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

Gérard Bailly, Romain Legrand, Martin Lenglet, Frédéric Elisei, Maëva Garnier and Olivier Perrotin
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab
F-38000 Grenoble, France
{first_name.family_name}@gipsa-lab.fr

Abstract
We developed a web app for ascribing verbal descriptions to expressive audiovisual utterances. These descriptions are limited to lists of adjective tags that are either suggested via a navigation in emotional latent spaces built using discriminant analysis of BERT embeddings, or entered freely by participants. We show that such verbal descriptions collected on-line via Prolific on massive French audiovisual data (742 participants, 8 970 tagged utterances up-to-now) provide Expressive MFciteultimodal Text-to-Speech Synthesis with precise verbal control over desired emotional content.

Keywords: verbal labelling, expressive audiovisual speech, TTS

1. Motivation

End-to-end Text-To-Speech (TTS) models achieve high standards in terms of naturalness, when applied to read literary texts (Perrotin et al., 2023). However, the control of expressivity, encountered in particular in simulated dialogues, remains a challenging issue. Stylistic variations have been successfully introduced by the encoding of reference speech signals (later called style embeddings) to bias the output of the text encoder. To allow for an explicit control of these variations, a projection of the style embeddings on a reduced sets of vectors such as the Global Style Tokens (GST) was introduced by Wang et al. (2018). The supervised training of these tokens to model specific emotions has then lead to an explicit control of expressivity using expressive tags (Wu et al., 2019), but that is limited to a small vocabulary. Recently, Kim et al. (2021) and Shin et al. (2022) have proposed to train such a style encoder with the addition of verbal tags collected via crowdsourcing, and encoded with Large Language Models such as BERT (Devlin et al., 2019), in parallel to the speech signal. They demonstrated that such co-constructed latent spaces combine interpolation capabilities (from the speech signal) with precise control of expressivity (from verbal tags).

One key issue then is how to collect verbal tags from expressive speech samples: free tagging may hinder verbal qualifiers that are often ”on the tip of the tongue” while forced choices made from a list of words or descriptions restrain felt emotions. *We describe here an original methodology to massively collect verbal tags from audiovisual data for supplying the style component of an expressive text-to-speech system (Lenglet et al., 2023) with verbal entries*, i.e. a list of adjectives – similar to didaskalia – describing how the utterance should be spoken.

2. State of the art

The seminal proposal of Kim et al. (2021) and Shin et al. (2022) to use verbal prompts as an alternative control of expressive TTS to predefined tags has triggered several works.

Recently, Liu et al. (2023) proposed to use a prompt encoder to extract prompt embeddings from natural language description. They invited nine professional annotators to describe the style of given utterance with a phrase or a sentence. They were told to focus on the speaking style only and ignore the linguistic content. The corpus contains 12 hours of speech data from eight female speakers, and one half was associated with prompts. Similarly, Yang et al. (2023) collected prompts in three steps: (1) one word to describe the overall perceived emotion of an utterance; (2) one word to describe the emotion level of the utterance; (3) a complete sentence in natural language to describe the style of the utterance.

Most labelling schemes impose annotators to draw emotional tags from a small set of labels, typically the six basic emotions (Ekman and Friesen, 1971) vs. 18 in the Geneva Emotion Wheel Scherer (2005) or a task-specific set (seven in Feng et al. (2022)). Labels collected during free-labelling tasks are often post-attributed to pre-defined categories (six in Widen and Russell (2008) vs. 12 in Vicari et al. (2000)).

