Discourse phrases classification: direct vs. narrative audio speech

Marie Tahon, Damien Lolive

Discourse phrases classification: direct vs. narrative audio speech

Marie Tahon[1]*, Damien Lolive*[2]

[1]LIUM / Le Mans University, Le Mans, France
[2]IRISA/ University of Rennes 1, Lannion, France
marie.tahon@univ-lemans.fr, damien.lolive@irisa.fr

Abstract

In the field of storytelling, speech synthesis is trying to move from a neutral machine-like to an expressive voice. For parametric and unit-selection systems, building new features or cost functions is necessary to allow a better expressivity control. The present article investigates the classification task between direct and narrative discourse phrases to build a new expressivity score. Different models are trained on different speech units (syllable, word and discourse phrases) from an audiobook with 3 sets of features. Classification experiments are conducted on the Blizzard corpus which features children English audiobooks and contains various characters and emotional states. The experiments show that the fusion of SVM classifiers trained with different prosodic and phonologic feature sets increases the classification rate from 67.4% with 14 prosodic features to 71.8% with the 3 merged sets. Also the addition of a decision threshold achieves promising results for expressive speech synthesis according to the strength of the constraint required on expressivity: 71.8% with 100% of the words, 79.9% with 50% and 82.6% with 25%.

Index Terms: discourse phrases, narration, classification, audiobook, expressive speech synthesis

1. Introduction

Speech synthesis usually consists in the conversion process of a written text to a speech sound, named as Text-To-Speech (TTS). Nowadays, TTS tries to move from a neutral and machine-like style to expressive speech with different speaking styles under various emotional states. Especially in the field of storytelling, Expressive Speech Synthesis (ESS) systems should be able to read books of different literary genres using various discourse modes and speaking styles.

Following this trend, in 2016 and 2017, the Blizzard synthesis Challenge proposed to the participants to build an expressive voice from children audiobooks in English [1, 2]. The system presented by IRISA team in 2017 [3] proposed an expressivity score which evaluates how expressive a speech segment is in the acoustic space of the speaker. This score is then introduced in a unit-selection system assuming that narrative parts are less expressive than dialogs. The purpose of this score is to favor the concatenation of units bearing similar expressivity levels. At this stage, the score has not been evaluated through classification but synthesis.

Audiobooks gather many characteristics which are suitable for ESS. They contain both a text of interest and the corresponding audio speech signal [4]. The reader usually uses different speaking styles and emotions [5]. He also personifies the different characters of the story by changing his way of speaking [6]. Some unsupervised approaches for automatic annotation of expressive styles have been investigated. Expressive clusters are associated with different voice styles considering glottal source parameters [7], prosodic features [8, 9], wavelets [10], spectral features [11] or voice quality [12]. For instance, in [13], the authors classified narrative structures using linguistic features. They obtained a F1 score of 71.5% on a specific narrative structure classification experiment. As far as we know, no experiments were conducted on both phonological and acoustic levels.

The present article investigates the classification of speech segments according to the two following classes: narrative and direct speech. Different models are trained on discourse phrases as direct or narrative parts from audiobooks with different sets of prosodic and phonologic features and also the state of the art Opensmile feature set [14]. Those models are then combined to improve the results.

Three data-driven approaches co-exist for TTS systems: unit selection, statistical parametric systems (mainly Hidden Markov Models or Deep Neural Networks) and hybrid systems. Generally speaking, while a parametric representation enables more flexibility, the unit-selection has the advantage of sound quality. For instance, interpolation between styles contained in the database is possible using such methods [15]. Several approaches are used to give the systems more flexibility like model adaptation to specific voices or prosodic styles [16, 17], symbolic constraints (diphone identity, position, etc.) [18] or prosodic constraints [19, 20]. These elements are usually used in the speech synthesizer directly in the cost function of unit selection systems or as input features during the construction of the acoustic model of parametric systems [21]. In the present work, we build a classifier whose prediction can be included as a feature or in the cost function in the form of an expressivity score assessing the expressivity level of a speech segment [3].

The remainder of the paper is organized as follows: section 2 presents the Blizzard audio corpus and feature sets. The different models used for classifying narrative and direct discourse phrases are described in section 3. The results are presented in section 4.

