Combining Manual and Automatic Prosodic Annotation for Expressive Speech Synthesis

Sandrine Brognaux[1,2,3,4], **Thomas François**[1,4], **Marco Saerens**[2]
[1]Cental, IL&C, Université catholique de Louvain (UCL), Belgium
[2]ICTEAM, Université catholique de Louvain (UCL), Belgium
[3] TCTS, Université de Mons (UMons)
[4] FNRS, Belgium
sandrine.brognaux@umons.ac.be, thomas.francois@uclouvain.be, marco.saerens@uclouvain.be

Abstract

Text-to-speech has long been centered on the production of an intelligible message of good quality. More recently, interest has shifted to the generation of more natural and expressive speech. A major issue of existing approaches is that they usually rely on a manual annotation in expressive styles, which tends to be rather subjective. A typical related issue is that the annotation is strongly influenced – and possibly biased – by the semantic content of the text (e.g. a shot or a fault may incite the annotator to tag that sequence as expressing a high degree of excitation, independently of its acoustic realization). This paper investigates the assumption that human annotation of basketball commentaries in excitation levels can be automatically improved on the basis of acoustic features. It presents two techniques for label correction exploiting a Gaussian mixture and a proportional-odds logistic regression. The automatically re-annotated corpus is then used to train HMM-based expressive speech synthesizers, the performance of which is assessed through subjective evaluations. The results indicate that the automatic correction of the annotation with Gaussian mixture helps to synthesize more contrasted excitation levels, while preserving naturalness.

Keywords: speech synthesis, expressive style annotation, machine learning

1. Introduction

Text-to-speech has long been centered on the production of an intelligible message of good quality. However, as early as the 90's, some early work already investigated the issue of generating more natural and expressive speech (Murray and Arnott, 1993). Today, this has become a major goal in the field as it comes as an answer to the widespread criticism towards current speech synthesizers, which have been said to be unnatural and to lack fundamental human components (Kim et al., 2004; Campbell, 2006).

In this framework, a large body of research has focused on the synthesis of basic emotions like happiness or anger (Yamagishi et al., 2004; Hirose et al., 2005; Qin et al., 2006), which can be seen as a simplistic reduction of the expressive naturalness lacking in speech synthesizers (Campbell, 2006). Conversely, we believe that synthesizers should include other types of expressiveness, such as various speaking styles (e.g. TV speech, sports commentaries, political discourse) or attitudes.

Surprinsingly, this topic has raised much less interest. Lorenzo-Trueba et al. (2013) and Obin et al. (2011) targeted HMM-based speech synthesis of various speaking styles like broadcast news or political speech. In the same vein, Eyben et al. (2012) proposed modeling different speaking styles found in audiobooks. The synthesis of sports commentaries has also aroused some interest (Krstulović et al., 2007; Picart et al., 2013; Brognaux et al., 2013). Specifically, it is proposed in Picart et al. (2013), to annotate basketball commentaries in various excitation levels like "Negative tension", "Excited" or "Neutral" to train style-specific synthesizers.

A major issue of such studies, is that they rely on a manual annotation which is usually rather subjective. In the framework of sports commentary synthesis (Brognaux et al., 2013), for instance, the manual macro-prosodic annotation of the corpus in terms of excitation level and valence was shown to achieve low inter-annotator agreement scores, with a Cohen's kappa of 0.38. It is likely that the manual annotation of excitation levels is strongly influenced by the semantic content of the corresponding speech segments. A successful shot, for instance, may have been considered as expressing a high degree of excitation by the annotators, while a section describing a player biography would be annotated as neutral[1], even if the corresponding speech segments did not display the characteristic acoustic features of the corresponding excita-

[1]By the term 'neutral', we refer in this work to the standard level of excitation in sports commentaries, which may appear more excited than the standard level in other types of speech.

tion level.

An interesting feature of the annotation in terms of level of arousal/excitation is that acoustic features (e.g. pitch, timing, energy and voice quality) of the speech signal seem to be highly correlated with the excitation level, which allows for a highly accurate automatic classification between high-activation and low-activation segments (Tato et al., 2002; Liscombe et al., 2003). Schröder et al. (2001) also found that activation is the emotional dimension that is most correlated with specific acoustic realizations. They showed, for example, that activated speech displays higher F0 mean and range, longer and faster increases and decreases in F0, and increased intensity. By capturing long-term dependencies between the acoustic observations derived from hierarchical functionals of prosodic, spectral, and voice quality features, Wöllmer et al. (2008) also highlighted that it is possible to predict the level of arousal with an accuracy similar to human performance.

