

Emotags: Computer-Assisted Verbal Labelling of Expressive Audiovisual Utterances for Expressive Multimodal TTS

Bailly G., R. Legrand, M. Lenglet, F. Elisei,
M. Garnier & O. Perrotin

GIPSA-Lab

Grenoble-Alpes Univ. & CNRS, France

LREC-COLING, Torino, Italy, 2024

Test Audio

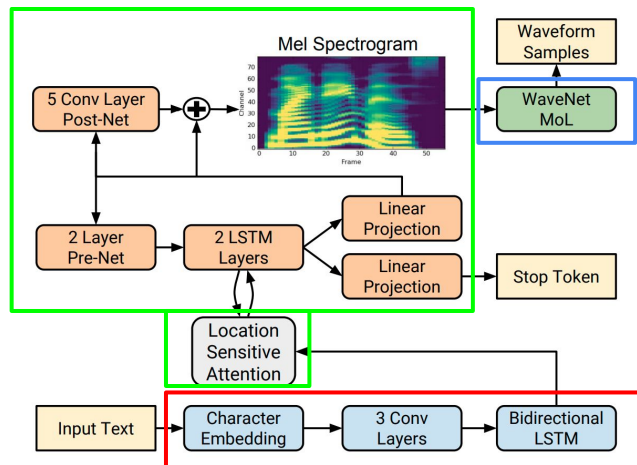


Technologie TTS actuelle

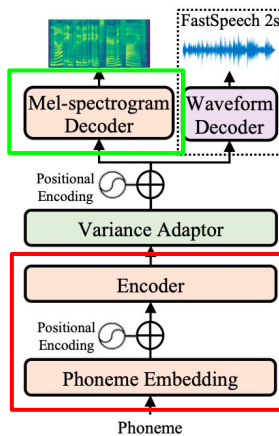
- Impressive quality e.g. FastSpeech2+Hifigan trained on 70 hours of French audiobooks
- Text2Spectrogram : **encoder-decoder** model
 - Tacotron2, Fastspeech2...
- Spectrogram2Signal-signal : **vocodeur neuronal**
 - WaveNet, WaveGlow, wavRNN, hifiGAN, waveGAN, LPCNet...



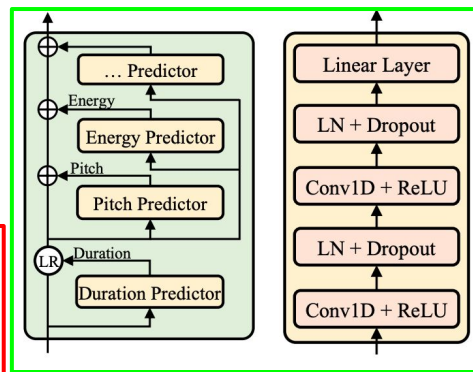
[Ren & al, ICLR 2021]



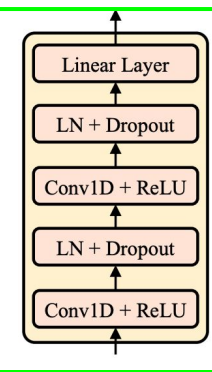
[Shen & al, ICASSP 2018]



