HMM-based Speech Synthesis of Live Sports Commentaries: Integration of a Two-Layer Prosody Annotation

Benjamin Picart [1], *Sandrine Brognaux* [2], *Thomas Drugman* [1]

[1] TCTS - Université de Mons, Belgium
[2] Cental, ICTEAM - Université Catholique de Louvain, Belgium

benjamin.picart@umons.ac.be, sandrine.brognaux@uclouvain.be, thomas.drugman@umons.ac.be

Abstract

This paper proposes the integration of a two-layer prosody annotation specific to live sports commentaries into HMM-based speech synthesis. Local labels are assigned to all syllables and refer to accentual phenomena. Global labels categorize sequences of words into five distinct speaking styles, defined in terms of valence and arousal. Two stages of the synthesis process are analyzed. First, the integration of global labels (i.e. speaking styles) is carried out either using speaker-dependent training or adaptation methods. Secondly, a comprehensive study allows evaluating the effects achieved by each prosody annotation layer on the generated speech. The evaluation process is based on three subjective criteria: intelligibility, expressivity and segmental quality. Our experiments indicate that: (i) for the integration of global labels, adaptation techniques outperform speaking style-dependent models both in terms of intelligibility and segmental quality; (ii) the integration of local labels results in an enhanced expressivity, while it provides slightly higher intelligibility and segmental quality performance; (iii) combining the two levels of annotation (local and global) leads to the best results. It is indeed shown that it obtains better levels of expressivity and intelligibility.

Index Terms: HMM-based Speech Synthesis, Speaking Style Adaptation, Expressive Speech, Prosody, Sports Commentaries

1. Introduction

Expressive speech synthesis based on Hidden Markov Models (HMMs) has been the focus of many studies in the last ten years (e.g. [1], [2], [3] [4]). Conversely to unit-selection based synthesis, HMM-based synthesis [5] [6] offers a rich playground in terms of controllability of the generated speech. However, current research presents a certain number of drawbacks.

First, expressivity is often exclusively generated via adaptation or training on corpora with the targeted expressivity (e.g. [1], [2]). Generally, the training or the feature adaptation is achieved globally, with no consideration of local phenomena specific to expressive speech, like accentuation. However, it is widely acknowledged that the accentual structure of a sentence and the realization of focusses play a crucial role in the expressive function. Fonagy [7] notably emphasized the greater accentual density in emphatic speech. A few isolated studies (e.g. [8]) have tried to integrate some expressive accentual information in speech synthesis. However, they were led on Japanese, which is a language with more restricted accentual patterns compared to French or English [9].

Besides the omission of local phenomena, very few attention has been paid to macro prosodic changes. Indeed, most current studies rely on acted corpora of each emotion. These corpora are very constant regarding expressivity, being stereotypical with respect to the considered emotion. However, various expressivity types follow each other in real human speech. As stated by [10], "a coherent speech corpus includes prosodic effects that go beyond the sentence level". These global prosodic changes should be modeled to improve the quality of the generated expressive speech (see [11]).

The generation of an expressive prosodic realization is of utmost importance when synthesizing sports commentaries. Several studies have focused on their prosodic analysis (i.e. for basketball, football and rugby [12], horse races [13], soccer [14] and football [15] [16]). One of their main findings lies in the fact that such speech databases are importantly characterized by variations at the local but also at the global or macro level [15] [16]. At a macro level, [15] proposes to divide the corpus into three main speaking styles. *Elaboration* corresponds to relatively neutral speech. Conversely, dramatic style is more related to a high arousal level and can be subdivided into the *building up of a suspense*, which relates to a rise in the arousal level, and *the presentation of a highlight*, i.e. the arousal climax. These various speaking styles were shown to display specific prosodic features. It was pointed out in [12], for example, that highly excited phases, like shots, tend to be realized with a significantly higher fundamental frequency. This phenomenon was also observed in horse races at the end of the race, when the excitation reaches a maximum [13]. Interestingly, [16] emphasized the fact that, besides the arousal degree, the valence of the expressivity may also influence the prosodic realization of the commentaries. The analysis of sequences happening just after a goal in football games indicates, indeed, that the prosodic realization depends upon whether the goal is for or against the supported team [15]. This could be explained by the fact that sports commentaries are deeply 'listener-oriented' and that this acoustic distinction helps the listener decode the action more quickly. On the whole, most studies tend to suggest that a prosody annotation of sports commentaries requires, besides local accentual information, a more global annotation level assigning a specific speaking style to the speech segments.

