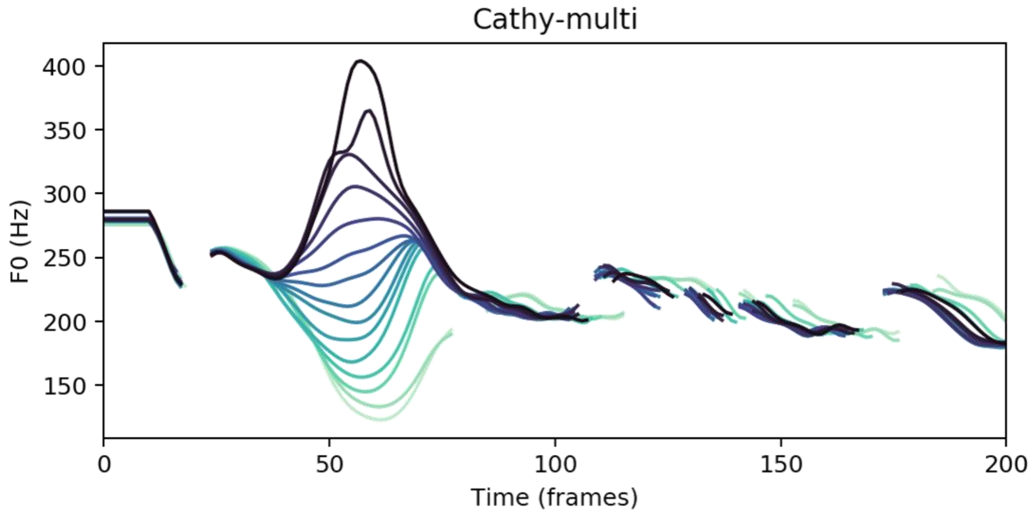
With lowered head he asked: Whered you go to?



Modifications: phoneme-level

F0 modification
based on offset from ground-truth labels

*To prolong (p r @ l **Q** N) and intensify the feeling he added...*



Adaptation to unseen speakers

Objectives:

- Adaptation to new, unseen target speaker with limited amount of data
- Maintain high quality
- Maintain the ability for fine-grained control

Process:

- Apply augmentation and z-normalization to the new speaker's data
- Fine-tune the model by replacing one of the existing speakers with the target speaker
- Even 5 minutes of target speaker's audio was enough

"His genius and ardour had seemed to foresee and to command his prosperous path."