Drawing from these findings, we propose to investigate, in this work, the possibility of using an automatic labeling of excitation levels, based on acoustic features only, to correct semantically-driven "outliers" in the manual annotation, that is, word segments that were wrongly annotated as neutral or excited due to their semantic content. More precisely, our objective is to determine whether the homogeneity of human annotations can be improved thanks to an automatic reclassification system, thereby providing a more acoustically consistent annotation. This is in line with previous studies by Brognaux et al. (2012), in which the authors proposed a post-processing technique to reduce the number of prosody labeling errors. It was shown that a predictor trained on a (partially erroneous) manual annotation could be used to check existing labels, generalize global tendencies and improve the annotation quality. Such an approach amounts to leveraging the existing manual annotation to generalize global tendencies and then automatically correct outliers.

This paper is organized as follows: Section 2. presents our corpus and its manual annotation. In Section 3., we discuss the level at which the excitation level annotation should be provided. Section 4. investigates two automatic classification methods to predict labels for excitation level and correct the manual annotation of the corpus. Finally, Section 5. evaluates the integration of the corrected macro-prosodic annotation in HMM-based speech synthesis.

2. Corpus design

This study is based on the *Sportic* corpus (Brognaux et al., 2013), which contains the commentaries of two basketball matches by a professional French commentator, recorded in sound-proof conditions. The speaker watched the games and commented them without any prompting. Both matches star the *Spirou*, a popular Belgian team, and are characterized by very tight final scores, which induced a high level of excitation. The issue with sport commentary corpora is usually the high level of background noise that precludes their precise acoustic analysis (Trouvain, 2011). Conversely, our corpus exhibits the advantage of being spontaneous and of high acoustic quality, therefore being suitable for speech synthesis.

The total corpus duration is 162 minutes, with a high proportion of silences. Several processes were applied. First, the corpus was transcribed into a text version whose phonetization was automatically produced by the eLite-HTS system (Beaufort, 2008). The phonetic transcription was manually corrected and then automatically aligned with the sound with the Train&Align tool (Brognaux et al., 2012). This alignment took advantage of the bootstrap option of Train&Align to reach alignment rates higher than 80 % with a 20ms tolerance threshold. The eLite NLP system produced other required annotation tiers (*e.g.* syllables, parts of speech, rhythmic groups). For the detection of sentence boundaries, which is a rather complex task as we do not have access to punctuation, the corpus was manually annotated to define segments corresponding to both a prosodic and a semantic completeness.

As regards the speaking style annotation of the corpus, which we refer to as the macro-prosodic annotation, we followed the method described in Brognaux et al. (2013). In short, groups of words were assigned an excitation level, based on a dimensional analysis of emotions in terms of valence and arousal (Mehrabian and Russel, 1974; Russell, 1980). A unique annotator took care of the whole corpus, but another expert also annotated 20% of the data already seen in order to evaluate agreement on this task. Inter-annotator agreement was low (kappa = 0.38), mainly due to logical confusion between contiguous levels. For the current study, we relied on a simplified version of the original annotation that results from a discretization of the arousal (excitation) level into three degrees: Neutral, Excited and ExMax. The distribution of the corpus across these three levels is as follows (with silences longer than 1 second excluded): Neutral (2955 sec.), Excited (1032 sec.) and ExMax (475 sec.).

3. Choice of a minimal unit

An excitation level, or macro-prosodic unit, must be assigned to a group of words. The main challenge is

therefore to determine where an excitation level begins and where it ends.

3.1. Minimal unit in the literature

This issue has been overlooked in the literature, notably because most studies are based on acted speech in which each sentence is pronounced with a prompted emotion (Amir et al., 2001). Even for more realistic data, classification is often related to pre-segmented chunks, based on the manual annotation of emotions (Schuller et al., 2003).