In all cases, no assistance was given to participants for writing the prompts: suggested emotional tags are just given as a large alphabetic

LREC-COLING 2024, pages 5689–5695

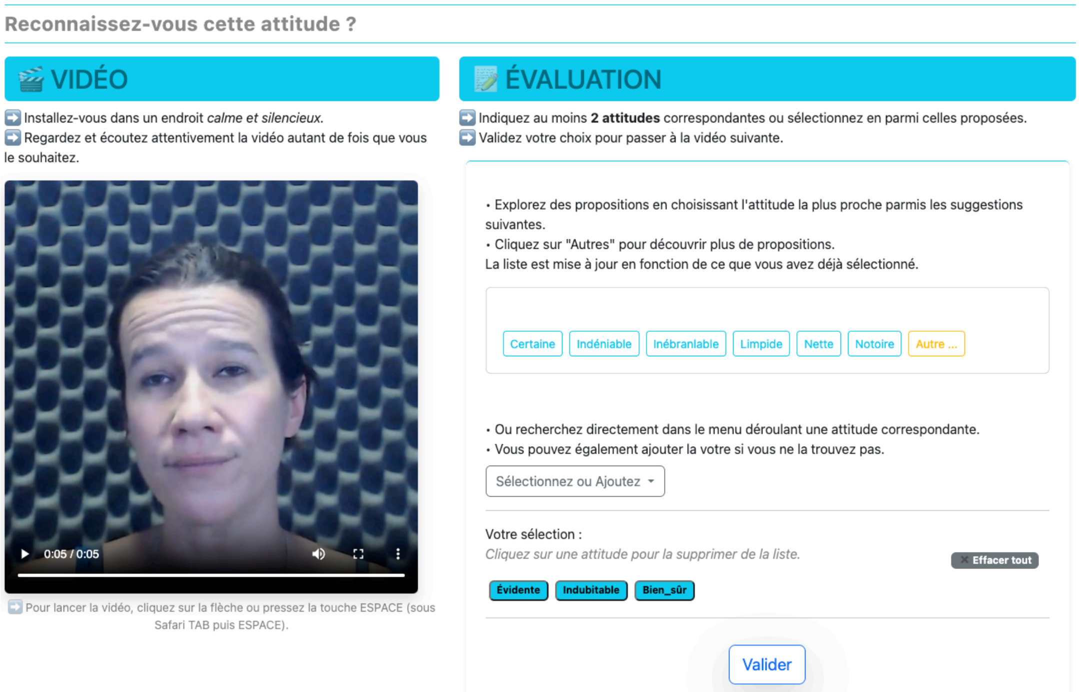

Figure 1: Evaluation page. Left: video can be freely played. Right: the evaluation panel proposes two ways to complement the list of adjectives (displayed in blue boxes at the bottom) : (1) a suggestion of 6 adjectives close to the previously selected one, if any, or far from the proposed ones if the ”autre” (other) button has been selected; (2) a selection among all registered adjectives or free input.

Table 1: Phone-aligned expressive dataset, with durations given in minutes.

Style	Train		Test	
	Dur.	#Utt	Dur.	#Utt
Angry	17.9	396	1.2	26
Sorry	15.2	328	1.0	25
Committed	15.1	342	0.7	15
Enthusiastic	16.4	304	0.6	15
Mischievous	12.3	343	0.8	22
Surprised	15.6	325	0.8	15
Obvious	27.4	495	1.2	24
Skeptical	16.9	405	0.7	16
Thoughtful	29.7	334	2.8	26
Comforting	25.4	401	1.9	30
Pleading	18.0	330	1.5	26
Narrative	205.0	4332	9.1	199
Total	**415.4**	**8345**	**22.3**	**439**

list (Dupre et al., 2015) or grouped by categories (Demszky et al., 2020). **We propose here** a system for incrementally suggesting possible descriptions. All judgments are based on audiovisual speech uttered with a large variety of attitudes, that reflect the position of speakers with regard to what they say (Bolinger and Bolinger, 1989).

3. Expressive audiovisual TTS

Data. A French female comedian uttered sentences extracted from the SIWIS database (Goldman et al., 2016) in a narrative mode as well as with 11 attitudes elicited by a short context during “exercice-in-style” sessions (Queneau, 2018). These utterances were phone-aligned and one half of these alignments were hand-checked (Table 1) and used to train an expressive TTS.
TTS. The global architecture of expressive audiovisual TTS is given in Figure 2. Its backbone is FastSpeech2 (Ren et al., 2020) to which several modules have been grafted:

- a phonetic predictor (Bailly et al., 2023)
- a visual decoder
- a GST module whose weights are cross-correlated with the instructed emotional tags
- a LST module that modulates the utterance-wise GST embeddings with word-wise local embeddings (Lenglet et al., 2023)

The final objective of this work is to replace the GST module – that is currently controlled by the limited set of instructed emotional tags – by a module driven by finer emotional tags, i.e a list of adjectives combined via BERT embeddings (see Fig. 4).