2. Material and data

2.1. Expressive audiobook corpora

Table 1: *Blizzard expressive corpus characteristics.*

Blizzard corpus	train	test	validation
# utterances	546	272	350
# words	29,190	10,771	17,331
# discourse phrases	3,256	980	1,727
# syllable	41,036	14,941	24,543
F_0 Hz (std)	192 (58)	187 (57)	188 (56)

The English Blizzard audiobook corpus [2] designed for ex-

pressive speech synthesis is described in Table 1. Children's audiobooks were read by a professional British female speaker. Around 6.5 hours of material were made available to participants of the challenge in 2017. This corpus is expressive as the reader embodies different characters, thus changing her voice to fit the suitable expressivity. Voice changes are notably remarkable through pitch variations (see the high F_0 standard deviation in table 1). Speech signals were segmented in phone units and aligned with the corresponding text. Additional linguistic, syntactic and phonological information has been automatically added to the corpus. Discourse phrases segmentation and annotation mainly relies on quotes which are present in the text. The data has been split into 3 parts on the number of books, 50% for training the models, 25% for development purposes and 25% for the future final speech synthesis evaluation. Additional experiments have been set up on the French SynPaFlex audiobook corpus [22], however, the full results will not be presented in this paper.

2.2. Acoustic features

Three sets of acoustic features are extracted at the syllable, word and speech style levels on both corpora.

Proso is a set of prosodic features developed by the team. It consists of low-level features such as fundamental frequency (F_0) in semitone (min, max, range, average, standard deviation, slope, discrete level), energy (min, max, range, average), duration, articulation and speech rates. The F_0 discrete level consists in assigning a level between 1 and 5 relatively to the whole corpus. In total, 14 features are extracted at word and discourse phrase levels. Only F_0 features are extracted at the syllable level.

Os385 is a reference set in affective computing. In this paper the Interspeech 2009 challenge configuration is used [14]. It consists of 12 low-level features combined with 14 functionals and their derivatives, in other words 384 features to which we add the segment duration. In order to enable Opensmile extraction possible, we restricted the studied items to those which last more than 50 ms.

Phono is a set of 6 phonological features developed by the team. It consists of the number of phonemes, the number of vowels and of consonants over the total number of phonemes, the number of opened syllables over the total number of syllables, the average phoneme frequency and the average vowel frequency. Frequencies are extracted from the whole corpus.

3. Discourse models

The aim is to train models able to distinguish narrative and direct phrases in read speech. Therefore, we present different techniques to predict whether an unknown sample belongs to narrative or direct speech class. The choice for the different statistical models tested in this paper is driven by speaker identification and emotion recognition techniques: Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). However, direct speech samples can also be considered as outliers among narrative samples and outlier detection techniques [23] such as one-class models, are also considered.

3.1. Two class models

The classification task consists in automatically labelling an unknown speech signal x. To do so, during the learning stage, two functions $\mathcal{M}_{dir}$ and $\mathcal{M}_{indir}$ are built to represent respectively the direct phrases and the narrative phrases in speech. When comes a new speech signal x, the difference $\Delta(x)$ between each model is computed (eq. 1). The predicted class $c(x)$ depends on the sign of Δ (eq. 2).

$$\Delta(x) = \mathcal{M}_{narr}(x) - \mathcal{M}_{dir}(x) \tag{1}$$

$$c(x) = \begin{cases} \text{narr if } \Delta(x) \geq 0 \\ \text{direct} \quad \text{if } \Delta(x) < 0 \end{cases} \tag{2}$$

GMM models We first try to classify narrative and direct speech using Gaussian Mixture Models using speaker identification techniques. These are trained on the three acoustic feature sets presented before. More precisely, two GMM λ_i are trained for each discourse phrase. The number of Gaussians were limited to 10 in order to better generalize the data. This number in then optimized according to the BIC criterion on training data. Furthermore, Gaussians weights which represent less than 25% of the data are set to zero, the others are updated in order that the total sum of weights remains equal to 1. This change is done to avoid too specialized gaussians. During the decoding part, the GMM function is defined according to $\mathcal{M}_i(x) = \log p(x|\lambda_i)$.

SVM models We also trained SVM models using standard protocols used in emotion recognition frameworks. SVM parameters (kernel, complexity and γ) are optimized using a grid search cross-validation technique using the unweighted average recall (UAR) measure. During the decoding part, the SVM function depends on the distance to hyperplane obtained while testing a new sample x with the model $\mathcal{M}_i(x) = dist_i(x)$.