The question of the minimal unit of analysis has also been investigated by Schuller et al. (2011). They point out that the speaking turn is often considered in conversational speech, but highlight the fact that turns can sometimes be rather long and contain several shorter emotional episodes. They distinguish two strategies to cope with this problem: *i.e.* the use of 'technical units', which simply corresponds to either time windows or fixed proportions of longer units; and the use of 'meaningful units', which are linguistically and semantically well-defined (e.g. syllables, words, or phrases). Batliner et al. (2010) investigated three different units: the word level or *ememe* (considered as the smallest meaningful emotional unit), the syntactic chunk, or the ememe chunks (cluster of adjacent ememes belonging to the same arousal class). The latter relies on a manual annotation of the corpus and is hardly automatizable, but their performance was shown to outperform that obtained with syntactic chunks. The authors indicate that the best compromise between automation and performance seems to rely on the use of word units.

Automatic clustering of various speaking styles in audiobooks has also pointed out the need for defining a minimal unit. Eyben et al. (2012) propose to work at the sentence level, which they consider to be the longest possible chunk in that respect. It is worth mentioning that segmenting sentences is more straightforward in the case of audiobooks as it can rely on the punctuation of the text, while we have no punctuation information in spontaneous speech. In a similar study, Székely et al. (2011) pointed out that significant reading style changes may occur within a single sentence. This was an incentive to choose interpausal units instead, which avoided abrupt changes of voice style within a speech segment.

3.2. Experimenting with the minimal unit

Facing this divergence in the literature, we carried out an analysis, based on the manual prosodic annotation of the *Sportic* corpus, to determine the best linguistic unit to use. We compared five different unit levels, namely interpausal units (IPU), rhythmic groups (RG), accentual phrases, words, and sentences[2]. IPU and RG are middle-sized chunks that should provide a good compromise of being short enough to ensure emotion stability, but of sufficient duration to compute global prosodic measures, such as articulation rate or accentual density, which have been shown to be correlated with the activation level (Murray and Arnott, 1993; Schröder et al., 2001). IPU relies on the manually checked phonetization of the corpus in which silences are indicated. The segmentation of RGs is provided by eLite-HTS, which proposes a phrasing algorithm based on an improved version of Liberman's chinks & chunks (Liberman and Church, 1992; Beaufort, 2008). As regards accentual phrases, we defined two variants of this unit. On the one hand, they were simply considered as sequences of words ending with a boundary tone (*i.e.* H, L, LL, HH or E as defined in (Brognaux et al., 2013)). On the other hand, we defined a specific type of accentual phrase (which could be considered as closer to intonational phrases) that only considers groups of words ending in higher level boundaries (*i.e.* LL and HH tones).

In our experiments, we computed the percentage of manually-annotated boundaries of excitation level segments in our corpus that corresponded to boundaries of each of the six minimal units considered. Results are shown in Figure 1. Bars with a red border (*i.e.* accentual phrases and sentences) correspond to information that can only be obtained with the help of a manual annotation. Ideally, our study should be based on automatically segmented units (in green) such as RGs or IPUs.

Our study shows that 91.5 % of the boundaries of excitation level segments also correspond to RG boundaries, while it is only the case for 82.8% of the IPU boundaries. Other types of considered units get an even lower correspondence percentage, which makes RG the best segmentation level for our purposes.

4. Automatic prediction of macro-prosodic labels

The experiments described in the previous section convinced us to use a segmentation of our corpus in rhythmic groups for our study. The next step consisted in automatically reclassifying all rhythmic groups of the corpus in terms of excitation levels based on acoustic features. To this aim, we defined and extracted a large set of acoustic features (see Section

[2]Dialog acts is another unit that could have been considered, but we did not do so as our corpus only includes commentaries which are soliloquies .