4. Computer-Assisted Verbal Labelling

Our labelling interface (see Fig.1) proposes three ways to assign emotional tags to each video clip:

- select among a large list of 132 predefined adjectives

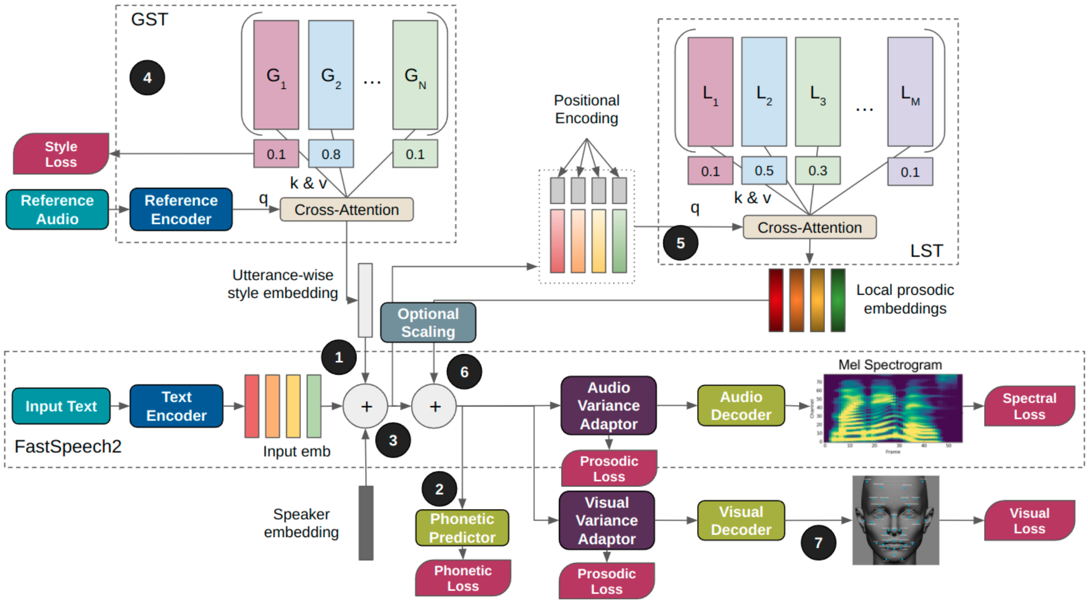

Figure 2: Architecture of the proposed expressive audiovisual text-to-speech system: a FastSpeech2 kernel (see 1) is augmented with a phonetic predictor (2), a speaker embedding (3), a GST module whose weights are cross-correlated with emotional tags (4), a LST module that modulates the utterance-wise GST embeddings with word-wise local embeddings (5) and overlap-and-add with the GST output (6) and a visual decoder (7). Note that GST and LST can be scaled to monitor style strengths.

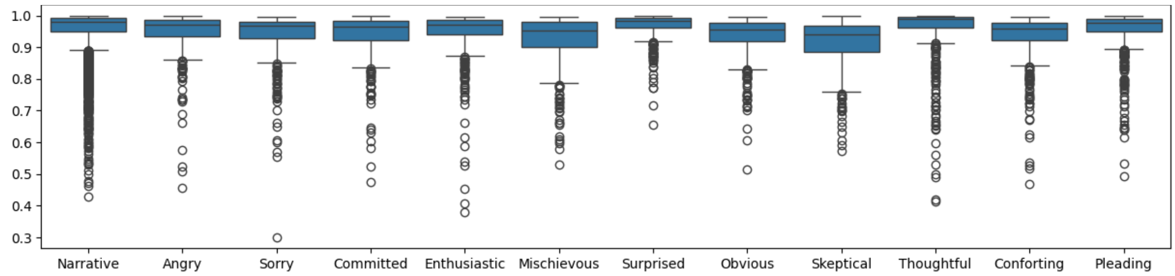

Figure 3: GST weights for the style embedding corresponding to each instructed style. Fscore of majority voting is close to 1.