Obama

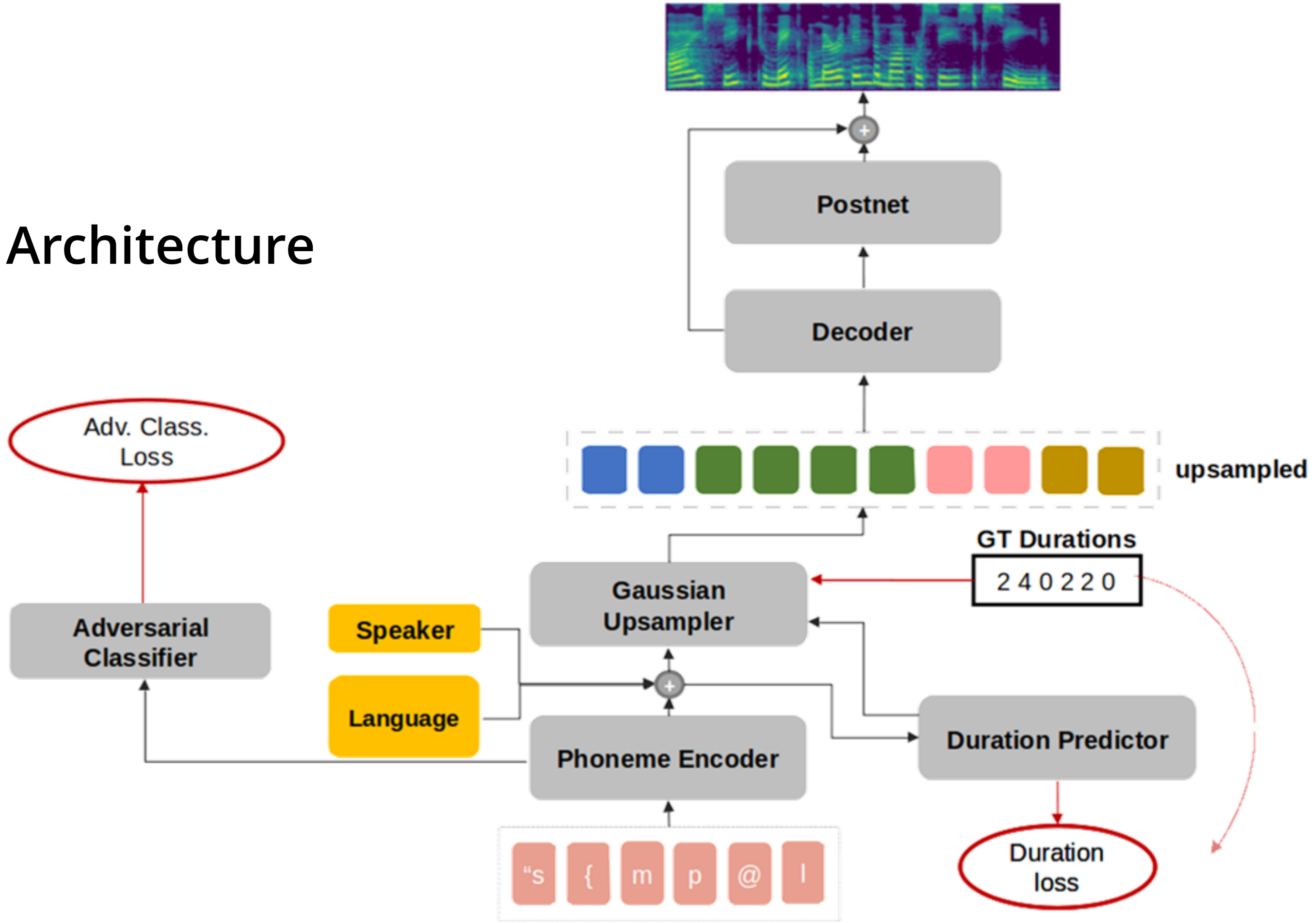
LJ Speech

Speaker Generation

- Multi-speaker TtS systems can closely imitate the voice color and style of the speakers in their training data
- The speaker identity representations that they learn correspond to real people.
- Extrapolate: generate **novel speakers** from multi-speaker / multi-lingual data
- Multiple approaches:
  - TacoSpawn: recurrent attention-based text-to-speech model that learns a distribution over a speaker embedding space
  - Transfer learning: d-vectors from speaker verification task as speaker representations for TTS
  - ...

Model and training data

Architecture



Training data

				speakers		
name	open	lng	hours	male	female	all
en96		en	342	56	40	96
LibriTTS [8]	✓	en	163	457	421	878
VCTK [9]	✓	en	25	46	62	108
ko87		ko	553	44	43	87
es8		es	96	4	4	8
de9		de	117	4	5	9
fr10		fr	95	4	6	10

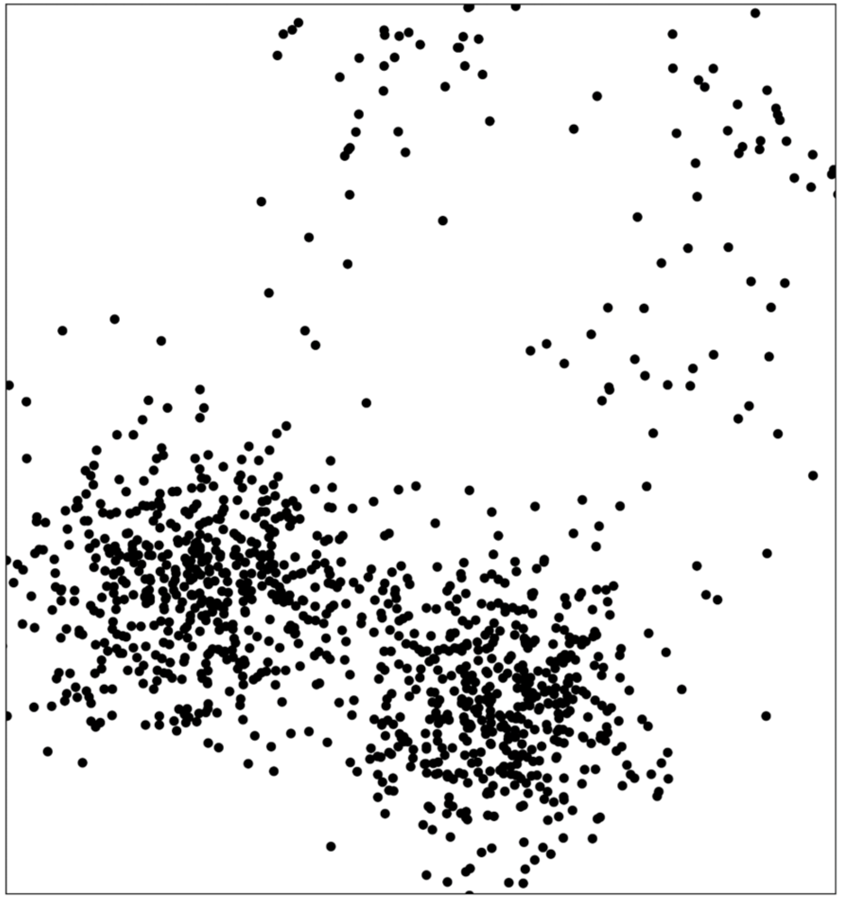
Generating Multilingual Gender-Ambiguous Text-to-Speech Voices