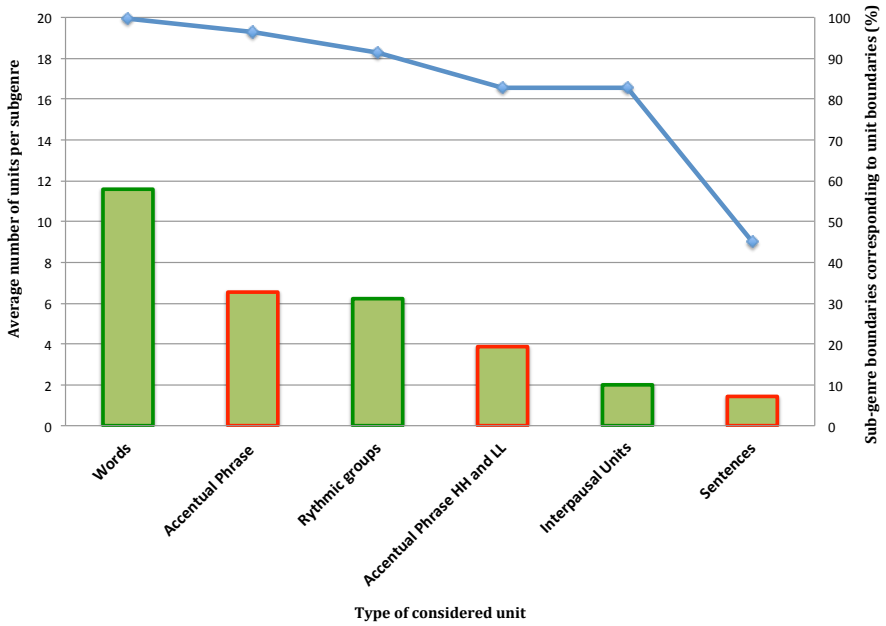

Figure 1: Connection between linguistic segmentation and prosodic macro changes.

4.1.). Then, we compared two classification algorithms: a Bayes classifier based on a simple Gaussian mixture and an ordinal logistic regression (see Section 4.2.).

4.1. Feature extraction

A set of 27 global acoustic features was automatically extracted at the rhythmic group level (3,999 RGs were considered). The set includes the following continuous features:

- **Pitch** : mean, median, mode, min, max, interquartile range, and variance.
- **Energy** : mean, median, mode, min, max, interquartile range, and variance.
- **Duration** : mean phone duration, mean syllable duration, mean vocalic nucleus duration, mean normalized syllable duration, and their variance.
- **Pauses** : length of the preceding and following silence. If the RG is not preceded and/or followed by a silence, this parameter is set to 0.
- **Complex values** : percentage of prominent syllables (as detected by Prosoprom (Goldman et al., 2007)), average difference of pitch between a syllable and the preceding syllable (*delta pitch*), in absolute value (as a measure of pitch dynamism).
- **Speaking rate** : number of syllables per second (silences do not need to be considered/excluded here as RGs never include silences).

Three post-processing operations were applied in sequence to the raw computed values: (1) a mean imputation of missing values, (2) the normalization of all values as z-scores, and (3) a dimensionality reduction with a principal component analysis (PCA) from which only the first 6 components were kept[3]. We emphasize the fact that no linguistic features were integrated into our feature set as our motivation was to automatically distinguish segments of speech that were *acoustically* different in order to improve the training of the synthesizer. Linguistic contexts that may have erroneously influenced the manual annotators should therefore be discarded.

4.2. Classification methods

To automatically predict the level of excitation based on the six features resulting from the PCA, we compared two classification algorithms: (1) a Bayesian approach based on a simple Gaussian mixture and (2) an ordinal logistic regression model, also called proportional odds model (Agresti, 2002).
The **Bayesian approach** re-assigned to each rhythmic group the excitation level ($e_{new}(v)$) that maximizes the following a posteriori probability:

[3]To select the n-relevant first dimensions, we used a popular rule of thumb, namely keeping the dimensionality accounting for 70% of the total variance.

$$e_{new}(v) = \arg\max_{e} P(e \mid v) = \frac{P(v \mid e) \cdot P(e)}{P(v)} \quad (1)$$

where $e \in E = \{Neutral, Excited, \text{and } ExMax\}$ and v is the feature vector with the 6 PCA-reduced dimensions. As we are interested in the argmax values, the quantity $P(v)$ may be neglected. The a priori probability $P(e)$ of each level of excitation e in the initial data is computed as:

$$P(e) = \frac{nb(e)}{\sum_{\forall e \in E} nb(e)} \quad (2)$$

where $nb(e)$ is the number of occurrences of level e in the corpus. As regards the conditional probability $P(v|e)$, it is computed as follows:

$$\begin{aligned} P(v \mid e) &= f(v, \mu_e, \Sigma_e) \\ &= \frac{1}{\sqrt{|\Sigma_e|(2\pi)^d}} e^{-\frac{1}{2}(v-\mu_e)^T \Sigma_e^{-1}(v-\mu_e)} \end{aligned}$$

where d is the dimensionality of v (6, in our case); μ_e is a 1-by-6 vector, containing average values of the 6 components for the level of excitation e. Σ_e is the corresponding 6-by-6 covariance matrix. This allows considering the mean and variance of each class. It should however be noted that the variance-covariance matrices of all three classes are rather similar and should not greatly influence the classification task.