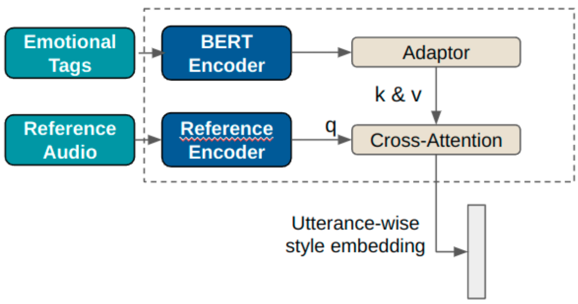

Figure 4: New foreseen style encoder combining audio and verbal input.

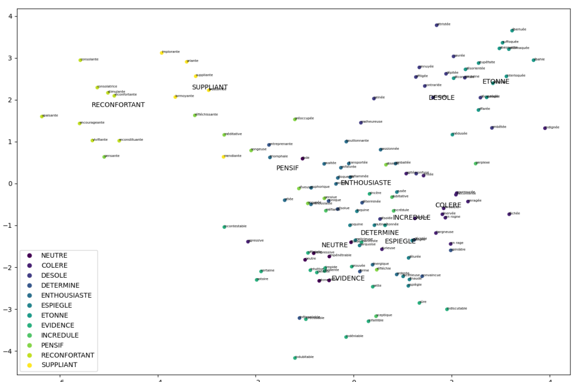

Figure 5: Projections of 132 adjectival synonyms of the nominal 12 attitudes on the two first discriminant axis of their FlauBERT embeddings.

- select among a small selection of 6 tags, extracted incrementally from the large set
- free text input. We encourage labellers to use feminine adjectives to qualify the emotional performance of the female comedian.

Building an emotional space from BERT embeddings. We beforehand collected a dozen of synonyms for each of the 12 instructed attitudes, resulting in a collection of 132 feminine adjectives. We performed a Linear Discriminant Analysis (LDA) of 1024 embeddings of the penultimate layer of the large cased model FlauBERT (Le et al., 2020): sub-tokens if any are summed-up. We thus obtain 11 discriminant dimensions of this “emotional” latent space. The projection of the 132 synonyms onto the first factorial plane is shown in Fig. 5.

Incrementally suggesting a selection of emotional tags. In order to ease navigation into this emotional space, we propose an interactive selection of emotional tags via a simple search policy based on the RMS distance between each pair of tags computed on the 11 loading factors. This method is inspired by the Nelder-Mead simplex algorithm using expansion / contraction of the search space (Singer and Nelder, 2009):

- 12 adjectives close to each centre of the groups are first proposed together with an “other” option
- if “other” is selected, 6 tags which are furthest from all tags already explored are further proposed
- otherwise the tag is stored and 6 tags which are closest to the selected tag are further proposed
- the participant can keep selecting tags as much as he/she wants

We tested this procedure by asking 30 subjects to recover an adjective not proposed in the first group of 12. An average of 2.3±1.2 clicks were necessary to retrieve the proper target: this indirectly shows the homogeneity of the emotional space that mirrors expected semantic distances.

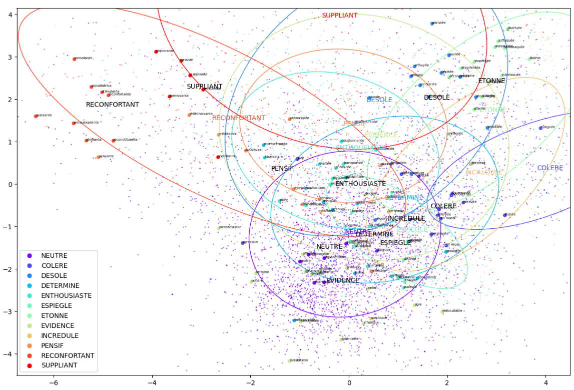

Figure 6: Projections of 8 970 tagged clips together with the dispersion ellipsis of the emotional group to which they were supposed to belong.