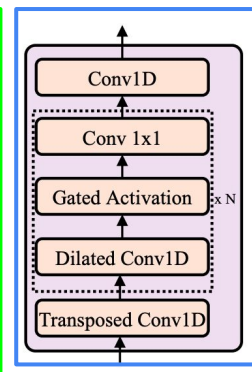
(a) FastSpeech 2



(b) Variance adaptor



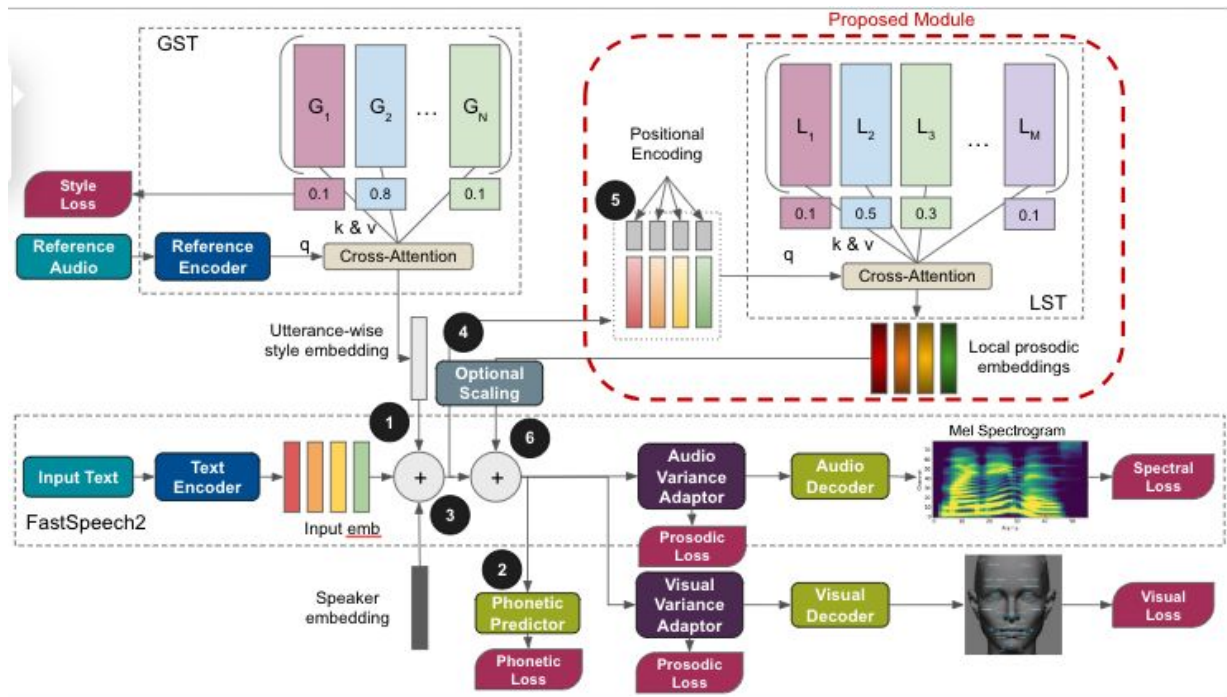
(c) Duration/pitch/energy predictor


















(d) Waveform decoder

Handling expressivity

- Reference encoder: summarizing the variance of the utterance not yet explained by the text
- GST: structuring this variance with tokens
- Crossentropy with labels if any...
- LST: handling the coordination of the contours with the text
- Added to the output of the text encoder for every input symbol as for speaker embeddings



GST: expressivity similar to speaker embeddings

Sentence	NEB	DG	AD	IZ	RO
Bonjour à tous. Je suis Suzie, l'avatar expressif de l'application Théradia.					
J'ai préparé une nouvelle séance d'exercices pour vous. Vous êtes prêt ?					
Félicitations ! Vous avez obtenu de super résultats ! On se revoit la semaine prochaine à la même heure ?					

























Expressive audiovisual corpus

- Professional comedian (AD)
 - 12 attitudes (contextual prompts + seeds) + narrative style
 - Contracted for 100 hours : 30 hours of useful speech
 - Extracts from SIWIS utterances (parliament & books)
 - Tracked by Dynamixyz® performer and mapped onto an avatar in RT