This paper is in the continuity with our previous work on the subject [17], in which a prosody annotation protocol specific to sports commentaries (basketball in particular) and relying on two annotation levels was developed. A local annotation is associated to the syllable level and aims at annotating accentual events. A global annotation classifies groups of words into specific speaking styles. The interested reader is referred to [17] for more details. This annotation protocol was developed with HMM-based speech synthesis in view, which is implemented in this work.

One way to perform HMM-based speech synthesis is to train a model, called *full data model*, using a database containing specific data (e.g. data corresponding to a particular emotion, degree of articulation of speech, etc.). Another way to build the models is to make use of adaptation techniques, which allow changing the voice characteristics and prosodic features of a source speaker into those of a target speaker [18]. These latter adapt the source HMM-based model with a limited amount of target speech data. The resulting model is called *adapted model*. The same concept holds for speaking style adaptation [19] [20]. This technique allows providing high quality speech synthesis using a limited amount of adaptation data [21].

Recently, Zen [22] proposed a new framework for estimating HMMs on data containing both multiple speakers and multiple languages. Speaker and language factorization attempts to factorize specific characteristics in the data and then models them using separate transforms. Another study [23] described a discrete/continuous HMM for modeling the symbolic and acoustic speech characteristics of speaking styles.

In this paper, we precisely aim at integrating efficiently the local and global annotations into an HMM-based speech synthesizer. The goal of this study is two-fold: i) quantifying the possible improvements brought by each annotation layer on various aspects of speech synthesis; ii) comparing different training methods regarding the integration of the global labels.

The paper is structured as follows. Section 2 presents the corpus used throughout this study. Section 3 summarizes the proposed annotation protocol and provides a brief overview of the acoustic analysis of both annotation levels (global and local). The integration of the proposed annotation protocol within HMM-based speech synthesis is investigated in Section 4 where some experiments are carried out in order to evaluate the quality of the generated speech across various aspects. Finally, Section 5 concludes the paper.

2. Database

This study is based on a corpus of live commentaries of two basketball games, uttered by a professional French commentator and recorded in sound-proof conditions. The speaker watched the game and commented it without any prompt. The speech signal was recorded with an AKG C3000B microphone. The audio acquisition system Motu 8pre was used outside the sound-proof room, with a sampling rate of 44.1 kHz. The issue with sports commentaries corpora is usually the high level of background noise which precludes their precise acoustic analysis [15]. Conversely, our corpus exhibits the advantage of being spontaneous and of high acoustic quality, being therefore suited for speech synthesis. Both matches star the Spirou Belgian team with very tight final scores, which induces a high level of excitation. The corpus lasts 162 minutes, silences included.

The corpus was orthographically transcribed and the phonetization was automatically produced by [24], with manual check. The phonetic transcription was automatically aligned with the sound with [25], taking advantage of the bootstrap option to reach alignment rates higher than 80% with a 20 ms tolerance threshold. The Elite NLP system [26] produced other required annotation tiers (e.g. syllables, parts of speech, rhythmic groups, etc.). Sentence boundaries form another important annotation level. Such a segmentation of a spontaneous speech corpus is rather complex as we do not have access to punctuation. The corpus was therefore manually annotated to define segments with both a prosodic and a semantic completeness.

3. Prosody Annotation Protocol

We defined in [17] a two-level prosody annotation framework to consider both accentual and macro prosodic phenomena. The objective was to determine a specific set of labels for both annotation tiers. To allow for their integration in a speech synthesis system, the labels required to comply with two main constraints. First, they had to be related to a specific function to facilitate their prediction from text at synthesis stage. Secondly, they had to be characterized by distinct acoustic realizations. It should be noted that existing systems like ToBI [27] could hardly be exploited as such for our study as their complexity makes it difficult to predict the labels from the text.