K. Markopoulos, G. Maniati, G. Vamvoukakis, N. Ellinas, G. Vardaxoglou, P. Kakoulidis, J. Oh, G. Jho, I. Hwang, A. Chalamandaris, P. Tsiakoulis, S. Raptis

Accepted to: Interspeech 2023

Embedding space

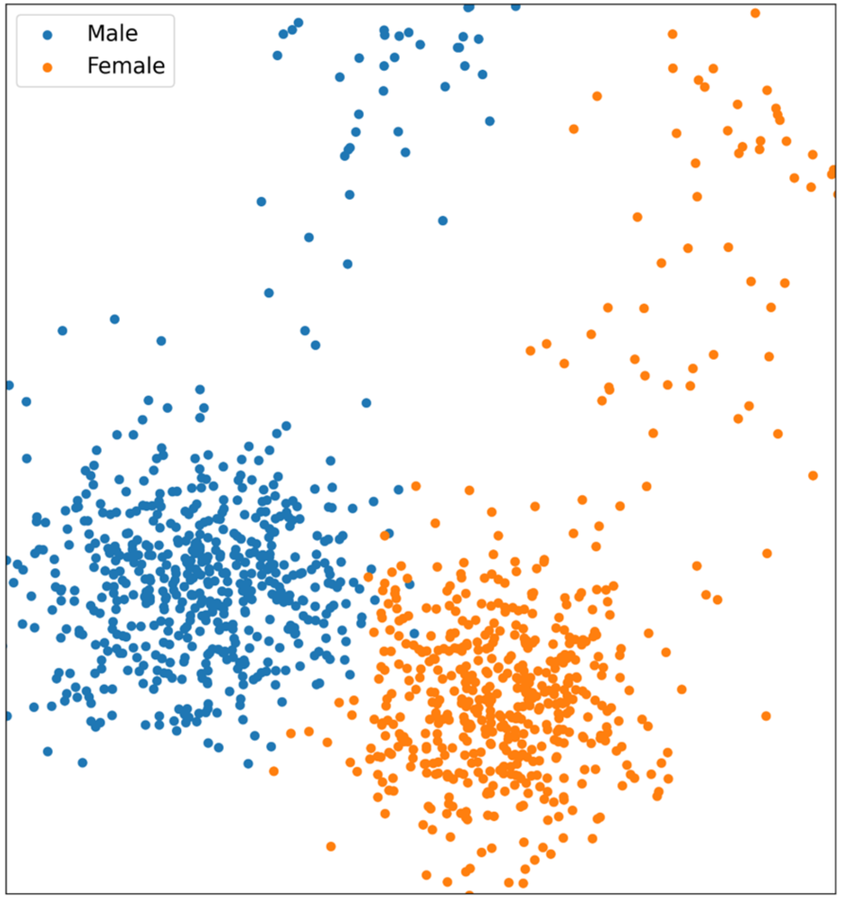
- Speaker embeddings: the "essence" of a speaker's voice (color, prosody, ...)
- "Similar" voices represented by nearby vectors
- 256-dim space
- PCA to reduce to 2D for display



Gender in the embedding space

- the first two PCA dimensions highly correlated to gender
- male and female speakers are almost linearly separable
- gender is one of the most important factors that explain the variance in the learned speaker embedding space

- additional sources of variation, e.g. acoustic conditions, recoding equipment, …