3.2. One class (OC) models

Founded on outliers detection techniques, one class models are also tested. A single model is trained on narrative utterances and should represent the "normal" speech. A pseudo-likelihood threshold ΔL is extracted from development data according to eq 3 where U_i is the set of samples belonging to class i. Finally, when comes a new sample x, the difference $\Delta(x)$ of the log-likelihood obtained on the one-class model and ΔL is computed (eq. 4). The predicted class $c(x)$ is defined by eq. 2. When $\Delta(x)$ is high, the sample is considered as abnormal (direct speech), on the contrary it is considered as normal (narrative speech).

$$\Delta L = \frac{1}{2}\left[\underset{x \in U_{narr}}{\text{mean}}\, \mathcal{M}(x) + \underset{x \in U_{dir}}{\text{mean}}\, \mathcal{M}(x)\right] \tag{3}$$

$$\Delta(x) = \mathcal{M}_{narr}(x) - \Delta L \tag{4}$$

We have also tested to train models on all narrative and direct phrases following the methods typically used for speaker identification. Indeed, a universal background model (UBM) [24] should capture the most general characteristics of speech. However, the results obtained with this method were not convincing, consequently we decided not to present them.

3.3. Random Forest models

Random Forests (RF) have the advantage of being very fast and of not overfitting the data. Overfitting is highly probable in our case since the training corpus is quite small (as most of emotional speech databases). Models are learnt with 50 estimators and the entropy criterion. The predicted class $c(x)$ is the one with the highest likelihood across the trees. An additional AdaBoost classifier has also been set up.

Table 2: *UAR (UAF1) classification results on Blizzard corpus after model optimization. Features are extracted on syllable and word units and also on discourse phrases (DP).*

Item	#utt narr/dial	Set	GMM	OC-GMM	SVM	OC-SVM	RF	RF-BOOST
Syllable	train: 11632/11632	Os385	63.8 (60.4)	57.1 (53.9)	**68.2 (61.1)**	55.8 (49.1)	60.0 (48.2)	64.4 (56.8)
	test: 11549/2819	Phono	52.8 (40.0)	48.3 (47.9)	**58.2 (53.9)**	50.3 (45.2)	51.8 (46.0)	52.2 (46.3)
		Proso	63.2 (59.8)	50.0 (44.6)	**66.5 (62.5)**	60.3 (52.7)	59.3 (49.8)	64.0 (56.6)
Word	train: 8763/8763	Os385	53.5 (30.7)	59.8 (59.6)	64.8 (51.2)	55.2 (41.7)	61.1 (48.9)	**65.9 (58.2)**
	test: 8455/2194	Phono	54.3 (51.7)	49.6 (43.7)	**57.7 (55.2)**	49.8 (45.7)	56.1 (49.1)	52.6 (47.1)
		Proso	64.5 (61.7)	58.5 (59.6)	**67.4 (63.1)**	61.1 (54.0)	61.2 (52.3)	65.4 (59.6)
DP	train: 1365/1365	Os385	68.7 (67.6)	56.8 (51.5)	**80.3 (80.4)**	59.2 (59.1)	73.8 (73.2)	79.3 (78.9)
	test: 502/378	Phono	69.6 (69.3)	58.9 (57.4)	**71.7 (70.7)**	57.9 (57.8)	66.7 (65.5)	66.6 (66.6)

4. Classification experiments

Models are trained on the balanced training subcorpus, optimized in cross-validation on the training subcorpus and evaluated on the test subcorpus. The remaining evaluation subcorpus is left for future evaluation with speech synthesis. Performances are measured on the test subcorpora with the unweighted average recall (UAR) and unweighted average F1 measure (UAF1).

4.1. Classification results

Classification rates obtained with the 6 models (GMM, OC-GMM, SVM, OC-SVM, RF, RF-BOOST), 3 feature sets (Os385, Phono, Proso) and 3 speech units (syllable, word, segment) are summarized in table 2. From this table, we learn that one-class models are not able to classify correctly discourse parts. Also, simple RF models underperform GMMs. However, the boosting technique helps in improving performance thus making RF-Boost better than GMMs. Finally SVMs seem to better classify discourse phrases whatever the feature set and the speech unit.