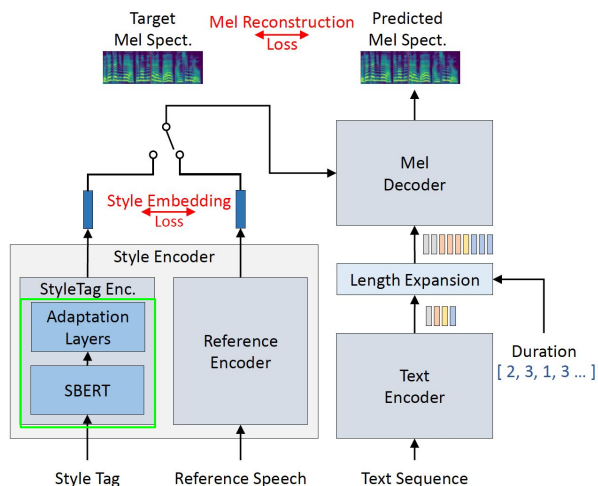
Expressivity by GST

- With DIRAC distributions of GST weights
 - Weight corresponding to each expression set to 1, others 0
- Instructed emotions can be mixed but
 - how to set up weights to reach a targeted emotional nuance?
 - how to cope with variability of training performances that can already be emotional nuances?
- Scientific challenge
 - Explain variability around prototypes
 - Not using weights (what is 80% enthusiasm + 20% skepticism?)
 - But using verbal input (eg. stupefaction, daze)

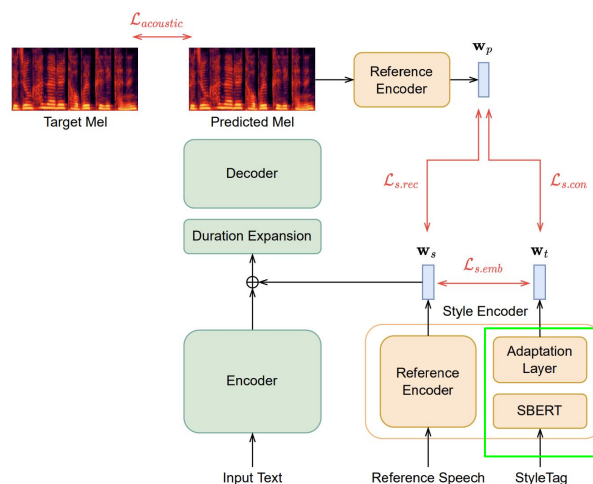
Angry		
Sorry		
Committed		
Enthusiastic		
Playful		
Surprised		
Obvious		
Skeptical		
Thoughtful		
Comforting		
Pleading		
Narrative		

From GST weights to verbal control

- Style encoder trained jointly by
 - Signal-based reference encoder
 - Verbal-based encoder driven by SBERT-encoded emotional tags
- Massive collection via crowdsourcing
 - 5 000 unique verbalizations in Shin et al (2022)



[Kim & al, Interspeech 2021]



[Shin & al, Interspeech 2022]

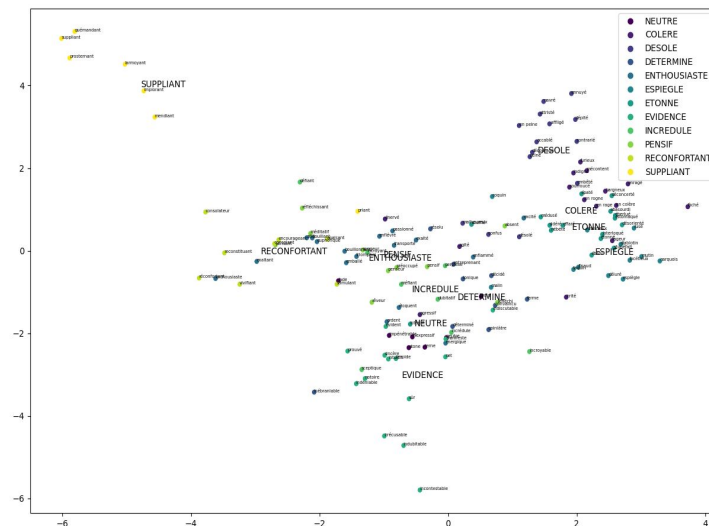
Emotags: scientific challenge & improvements

- Collect massive data
 - Combining free and semi-directed selection
- Explain variability around instructed emotions
 - Associating GST weights with lexical nuances (stunned, amazed, overwhelmed...)
- In French and audiovisual!!
- Beyond basic emotions
 - 12 attitudes