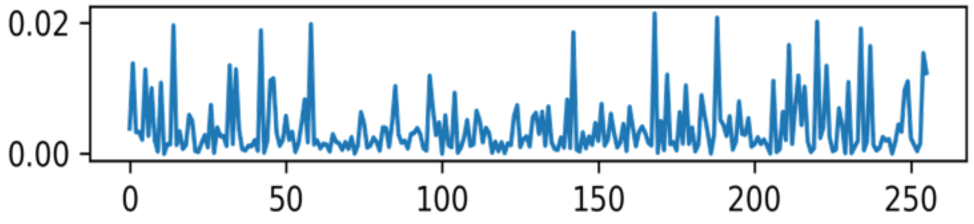


Gender in the embedding space

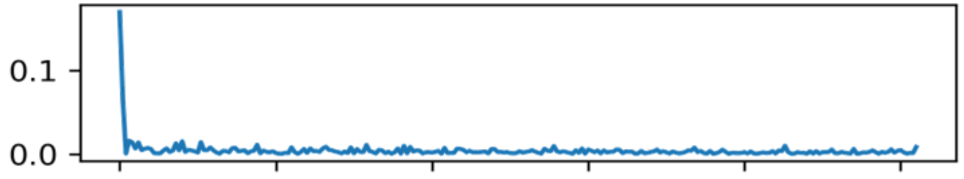
- How much part of the gender information is captured in the first 2 dimensions?
- Correlation ratio:
  η = the weighted variance of the mean of each category (male/female) over the variance of all samples

- Gender information spread across dimensions in the original space, but concentrated mainly in the first 2 in the PCA space

Original embedding space



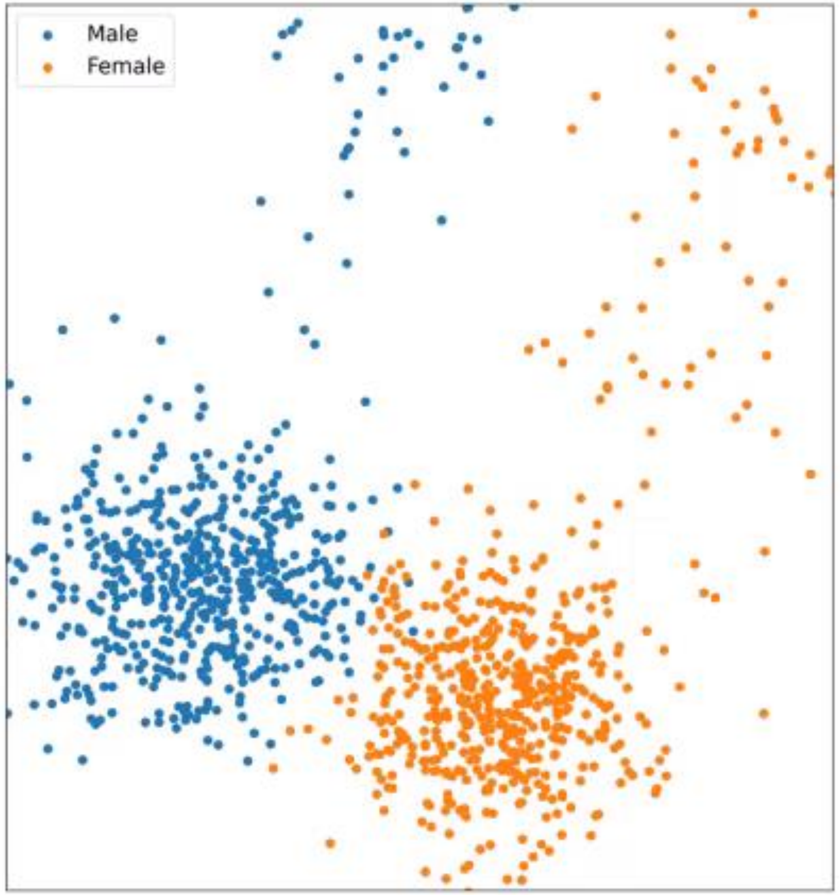
After PCA



-- embedding dimensions →

Generating novel speakers

- All sources of variation (aside from linguistic content) entangled in the embedding space
- Point in the embedding space → plausible novel speakers "similar" to their neighbors
- First two PCA dimensions capture most speaker variability, so it makes sense to use these to guide the sampling
- How do we recover the rest of the dimensions?
  - Assume 0 for all the rest in the PCA space (i.e. assume their mean value); or
  - Find the closest male and female speaker and perform weighted interpolation