Concerning the feature sets, it appears that the difference in the number of features (385 vs. 14) has no significant impact on the results. For example with SVM models, Proso features are better than Os385 in classifying word units, whereas it is the contrary on syllable units. In all cases, Phono set reaches the worst results while being over the random guess. In section 4.3, we will see that the fusion of phonological and acoustic features is of interest.

No significant differences are noticeable between syllable and word speech units. However, the results obtained on discourse phrases are significantly higher. A reason can be the small number of test segments (see Table 2). But it is also known that emotion detection systems are usually designed at the segment level because it enables to model the whole prosodic contours of the sentence. It appears that such an approach could be relevant for discourse phrases classification also. Considering the limited number of samples for discourse phrase units, only word units classification will be investigated in the rest of the paper.

The classification results obtained on the French corpus, follow the same trends. SVM classification with Os385 reaches the best performances on syllables (65.6%), words (67.8%) and discourse phrases (76.7%). The followings describe two additional techniques we propose for improving classification of words units with SVMs. We remind the reader, that the classification of discourse speech units aims at designing an expressive speech synthesis system in the form of an expressivity score.

4.2. Addition of a threshold value and intermediate class

In GMM models decoding, we can easily integrate a log-likelihood threshold ε as it is usually done in speaker identification [24]. Together we propose to add an intermediate discourse class: words which do not strongly belong to direct or narrative phrases, are classified in this intermediate class, following eq. 5. This threshold can also be defined as a distance to hyperplane in the case of SVM models. If the distance is too small, the item belongs to the intermediate class.

$$c(x) = \begin{cases} \text{indirect} & \text{if } \Delta > \varepsilon \\ \text{direct} & \text{if } \Delta < -\varepsilon \\ \text{other} & \text{if } |\Delta| \leq \varepsilon \end{cases} \quad (5)$$

The introduction of this threshold brings two issues: first how to set the ϵ threshold value, second how to evaluate models. Indeed no ground truth is available for the intermediate class, thus making the classification rate obtained on the three classes out of purpose. In speaker verification techniques, the threshold is usually obtained using a set of impostors. In our case, the classification rate (here UAR) obtained on the direct and narrative words (without taking into account words classified in the intermediate class) will be used to evaluate models. As a consequence, the required threshold is a compromise between the UAR, and the ratio R between the number of words being classified as belonging to either direct or narrative phrases and the total number of words. In order to compare different settings, we propose to normalize the threshold between 0 and 100. In the results presented in table 3, only SVM models are used since they outperform the other tested models on this task.

This work aims at controlling expressivity in unit-selection TTS systems through discourse phrases. Therein, according to the context of the text to synthesize, the system should automatically select adequate speech units before concatenation. For example, if the text should be declaimed by a character of the story, the system should select speech units preferably in words belonging to the direct class. ϵ threshold must be adapted to the strength of the expressive constraint the user wants. If the user wants to emphasize the differences between dialogs and narration, he should select a high threshold, thus reducing the number of words correctly classified as belonging to direct or narrative phrases (R decreases) and increasing the UAR.

In order to have a general idea, the results obtained when the intermediate class contains $R = 50\%$ and $R = 25\%$ of the words are reported in the first three lines of Table 3. The introduction of an intermediate class induces an average increase of 5.4 percentage point (pp) when $R \simeq 50$ and of 10.1 pp when

Table 3: *Classification results with SVMs without and with a threshold $\hat{\varepsilon}$ in the decision value and weights optimization for each model fusion. $\hat{\varepsilon}$: normalized threshold, UAR: unweighted averaged recall [percentage points increase], R: relative number of items correctly classified in direct or indirect speech classes.*

	Set	#feat.	UAR ($\hat{\varepsilon}=0$)	$\hat{\varepsilon}$	UAR	R (%)	$\hat{\varepsilon}$	UAR	R (%)
Single	Os385	385	64.8 [0.0]	7	69.3 [4.5]	53.8	11	77.4 [12.8]	27.6
	Phono	6	58.2 [0.0]	19	62.3 [4.1]	51.6	24	65.4 [7.2]	26.1
	Proso	14	67.4 [0.0] [0.0]	22	75.2 [7.8]	50.0	27	77.7 [10.3]	24.5
Fusion balanced weights	Proso+Os385	399	68.9 [1.5]	18	78.5	52.5	30	80.3	24.5
	Proso+Phono	20	68.1	25	73.7	50.2	39	79.2	25.0
	Os385+Phono	391	58.2	19	62.5	52.1	24	66.1	26.5
	Os385+Phono+Proso	419	70.4 [3.0]	26	77.9	50.4	**40**	**82.8**	**26.0**
Fusion optimized weights	Proso+Os385	399	70.4 [3.0]	14	77.9	49.8	24	81.0	24.7
	Proso+Phono	20	69.1	25	74.8	51.9	40	79.4	24.4
	Os385+Phono	391	68.5	10	74.1	50.0	18	77.8	25.7
	Os385+Phono+Proso	419	**71.8** [4.4]	**20**	**79.9**	**50.4**	32	82.6	25.1