For the **logistic regression** approach, we trained an ordinal logistic regression model (Agresti, 2002) that takes advantage of the natural ordering across the response categories. Based on that model, a level of excitation is also predicted for each vector.

Tables 1 and 2 show the contingency matrices resulting from the reclassification of the 3,999 RGs by both models[4]. We see that the logistic model is biased towards Neutral, with a high reclassification of manual ExMax labels. The Bayesian approach is more balanced, but also tends to reclassify Excited RG as Neutral.

For each model, a Krippendorf alpha (Hayes and Krippendorff, 2007) for ordinal data was computed between the model's predictions and the manual labels to assess the proportion and degree of reclassification. We obtained an alpha of 0.4016 for the Bayes approach and of 0.3030 for the logistic regression, which seems to indicate that the Bayes model remains closer to the original annotation.

As the results provided by both methods tend to go in the same direction, we also compared the classes they predict for each rhythmic group (RG). The comparison indicated that both predictors assign an identical label in 70.48% of the cases. Interestingly, the adjacent accuracy, defined as the proportion of prediction with maximum one level of error (Heilman et al., 2008), reaches 99.02%, which shows their high consistency.

5. Perception evaluation

The main goal of this paper is to test whether automatically re-annotating a corpus manually annotated in terms of excitation levels can improve the quality of speaking style-adapted HMMs for speech synthesis (here one for each excitation level). To this aim, we built three synthesizers of a male voice, respectively based on the manual annotation (*Baseline*), the automatic re-annotation with Bayes (*Auto1*) and the automatic re-annotation with logistic regression (*Auto2*). All implementations were done with the HTS toolkit (version 2.1) (Zen et al., 2007). The training of each synthesizer followed the same protocol: an average-voice model, trained on the whole corpus, was adapted with Constrained Maximum Likelihood Linear Regression (CMLLR) (Gales, 1998) to the sub-part of the corpus corresponding to each excitation degree. The linearly transformed models were further optimized using MAP adaptation (Yamagishi et al., 2009). This method was shown in Picart et al. (2013) to produce the best results in contrast to other training and adaptation approaches. The synthesis set is made of rhythmic groups (RG) or sequences of rhythmic groups which are assigned an identical label in the manual annotation and are not separated by a silence. For each synthesizer, 90% of the corresponding data was used for the training, leaving around 10% for the test.

Based on this test set, two perception tests were carried out: the first assessed the synthesis naturalness, whereas the second evaluated whether the three levels of excitation were better discriminated by one of the three synthesizers. 20 native French speakers, mainly naive listeners, participated in the evaluation. Each of them was provided with 18 items (consisting in pairs of sentences) for each test, selected from the test set with a stratified sampling, in order to balance excitation levels and synthesizers.

In the first test, which focused on voice quality, the two sentences forming an item were generated with a different synthesizers. Listeners were then asked to indicate which version seemed more natural in the con-

[4]As we are primarily interested in automatically re-annotating the corpus rather than defining a maximally accurate model, results are reported on all data.

		Automatic annotation			
		Neutral	Excited	ExMax	Total
Manual annotation	Neutral	2,018	338	101	2,457
	Excited	627	315	118	1,060
	ExMax	129	112	239	480
	Total	2,774	765	458	3,997

Table 1: Contingency table between manual annotation and automatic reclassification with Gaussian mixture models.

		Automatic annotation			
		Neutral	Excited	ExMax	Total
Manual annotation	Neutral	2,253	191	13	2,457
	Excited	813	229	18	1,060
	ExMax	184	200	96	480
	Total	3,250	620	127	3,997

Table 2: Contingency table between manual annotation and automatic reclassification with logistic regression.

text of sports commentaries. The voice quality scale ranged from -3 (sentence A is much less natural than sentence B) to 3 (sentence A is much more natural than sentence B). If both versions sounded equally natural (or if they both sounded mediocre), subjects could rate them as "equivalent". The preferences expressed by the subjects during this first test were very slight. Significance measures were computed, with a unilateral signed rank sum test comparing the average percentage of preferences on the 20 testers, and no preference was shown to be significant ($p = 0.10$, $p = 0.06$ and $p = 0.36$ for pairs Auto1/Baseline, Auto2/Baseline and Auto1/Auto2 respectively). This indicates that using the corrected annotation does not enhance the naturalness of the synthesized speech.