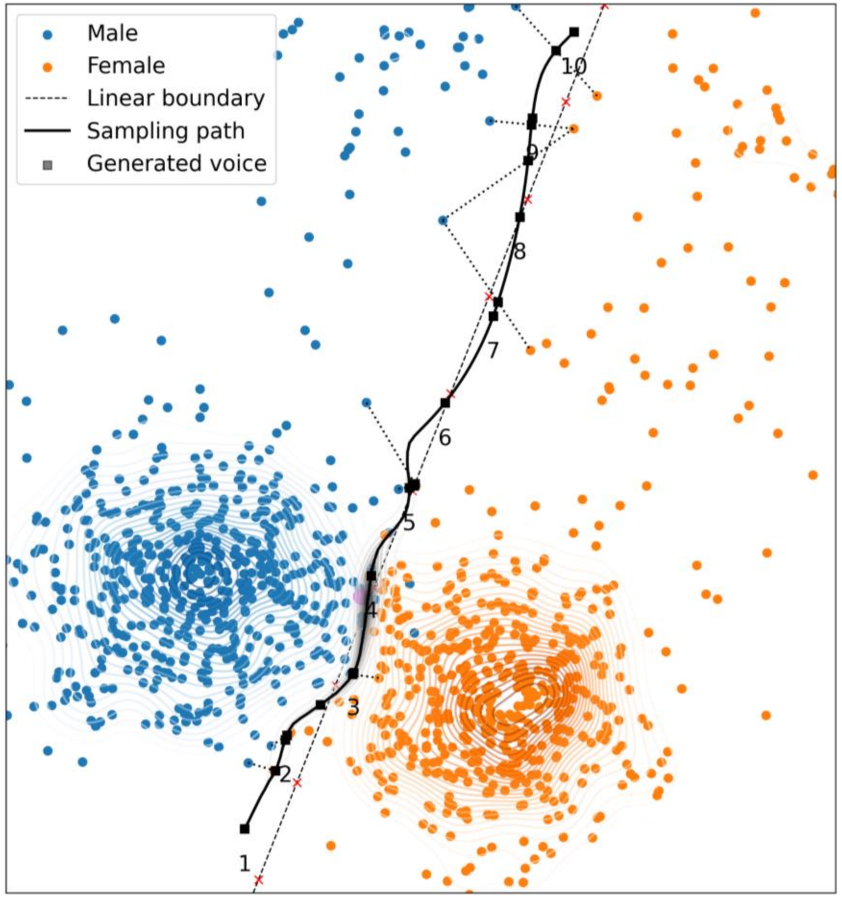
Browsing the embedding space

“ *This 48th ICASSP is the first post-pandemic edition, celebrating the return to an in-person experience and the 75th anniversary of SPS. We are looking forward to welcome back the whole signal processing community in a single venue, after three very challenging years. ICASSP’s main theme this year is “Signal Processing in the AI era,” promoting the creative synergy between signal processing and machine learning.* ”

Gender perception in generated voices

- Gender information is very prominent in the first 2 PCA dimensions, so we can experiment on:
  - How amenable is voice gender to our control?
  - How is voice gender perceived across different demographics?

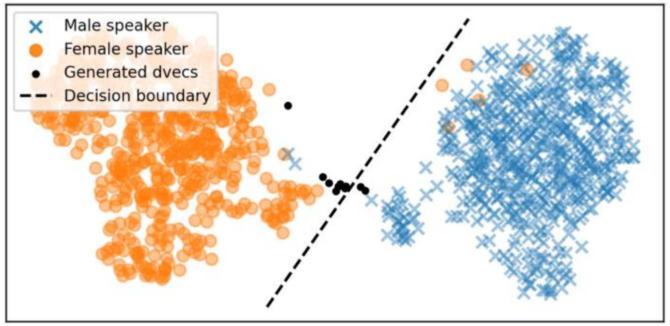
- Process:
  - **estimate the density** of male and female speakers;
  - find the **boundary area**: where male and female densities are relatively high and comparable
  - **sample** from the boundary and generate the respective speaker embeddings;
  - run **objective metrics** to measure gender ambiguity, speaker diversity, voice consistency
  - perform **listening tests** on samples synthesized by the corresponding generated speaker embeddings with subjects from different demographics