Building an emotional space

- Data
 - 12000 audiovisual clips of AD performances
 - 12 instructed attitudes
- Emotional space
 - A priori selection of 132 synonyms
 - A dozen for each of 12 attitudes
 - LDA of Large Flaubert embeddings
 - 1024 to 11 discriminant axis



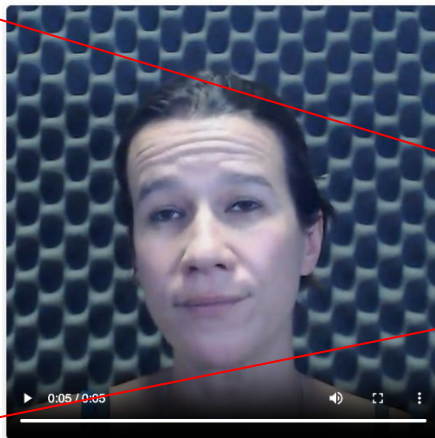
Collecting tags (free OR semi-directed selection)

- Incremental suggestions via navigation in the emotional space
 - 12 seed adjectives + “other”
 - If “other”, 6 most distant adjectives from these already suggested
 - otherwise, 6 closest adjectives from the selection
- Direct selection or free input

Reconnaissez-vous cette attitude ?

VIDÉO

- ➡ Installez-vous dans un endroit *calme* et *silencieux*.
- ➡ Regardez et écoutez attentivement la vidéo autant de fois que vous le souhaitez.



➡ Pour lancer la vidéo, cliquez sur la flèche ou pressez la touche ESPACE (sous Safari TAB puis ESPACE).

ÉVALUATION

- ➡ Indiquez au moins **2 attitudes** correspondantes ou sélectionnez en parmi celles proposées.
- ➡ Validez votre choix pour passer à la vidéo suivante.

- Explorez des propositions en choisissant l'attitude la plus proche parmi les suggestions suivantes.
 - Cliquez sur "Autres" pour découvrir plus de propositions.
- La liste est mise à jour en fonction de ce que vous avez déjà sélectionné.

Certaine Indéniable Inébranlable Limpide Nette Notoire Autre ...

- Ou recherchez directement dans le menu déroulant une attitude correspondante.
- Vous pouvez également ajouter la votre si vous ne la trouvez pas.

Sélectionnez ou Ajoutez ▾

Votre sélection :

Cliquez sur une attitude pour la supprimer de la liste.

Effacer tout

Evidente Indubitable Bien_sûr

Valider

Results

- https://gricad-gitlab.univ-grenoble-alpes.fr/web/emotags-results/-/blob/main/analyse_prolific.ipynb
- Up to now, 794 annotators
 - 8988/12461 – 72.13% tagged utterances
 - 119,275 tags in total
 - 13,02 suggested (using 4,43 “other”) vs. 0.25 free tags per clip
 - 132 suggested vs. 320 free (count >2) adjectives