In the second test, the task was to discriminate between different levels of excitation for a given synthesizer. The test also consisted in 18 pairwise comparisons between combination of sub-genres, *i.e.* Neutral/Excited, Excited/ExMax and Neutral/ExMax. The two synthesized sentences, in each pair, were this time synthesized with the same synthesizer, but with a different level of excitation. For each pair, the listener was asked in which version the commentator sounded more excited. The scale ranged from -3 (sentence A sounds much less excited than sentence B) to 3 (sentence A sounds much more excited than sentence B). In case of similar level of excitation, the "equivalent" label could be chosen.

Results of the second perception test offered more insightful results. Figure 2 indicates the percentage of cases in which the synthesis of each excitation level (*i.e.* Neutral, Excited and ExMax) was considered as the most excited, in the pairwise comparison. Ideally, Neutral should be assigned 0%, Excited 50% (*i.e.* more excited than neutral, but less excited than ExMax), and ExMax 100%. The results clearly indicate that the baseline suffers from a lack of discrimination between Neutral and Excited syntheses ($p = 0.09$ with a one-sided signed rank sum test). The logistic model (*Auto2*) tends to discriminate more effectively between Neutral vs. Excited and ExMax, but has trouble distinguishing between the two latter ones. The *Auto1* model (Bayes) offers the best discrimination between the three levels of excitation, being close to the optimal situation.

These observations indicate that the automatic correction of the manual annotation of the corpus, based on simple Gaussian mixture, clearly allows for better discrimination between the synthesis of the three excitation levels, while offering a perceived naturalness similar to the baseline.

6. Conclusion

In this paper, we have investigated the assumption that human annotation of basketball commentaries for excitation levels can be automatically improved on the basis of acoustic features. We presented two techniques for label correction, using Gaussian mixture models and a proportional-odds logistic regression. Our perception evaluation showed that the corpus reannotated with the Gaussian mixture models helps to synthesize more contrasted excitation levels while preserving similar naturalness.

Further studies on this topic could investigate the integration of contextual information into the models, i.e. by considering the sequential aspect of the units to better predict the excitation level. To this aim, the approach based on Gaussian mixture models could be in-

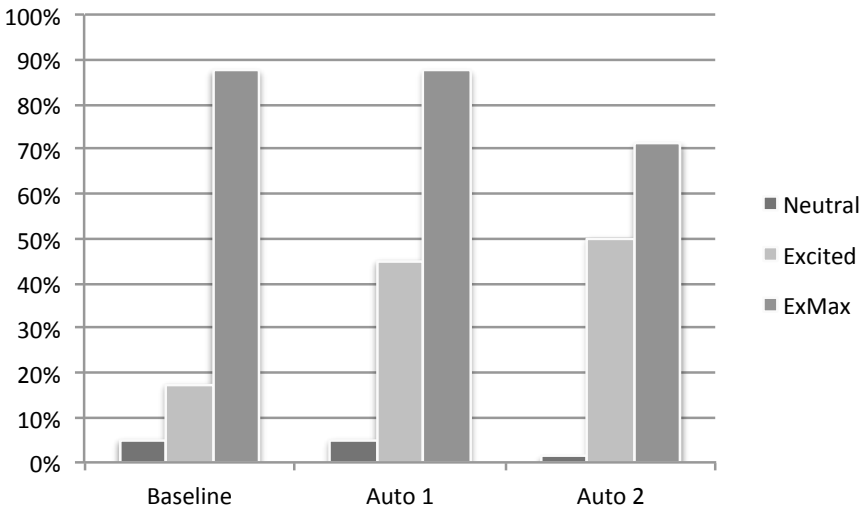

Figure 2: Percentage of cases in which each excitation level is perceived as the most excited

tegrated into an HMM framework, in which the emission probabilities would be extracted from the model. We could also investigate the integration of linguistic features into the model although we tend to believe that integrating non-acoustic predictors might degrade the quality of the resulting synthesizers.