5. Evaluation

We aim at collecting at least 10 tags for each 12 461 clips of our audiovisual database. Subjects were recruited via the crowdsourcing platform Prolific[1] and social networks. They were asked to tag 60 clips randomly picked in the database. A minimum of 2 suggested or free tags per clip was imposed to access to the next clip. A session lasted approximately 30 minutes and the participants were paid 6£. At the date of submission, 8 970 (7198%) clips have been tagged by at least one subject. We collected 111 399 tags. A Jupyter notebook provides stats about the current data collection [2].

After hand correction, we kept 288 free tags that were proposed by at least two subjects. We then averaged the embeddings of the selected tags (either suggested or free) for each of the 8 970 tagged clips. The projection of their embeddings onto the first factorial plane is given in Fig.6.

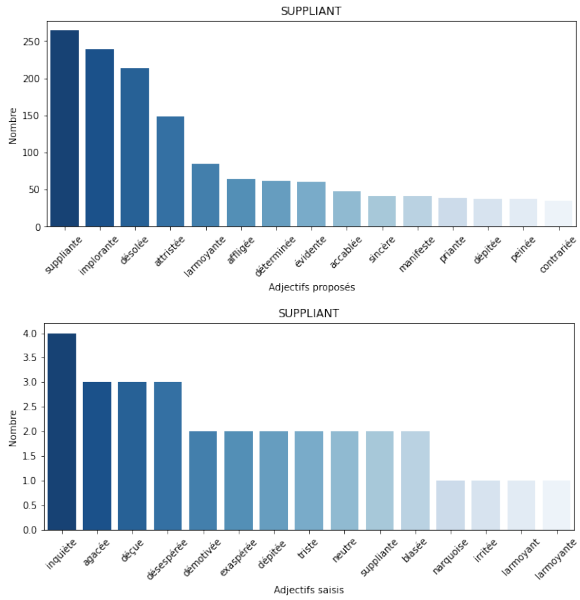

Figure 7: Use of *suggested* (top) vs *free* (bottom) tags for ”Pleading”. Note the scale difference.

Effectiveness of the suggestions We collected 12.18 suggested vs 0.24 free tags per clip. Suggested tags were found by selecting 4.09 “other” per clip. This shows that:

- The iterative suggestion system is quite effective: only 2% of tags were given as free text
- The refinement process is also quite effective: only 30% of the sets of suggested tags are discarded
- As evidenced by the high separability of emotions by the GST (see Fig. 3), the most popular tags for each group of clips do correspond to the nominal emotion tag in the 12 adjectives chosen to bootstrap suggestions (see Fig. 7).

Emotional coverage Using the average Fréchet distance (Brechet et al., 2009) between

[1] https://prolific.com

[2] https://gricad-gitlab.univ-grenoble-alpes.fr/web/emotags-results/-/blob/main/analyse_prolific.ipynb

distributions of tagged clips and distributions of suggested synonyms for each instructed emotion as an indicator of displacements of dispersion ellipsoids, we noticed that the “Narrative” (-9.25) was the only group to shrink whereas all others expand/move away from their expected emotional space, in particular “Mischievous” (+85.57), “Angry” (+79.33) and ”Enthusiastic” (+63.87). “Comforting” (+14.59) and “Sorry” (+16.28) are the most stable ones.

Style	GST	LDA	PCA	#Utt
Angry	.80	.70	.74	400
Sorry	.91	.71	.80	353
Committed	.68	.65	.69	324
Enthusiastic	.79	.71	.75	439
Mischievous	.72	.63	.69	252
Surprised	.76	.67	.69	395
Obvious	.71	.58	.61	444
Skeptical	.72	.58	.62	458
Thoughtful	.78	.70	.74	399
Comforting	.86	.74	.75	432
Pleading	.81	.69	.71	517
Narrative	.77	.59	.51	4557
Overall	.77	.62	.59	8 970

Table 2: R2 values for the prediction of audio embeddings from PCA projections of GST output (95% of variance explained by 11 dimensions) vs. LDA projections of synonyms (11 dimensions), PCA projections of tags (95% of variance explained by 38 dimensions). Note that GST weights – contrary to the explicit verbal tags – have no semantics.