$R \simeq 25$. In some cases, the UAR reaches an optimal maximum at a given R then decreases. This means that classification performances can not be improved by a sole threshold.

4.3. Fusion of the different models

SVM models trained with different feature sets are merged at the prediction function level. Precisely, distances to hyperplanes are merged using a weighted sum. Labels are still predicted according to eq. 5. The weights are either set to 1 (balanced case) or optimized for each ε (weighted case) as detailed in Table 3. The fusion of predicted classes have also been tested but no significant results emerged from that. The fusion of acoustic and prosodic features allows to improve model's performance of 1.5 pp. from Proso only ($\hat{\varepsilon}=0$) and the addition of phonological features deeper improves the performance of 3.0 pp. from Proso only. Weights optimization helps to improve even more the classification rates thus reaching a global improvement of 4.4 pp. Also, the addition of a threshold leads to better performance. However, it seems that weight optimization has very few impact when the threshold increases.

5. Discussion

We can see on Figure 1 that in reality the prediction function $\mathcal{M}$ better fits with a continuous representation than discrete categories. However discrete classification allows to find the best representation for discourse phrases. Blue crosses correspond to samples labelled as direct (left) or narrative (right) phrases while red points correspond to the intermediate class. Interestingly, when the prediction function decreases, F_0 values are more scattered. Indeed, it is related to the high variations in intonation in direct phrases.

6. Conclusion

The global aim of the presented experiments is to control expressivity in a unit-selection TTS system through discourse phrases. To do so, the speech units contained in the voice corpus are classified as belonging to either direct, narrative or intermediate phrases. Different models (two classes, one class), feature sets (acoustic, prosodic or phonological) and speech units (syllable, word and discourse phrase) are tested. SVM with prosodic features gives the best performance. The experiments

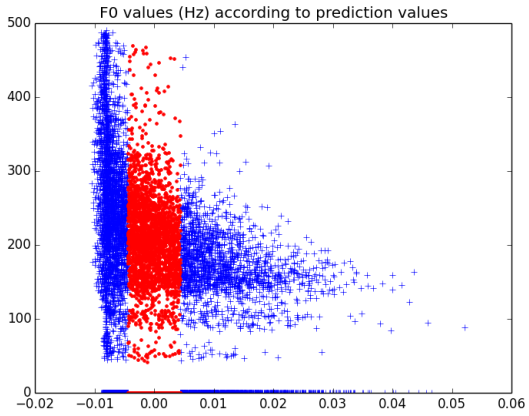

Figure 1: *$F_0(x)$ w.r.t. $\mathcal{M}(x)$ obtained with SVM trained with Os385 on words units.*

at the word level show that the fusion of the 3 feature sets increases the classification rate from 4.4 pp. Also the addition of a decision threshold achieves promising results for discourse phrases classification: 71.8% with 100% of the words, 79.9% with 50% and 82.6% with 25%.

The results show that classification rates were better at the phrase level than at the word or syllable level, in agreement with emotion recognition studies. To solve the problem of the relatively low number of units in that case, we could consider some context at the word level.

This work is part of a project on expressive speech synthesis. With this in mind, we plan to synthesize speech using the proposed expressivity score. To do so, we need to introduce a new cost to the concatenation and target costs used in our unit-selection system.

7. Acknowledgements

This study has been realized under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015 and also with the support of the funding AtlanSTIC 2020 from the region Pays de la Loire.

8. References

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2016," in *Blizzard Challenge (satellite of Interspeech)*, Cupertino, USA, 2016.

[2] S. King, L. Wihlborg, and W. Guo, "The Blizzard Challenge 2017," in *Blizzard Challenge (satellite of Interspeech)*, Stockholm, Sweden, 2017.