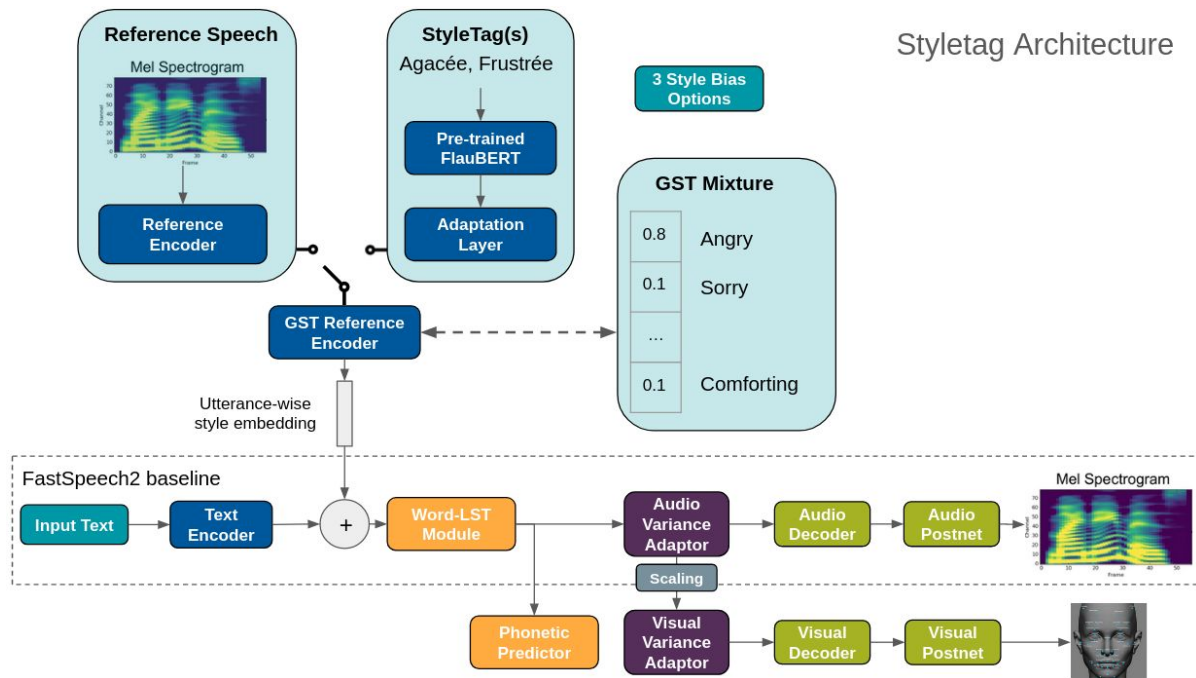
Predicting acoustic embeddings

- Weighted GST, original FDA space vs. emotion-specific PCA
 - Significant part of the variance explained by verbal tags

Style	GST	LDA	PCA	Count	Style	GST	LDA	PCA	Count
Angry	.80	.70	.74	401	Obvious	.71	.58	.61	446
Sorry	.91	.71	.80	353	Sceptical	.72	.58	.62	460
Committed	.68	.65	.69	325	Thoughtful	.78	.70	.74	399
Enthusiastic	.79	.71	.75	439	Comforting	.86	.74	.75	433
Mischievous	.72	.63	.69	253	Pleading	.81	.69	.71	518
Surprised	.76	.67	.69	395	Narrative	.77	.59	.51	4566