7. Acknowledgements

S. Brognaux and T. François are supported by FNRS respectively as an "Aspirant FNRS" and a "Chargé de recherches FNRS". The project was partly funded by the Walloon Region WIST 3 SPORTIC and partly funded by the Region of Bruxelles-capitale (INNOVIRIS). We also want to thanks Dr. André Bitar for his valuable comments.

8. Bibliographical References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience, New York, 2nd edition.

Amir, N., Kerret, O., and Karlinski, D. (2001). Classifying emotions in speech: A comparison of mehtods. In *Proc. of Interspeech*.

Batliner, A., Steidl, S., Seppi, D., and Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. *Advances in Human-Computer Interaction*, 3.

Beaufort, R. (2008). *Application des Machines à Etats Finis en Synthèse de la Parole. Sélection d'unités non-uniformes et Correction orthographique*. Ph.D. thesis, FUNDP, Namur.

Brognaux, S., Roekhaut, S., Drugman, T., and Beaufort, R. (2012). Train&Align: A new online tool for automatic phonetic alignments. In *Proc. IEEE Workshop on Spoken Language Technologies (SLT)*.

Brognaux, S., Drugman, T., and Beaufort, R. (2012). Automatic detection of syntax-based prosody annotation errors. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*.

Brognaux, S., Picart, B., and Drugman, T. (2013). A new prosody annotation protocol for live sports commentaries. In *Proc. of Interspeech*.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Acoustics Speech and Language Processing*, 14(4):1171–1178.

Eyben, F., Bucholz, S., and Braunschweiler, N. (2012). Unsupervised clustering of emotion and voice styles for expressive TTS. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer, Speech and Language*, 12(2):75–98.

Goldman, J.-P., Avanzi, M., Lacheret-Dujour, A., Simon, A. C., and Auchlin, A. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken French. In *Proc. of Interspeech*, pages 98–101.

Hayes, A. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1:77–89.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proc. of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*.

Hirose, K., Sato, K., Asano, Y., and Minematsu, N.

(2005). Syntesizing of F0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis. *Speech Communication*, 46:385–404.

Kim, Y. J., Syrdal, A., and Jilka, M. (2004). Improving TTS by higher agreement between predicted versus observed pronunciations. In *Proc. of the 5th ISCA Workshop on Speech Synthesis (SSW5)*.

Krstulović, S., Hunecke, A., and Schröder, M. (2007). An hmm-based speech synthesis system applied to German and its adaptation to a limited set of expressive football anouncements. In *Proc. of Interspeech*.

Liberman, M. and Church, K. (1992). Text analysis and word pronunciation in text-to-speech synthesis. In Sadaoki Furui et al., editors, *Advances in Speech Signal Processing*, pages 791–831. Dekker, New York.

Liscombe, J., Venditti, J., and Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. In *Proc. of Eurospeech*.

Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., and Montero, J. M. (2013). Towards speaking style transplantation in speech synthesis. In *Proc. of the 8th ISCA Workshop on Speech Synthesis (SSW8)*.

Mehrabian, A. and Russel, J. A. (1974). *An Approach to Environmental Psychology*. MIT Press.

Murray, I. and Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93(2):1097–1108.

Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011). Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation. In *Proc. of Interspeech*.

Picart, B., Brognaux, S., and Drugman, T. (2013). HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation. In *Proc. of the 8th ISCA Workshop on Speech Synthesis (SSW8)*.

Qin, L., Ling, Z.-H., Wu, Y.-J., Zhang, B.-F., and Wang, R.-H. (2006). HMM-based emotional speech synthesis using average emotion models. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 233–240.

Russell, J. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proc. of Interspeech*.

Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087.

Székely, E., Cabral, J. P., Cahill, P., and Carson-Berndsen, J. (2011). Clustering expressive speech styles in audiobooks using glottal source parameters. In *Proc. of Interspeech*, pages 2409–2412.

Tato, R., Santos, R., Kompe, R., and Pardo, J. M. (2002). Emotional space improves emotion recognition. In *Proc. of Interspeech*.

Trouvain, J. (2011). Between excitement and triumph - Live football commentaries in radio vs. TV. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of Interspeech*.

Yamagishi, J., Tachibana, M., Masuko, T., and Kobayashi, T. (2004). Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5–8.

Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., and Renals, S. (2009). Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Audio, Speech and Language Processing*, 17(6):1208–1230.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 294–299.