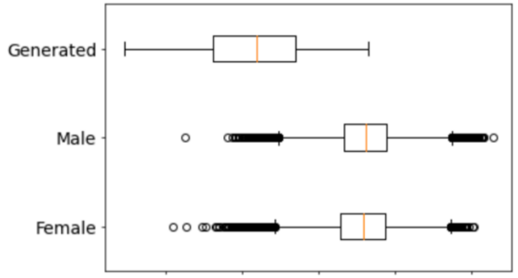
Gender perception in generated voices

Objective metrics

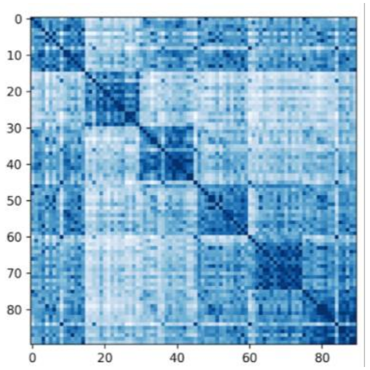
- **Gender ambiguity**.
  Generated utterances on the d-vectors space (UMAP on 2D).



- **Speaker diversity**.
  Diversity in the original and generated voices (distances in the d-vectors space)



- **Voice consistency**.
  Distance matrix of sentences synthesized by different generated speakers in different languages

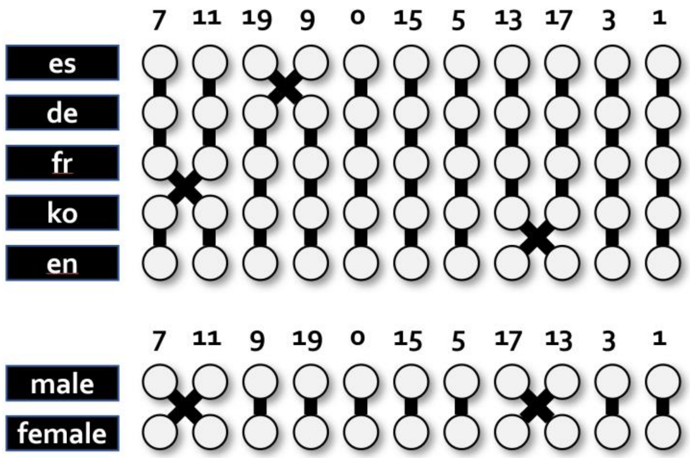


Gender perception in generated voices

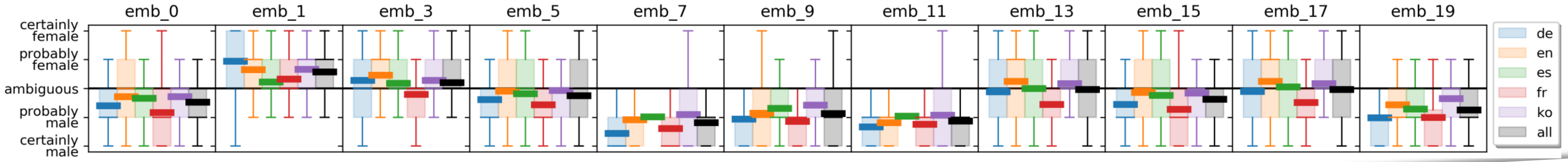
Subjective listening tests

- MOS scores listening tests: naturalness and gender perception
  - English , Korean, Spanish, French, German
- Subjects:
  - 114 subjects, 15432 ratings for gender perception
  - 102 subjects, 14136 ratings for naturalness.
  - 58.4% males, 31.7% females and 9.9% of undisclosed gender.
- Main conclusions:

  The ordering of systems' perceived gender (more/less male/female) is largely consistent across both dimensions.



So, gender perception seems to be shared among listeners of different gender and different native language



Applications

Applications

An increasing supply of AI-backed voice generation technologies and platforms with a significant impact on...