Verbal vs audio embeddings To assess whether the verbal tags embedding space can model as much variability as the reference audio encoder, we attempted to predict the 128-dimensional audio embeddings from verbal tags embeddings, using a linear regression. A high coefficient of determination (R^2) indicates that there is a linear mapping between both representation spaces, therefore encoding similar information. Prior to computing the linear regression, we reduced the dimension of the verbal tags embedding space to remove correlated dimensions, in two different ways: 1) use of a 11-dimensional LDA projection of the verbal tag embedding space (LDA condition) ; 2) use of a 38-dimensional PCA projection (95% of the variance) of collected tags (PCA condition). This is to be compared to the regression with the 12-dimensional weights of the GST (GST condition). The coefficients of determination (R^2) for each condition and each speaking style are reported in Fig. 2.

The high R^2 for GST is expected since the reference encoder was trained for discriminating between instructed styles. But no semantics is associated to GST weights: the aim of our work is precisely to explicitly fine-control variability via verbal tags. Around 87%=$\sqrt{.75}$ of the variance captured by GST is explained by verbal tags.

For both LDA and PCA conditions, the coefficients of determination (R^2) lay around .65 for each style. The R^2 of PCA projections are significantly above those obtained from the LDA projection, except for Narrative clips. This performance is rather encouraging since the reference encoder was trained for optimal GST projection and with cross-entropy loss. Tag embeddings delivered by the fine-grained verbal description of each utterance used as alternative input to the style encoder will certainly increase this fit. Note that the computation of audiovisual embeddings would potentially increase R^2.

6. Conclusions and perspectives

We hereby propose a system from ascribing verbal descriptions to expressive audiovisual utterance that are either iteratively suggested via a navigation in so-called emotional latent spaces or entered freely by subjects. We show that such a labelling system can be deployed at a large scale to efficiently collect relevant verbal descriptions. This procedure can be easily extended to other labelling tasks and other languages, as long as a BERT system is available for it.

Once the targeted collection of 10 tags per utterance has been reached and that verbal and prosodic descriptions can be related, the next step is to train and evaluate the control of expressive audiovisual TTS with such mixed – i.e. verbal vs. signal – style input (similar to (Kastner et al., 2019) for text vs. phonetic input) to enable fine-grained verbal descriptions of desired expressivity. Note that we will keep GST (see Lenglet et al., 2023) as the back-end of the style encoder so that to cluster emotions and regress style embeddings with the output phonetic bias that is added at the output of the text encoder: in fine, the objective is to replace the objective is to ”replace” GST weights (i.e. 0.8 ”anger” + 0.2 ”doubtfull”) by a verbal nuance (e.g. ”indignant”).

Acknowledgements

This research has received funding from the BPI project THERADIA and MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

Bibliographical References

Gérard Bailly, Martin Lenglet, Olivier Perrotin, and Esther Klabbers. 2023. Advocating for text input in multi-speaker text-to-speech systems. In *Speech Synthesis Workshop*, pages 1–7.

Dwight Bolinger and Dwight Le Merton Bolinger. 1989. *Intonation and its uses: Melody in grammar and discourse*. Stanford university press.

Claire Brechet, René Baldy, and Delphine Picard. 2009. How does sam feel?: Children's labelling and drawing of basic emotions. *British Journal of Developmental Psychology*, 27(3):587–606.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Annual Meeting of the ACL*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL/HLT*, pages 4171–4186.

Damien Dupre, Daniel Akpan, Elena Elias, Jean-Michel Adam, Brigitte Meillon, Nicolas Bonnefond, Michel Dubois, and Anna Tcherkassof. 2015. Oudjat: A configurable and usable annotation tool for the study of facial expressions of emotion. *International Journal of Human-Computer Studies*, 83:51–61.

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *LREC*, pages 4096–4113.