[3] P. Alain, N. Barbot, J. Chevelu, G. Lecorvé, D. Lolive, C. Simon, and M. Tahon, "The irisa text-to-speech system for the blizzard challenge 2017," in *Blizzard Challenge (satellite of Interspeech)*, Stockholm, Sweden, 2017.

[4] M. Charfuelan and I. Steiner, "Expressive speech synthesis in MARY TTS using audiobook data and EmotionML," in *Interspeech*, Lyon, France, 2013, pp. 1564–1568.

[5] E. Székely, J. Kane, S. Scherer, C. Gobl, and J. Carson-Berndsen, "Detecting a targeted voice style in an audiobook using voice quality features," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4593–4596.

[6] J. Y. Zhang, A. W. Black, and R. Sproat, "Identifying speakers in children's stories for speech synthesis," in *EuroSpeech*, Geneva, Switerland, 2003, pp. 2041–2044.

[7] E. Székely, J. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Interspeech*, Firenze, Italy, 2011, pp. 2409–2412.

[8] D. Doukhan, A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro, "Prosodic analysis of a corpus of tales," in *Interspeech*, Firenze, Italy, 2011, pp. 3129–3132.

[9] F. Eyben, S. Buchholz, N. Braunschweiler, V. W. J. Latorre, M. J. F. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive tts," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4009–4012.

[10] Éva Székely, T. G. Csapó, T. Bálint, P. Mihajlik, and J. Carson-Bernsden, "Synthesizing expressive speech from amateur audiobook recordings," in *Spoken Language Technology Workshop (SLT)*. Miami, Florida, USA: IEEE, 2012, pp. 297–302.

[11] L. Chen and M. Gales, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Interspeech*, Portland, USA, 2012, pp. 959–962.

[12] R. M. no and F. Alías, "The role of prosody and voice quality in indirect storytelling speech: Annotation methodology and expressive categories," *Speech Communication*, vol. 85, pp. 8–18, 2016.

[13] J. Ouyang and K. McKeown, "Towards automatic detection of narrative structure," in *LREC*, Reykavik, Island, 2014, pp. 4624–4631.

[14] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interpseech*, Brighton, UK, 2009, pp. 312–315.

[15] M. Schröder, *Expressive Speech Synthesis: Past, Present, and Possible Futures*. London: Springer London, 2009, pp. 111–126.

[16] H. Kanagawa, T. Nose, and T. Kobayashi, "Speaker-independent style conversion for HMM-based expressive speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 7864–7868.

[17] Y.-Y. Chen, C.-H. Wu, and Y.-F. Huang, "Generation of emotion control vector using MDS-based space transformation for expressive speech synthesis," in *Interspeech*, San Fransisco, USA, 2016, pp. 3176–3180.

[18] P. Alain, J. Chevelu, D. Guennec, G. Lecorvé, and D. Lolive, "The IRISA Text-To-Speech system for the Blizzard Challenge 2016," in *Blizzard Challenge (satellite of Interspeech)*, Cupertino, USA, 2016.

[19] I. Steiner, M. Schröder, M. Charfuelan, and A. Klepp, "Symbolic vs. acoustics-based style control for expressive unit selection," in *ISCA Speech Synthesis Workshop (SSW7)*, Kyoto, Japan, 2010.

[20] M. S. Ribeiro, O. Watts, J. Yamagishi, and R. A. J. Clark, "Wavelet-based decomposition of F0 as a secondary task for DNN-based speech synthesis with multi-task learning," in *International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016, pp. 5525–5529.

[21] S. Pammi and M. Charfuelan, "HMM-based sCost quality control for unit selection speech synthesis," in *ISCA Speech Synthesis Workshop*, Barcelona, Spain, 2013, pp. 53–57.

[22] A. Sini, D. Lolive, G. Vidal, M. Tahon, and E. Delais-Roussarie, "SynPaFlex-Corpus: An expressive French audiobooks corpus dedicated to expressive speech synthesis." in *LREC*, Myazaki, Japan, 2018.

[23] C. Fayet, A. Delhay, D. Lolive, and P.-F. Marteau, "Big five vs. prosodic features as cues to detect abnormality in SSPNET-personality corpus," in *Interspeech*, Stockholm, Sweden, 2017, pp. 3281–3285.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, vol. 10, 2000, pp. 19–41.