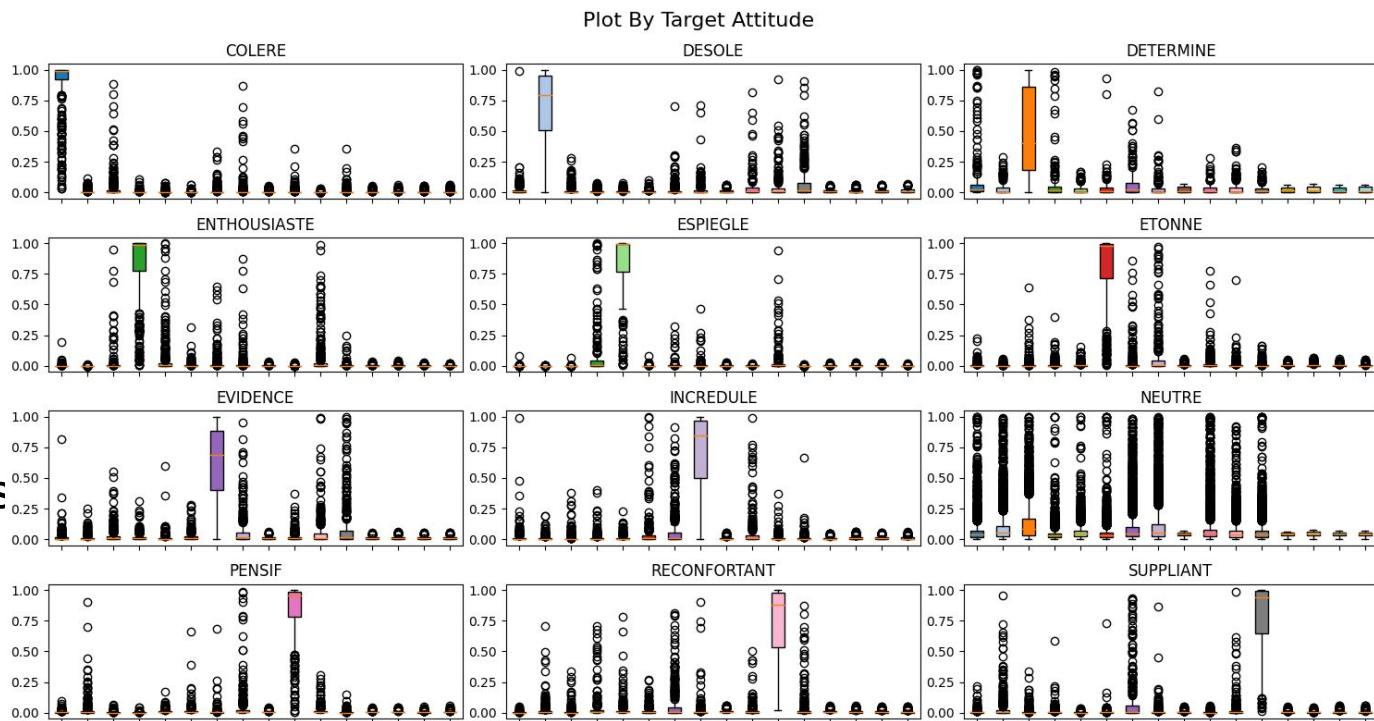
Expressive AVTTS with style driven by verbal input

- Reference encoder and verbal tags predict weights of the GST
 - 11 tokens for expressive instructions + 4 free tokens for narration



GST weights in the training corpus

- Effects of cross-entropy on weights of instructed attitudes but...
- Spreading of activations especially for «Narrative»
- 4 unused extra tokens



Demo

- « Pure » style tag
 - « En colère » (angry)
 - COLERE = 0.99



Selection TTS : Multi Speaker/Style

Selection Vocodeur : Hifi-GAN V2 FR 570000 Waveglow NEB

Speaker : NEB DG RO IZ AD MLB YB

StyleTag : en colère

Style : COLERE
 DESOLE
 DETERMINE
 ENTHOUSIASTE
 ESPIEGLE

Input Text : Je vous souhaite la bienvenue à cette présentation du proje

Synthèse

Durée audio : 3.518s
Durée TTS : 0.228s | 6% de la durée audio
Durée Vocodeur : 0.166s | 5% de la durée audio
Durée Denoiser : 0.039s | 1% de la durée audio
Durée Totale Synthèse : 0.432s | 12% de la durée audio

Play

GST weights

COLERE: 0.99
DESOLE: 0.00
DETERMINE: 0.00
ENTHUSIASTE: 0.00
ESPIEGLE: 0.00
ETONNE: 0.00
EVIDENCE: 0.00
INCREDULE: 0.01
NEUTRE: 0.00
PENSIF: 0.00
RECONFORTANT: 0.00
SUPLIANT: 0.00
TOKEN13: 0.00
TOKEN14: 0.00
TOKEN15: 0.00
TOKEN16: 0.00

Demo

- « Pure » style tag
 - « Etonnée » (surprized)
 - ETONNEE=.99



Selection TTS : Multi Speaker/Style

Selection Vocodeur : HiFi-GAN V2 FR 570000 Waveglow NEB

Speaker : NEB DG RO IZ AD MLB YB

StyleTag : étonnée

Style : COLERE DESOLE DETERMINE ENTHOUSIASTE ESPIEGLE

Input Text Je vous souhaite la bienvenue à cette présentation du proje Synthèse

Durée audio : 3.657s
Durée TTS : 0.302s | 8% de la durée audio
Durée Vocodeur : 0.229s | 6% de la durée audio
Durée Denoiser : 0.051s | 1% de la durée audio
Durée Totale Synthèse : 0.582s | 16% de la durée audio