Jean-Philippe Goldman, Pierre-Edouard Honnet, Rob Clark, Philip N Garner, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Tiago Macedo, Beat Pfister, Manuel Sam Ribeiro, et al. 2016. The SIWIS database: a multilingual speech database with acted emphasis. *Interspeech*, pages 1532–1535.

Kyle Kastner, João Felipe Santos, Yoshua Bengio, and Aaron Courville. 2019. Representation mixing for tts synthesis. In *ICASSP*, pages 5906–5910.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021. Expressive Text-to-Speech Using Style Tag. In *Interspeech*, pages 4663–4667.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for French. In *LREC*, pages 2479–2490.

Martin Lenglet, Olivier Perrotin, and Gérard Bailly. 2023. Local Style Tokens: Fine-Grained Prosodic Representations For TTS Expressive Control. In *Speech Synthesis Workshop*, pages 120–126.

Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Zhifei Li, and Lei Xie. 2023. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. *arXiv:2305.19522*.

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, and Gérard Bailly. 2023. The Blizzard Challenge 2023. In *Blizzard Challenge Workshop*, Grenoble, France.

Raymond Queneau. 2018. *Exercises in style*. Alma Books.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv:2006.04558*.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Yookyung Shin, Younggun Lee, Suhee Jo, Yeongtae Hwang, and Taesu Kim. 2022. Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS. In *Interspeech*, pages 2313–2317.

Saša Singer and John Nelder. 2009. Nelder-mead algorithm. *Scholarpedia*, 4(7):2928.

S Vicari, J Snitzer Reilly, P Pasqualetti, A Vizzotto, and C Caltagirone. 2000. Recognition of facial expressions of emotions in school-age children: the intersection of perceptual and semantic categories. *Acta Paediatrica*, 89(7):836–845.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *ICML*, pages 5180–5189.

Sherri C Widen and James A Russell. 2008. Children acquire emotion categories gradually. *Cognitive Development*, 23(2):291–312.

Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu, and Lirong Dai. 2019. End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *APSIPA Annual Summit*, pages 623–627.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Guangzhi Lei, Chao Weng, Helen Meng, and Dong Yu. 2023. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *arXiv:2301.13662*.

7. Appendice

Style	Suggested adjectives
Angry	en colère, en rage, énervée, agitée, agressive, courroucée, en rogne, enragée, fâchée, furieuse, hargneuse, indignée, irritée, mécontente, rageuse, autoritaire
Sorry	désolée, accablée, affligée, attristée, confuse, contrariée, découragée, dépitée, embêtée, en peine, ennuyée, malheureuse, navrée, peinée
Committed	déterminée, convaincue, décidée, entreprenante, ferme, inébranlable, opiniâtre, résolue, tonique
Enthousiastic	enthousiaste, éloquente, ardente, bouillonnante, emballée, énergique, enfiévrée, enflammée, euphorique, exaltée, excitée, pas sionnée, transportée, triomphale, zélée
Mischievous	espiègle, coquine, délurée, facétieuse, finaude, malicieuse, maligne, mutine, narquoise, rusée, taquine
Surprized	étonnée, ébahie, abasourdie, éberluée, déconcertée, désorientée, effarée, estomaquée, hébétée, interloquée, médusée, sidérée, stupéfaite, suffoquée
Obvious	évidente, certaine, incontestable, indéniable, indiscutable, indubitable, infaillible, intuitive, irrécusable, limpide, manifeste, nette, notoire, officielle, patente, prouvée, sûre, sincère
Skeptical	incrédule, dubitative, méfiante, perplexe, sceptique
Thoughtful	pensive, absente, méditative, occupée, pensante, préoccupée, rêveuse, réfléchie, réfléchissante, songeuse
Comforting	réconfortante, apaisante, consolante, consolatrice, encourageante, reconstituante, stimulante, vivifiante
Pleading	suppliante, implorante, larmoyante, mendiante, priante, prosternée
Narrative	neutre, atone, fade, froide, impénétrable, inexpressive, terne

Table 3: List of synonyms for each emotional group, used as suggestions in the navigation system