...the **creative industries**:

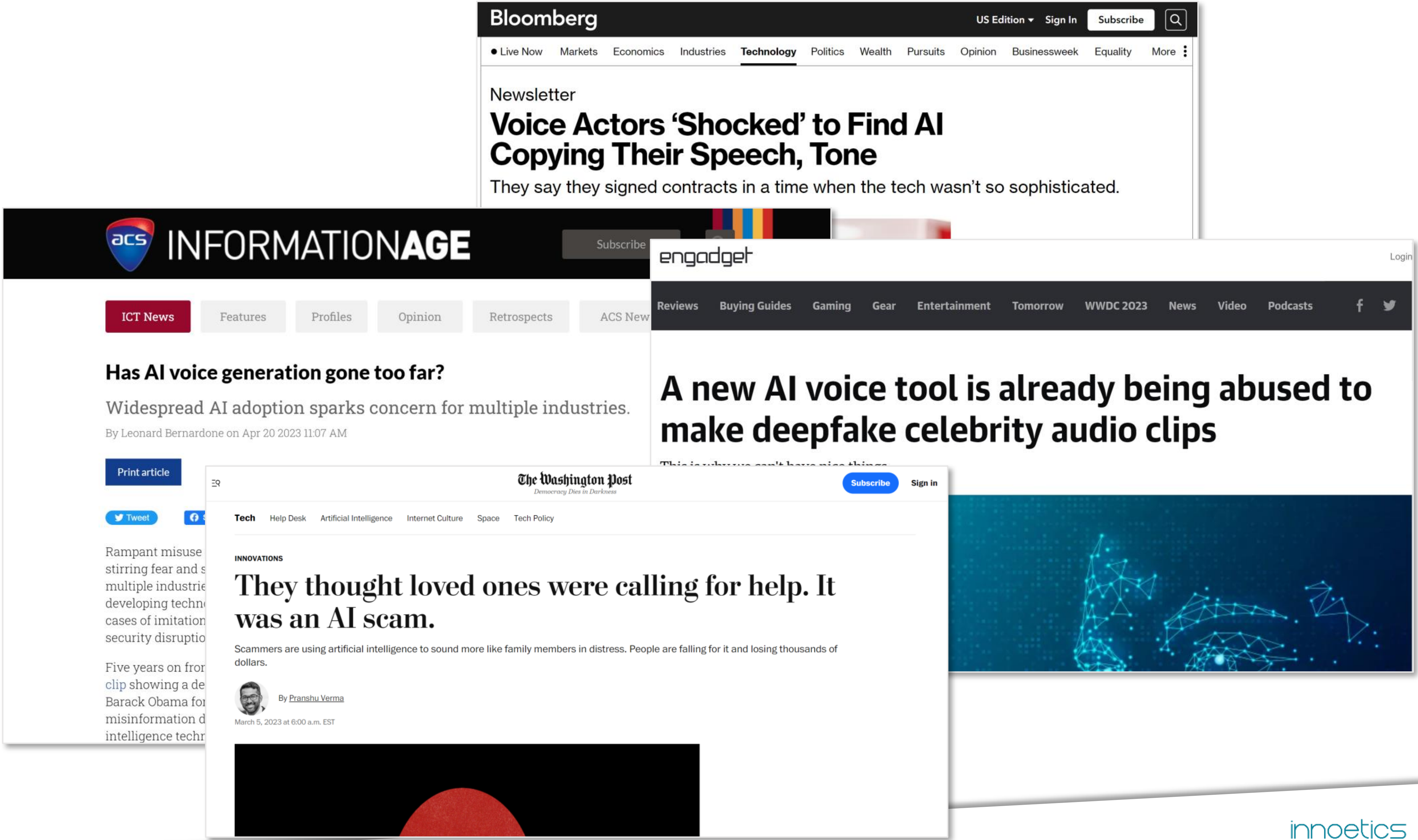
- Audiobooks
- Post-production for cinema
- Post-production for user generated content
- Voice cloning and personalized voices
- Gaming
- Singing
- ...

...our **personal lives**:

- Accessibility
- Education and language learning
- Personal communication

■ Ethics reshape

Sings of a missing framework



Regulation forming

- **AI Act**: draft EU Directive on AI applying to the development, deployment, and use of AI in the EU or when it will affect people in the EU.
  - Regulates **applications and the use of technology** (not the technology itself)
  - Adopts a risk-based approach:
    - unacceptable risk,
    - high risk,
    - limited, or
    - minimal risk
  - Example of "unacceptable risk" in the EU's site:
    "All AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned, from social scoring by governments to toys using voice assistance that encourages dangerous behaviour."
- US congressional hearing on "Oversight of AI"
- Digital "likeness"

Voice is special

- Voice is **trust**
- Voice is **emotional**
- Voice is **affective**

Thank you

Spyros Raptis
Innoetics | Samsung Greece

s.raptis@samsung.com
/in/spyros-raptis