Play

GST weights

COLERE: 0.00
DESOLE: 0.00
DETERMINE: 0.00
ENTHOUSIASTE: 0.00
ESPIEGLE: 0.00
ETONNE: 1.00
EVIDENCE: 0.00
INCREDULE: 0.00
NEUTRE: 0.00
PENSIF: 0.00
RECONFORTANT: 0.00
SUPPLIANT: 0.00
TOKEN13: 0.00
TOKEN14: 0.00
TOKEN15: 0.00
TOKEN16: 0.00

Demo

- « Nuanced » style tag
 - « Bienveillante » (kind)
 - ENTHOUSTIASTIC=.61,
COMFORTING=.38



Selection TTS : Multi Speaker/Style

Selection Vocodeur : Hifi-GAN V2 FR 570000 Waveglow NEB

Speaker : NEB DG RO IZ AD MLB YB

StyleTag : bienveillante

Style : COLERE DESOLE DETERMINE ENTHOUSTIASTE ESPIEGLE

Input Text : Je vous souhaite la bienvenue à cette présentation du proje Synthèse

Durée audio : 3.669s
Durée TTS : 0.224s | 6% de la durée audio
Durée Vocodeur : 0.185s | 5% de la durée audio
Durée Denoiser : 0.039s | 1% de la durée audio
Durée Totale Synthèse : 0.448s | 12% de la durée audio

Play

GST weights

COLERE: 0.00
DESOLE: 0.00
DETERMINE: 0.00
ENTHOUSTIASTE: 0.61
ESPIEGLE: 0.00
ETONNE: 0.00
EVIDENCE: 0.00
INCREDULE: 0.00
NEUTRE: 0.00
PENSIF: 0.00
RECONFORTANT: 0.38
SUPPLIANT: 0.00
TOKEN13: 0.00
TOKEN14: 0.00
TOKEN15: 0.00
TOKEN16: 0.00

Demo

- « Nuanced » style tag
 - « Dubitative » (doubtful)
 - SKEPTICAL=.61, THOUGHTFUL=.20, MISCHIEVOUS=.13



Selection TTS : Multi Speaker/Style

Selection Vocodeur : Hifi-GAN V2 FR 570000 Waveglow NEB

Speaker : NEB DG RO IZ AD MLB YB

StyleTag : dubitative

Style : COLERE DESOLE DETERMINE ENTHOUSIASTE ESPIEGLE

Input Text : Je vous souhaite la bienvenue à cette présentation du proje Synthèse

Durée audio : 3.762s
Durée TTS : 0.225s | 6% de la durée audio
Durée Vocodeur : 0.174s | 5% de la durée audio
Durée Denoiser : 0.040s | 1% de la durée audio
Durée Totale Synthèse : 0.439s | 12% de la durée audio

Play

GST weights

COLERE: 0.00
DESOLE: 0.00
DETERMINE: 0.00
ENTHOUSIASTE: 0.00
ESPIEGLE: 0.13
ETONNE: 0.00
EVIDENCE: 0.03
INCREDULE: 0.61
NEUTRE: 0.00
PENSIF: 0.20
RECONFORTANT: 0.00
SUPPLIANT: 0.00
TOKEN13: 0.00
TOKEN14: 0.01
TOKEN15: 0.00
TOKEN16: 0.00

Conclusions and perspectives

- Methodology for the semi-directed collection of verbal tags via crowdsourcing
 - Navigation into an emotional space
- Expressive AVTTS alternatively driven via :
 - A reference signal
 - Weighted reference tokens (restricted set of emotional labels)
 - Full verbal description (set of adjectives)
- Short-term:
 - Tag the entire database: 1000 subjects
 - Evaluate this fine control
- Mid-term:
 - Transferring to other speakers/languages
 - Enriching the emotional palette