Our local tier contains a small amount of labels (Table 1). Each label fulfills a distinct and specific function. Five labels are related to non-emphatic stresses [28] and are assigned to the end of accentual phrases. They are characterized by a pitch level, H for rising or higher pitch vs. L for falling or lower pitch. They can also be distinguished by the level of boundary they determine, similarly to boundary tones in [27]. To facilitate the automatic annotation of the labels (from the text or from the acoustics), these two levels are distinguished according to the presence or absence of a subsequent silence. Conversely to H and L syllables, HH and LL syllables are directly followed by a silent pause. A specific tag E is assigned to the final boundary of player names enumerations, which are very common in sports commentaries and may display a specific acoustic realization. A focus stress (F) relates to emphatic stresses. An hesitation label (He), and a creaky label (C) allow avoiding the degradation of the models at training time. Indeed, hesitations are realized with long durations whilst creaky syllables are characterized, among others, by a very low pitch [29]. If these syllables are not singled out, their prosodic features may influence the synthesized prosody. All remaining syllables are assigned a NA symbol.

Table 1: *List of local labels.*

Stresses		Unstressed	Other
Not emphatic	Emphatic		
H HH L LL E	F	NA	He C

The global tier is inspired by [15] and [16] and is assigned to groups of words, conversely to the local annotation which assigns a symbol to each syllable. It basically classifies the speech segments into speaking styles, based on a dimensional analysis of emotions [30]. The valence and the arousal levels drove us to define five speaking styles (Figure 1).

We showed in [17] that the different labels are, as required, associated with specific acoustic realizations. While 'F' labels tend to be realized with a higher pitch level but low syllable lengthening, non-emphatic stresses are usually characterized by an important lengthening of the syllable. Regarding global labels, the arousal level was shown to be correlated with the fundamental frequency, highly excited segments being realized with a significantly higher pitch. Inter-annotator rates were also computed. They reached a Cohen's kappa score [31] of 0.66 for the local labels, which is comparable to the rate obtained for ToBI [27]. The global annotation achieved lower rates but with logical interchanges between the labels [17].

4. Methodology and Experiments

In order to assess the validity of our local and global labels definition, several HMM-based speech synthesizers [5] were built,

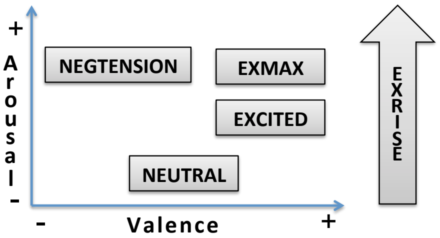

Figure 1: *The global labels on a dimensional scale.*

relying on the implementation of the HTS toolkit[1] (version 2.1) publicly available. For each synthesizer, 90% of the corresponding database was used for the training (called the *training set*), leaving around 10% for the synthesis (called the *synthesis set*). Note that both the training and synthesis sets were manually annotated with our two-layer prosody annotation. As filter parameterization, we extracted the Mel Generalized Cepstral (MGC) coefficients traditionally used in parametric synthesis. As excitation modeling, the Deterministic plus Stochastic Model (DSM [32]) of the residual signal was used to improve naturalness.

The influence of the integration of the local and global labels is first analyzed and quantified independently in Sections 4.1 and 4.2, respectively. Using the conclusions drawn from these latter evaluations, the combination of both local and global labels is studied in Section 4.3.

4.1. Integration of the Local Labels

This section is devoted to the integration of the single local annotation layer into HMM-based speech synthesis.

4.1.1. Method

The first synthesizer is trained on the entire training set of the database (Section 2), regardless of the speaking styles. This is our baseline system, called *Base*. The only contextual information provided during the training and synthesis stages is a manually-checked phonetic transcription, embedded as standard HTS labels [5].

The same training procedure is applied to the second synthesizer, called *Loc*. It makes use of the same phonetic transcription, but complemented in this case with specific contextual information from the local prosody annotation level (Table 1), replacing the unused ToBI field in the standard HTS labels.

4.1.2. Evaluation Protocol

A first Mean Opinion Score (MOS) test is conducted in order to quantify the impact of the local annotation layer in comparison with the baseline system. For this evaluation, participants were asked to listen to two versions of the same sentence synthesized by the two following models (randomly shuffled): (i) the baseline system (*Base*); (ii) the system integrating local labels (*Loc*). Each sentence was scored according to three criteria: intelligibility, expressivity and segmental quality. Listeners were given two continuous MOS scales (one for each criterion) ranging from 1 (meaning "poor") to 5 (meaning "excellent"). These scales were extended one point further on both sides (ranging therefore from 0 to 6) in order to prevent border effects.

The test consists of 10 pairwise comparisons. Sentences were randomly chosen amongst the synthesis set of the database. These sentences were not divided in speaking styles, which means that a sentence may correspond to a sequence of various speaking styles. 10 native French-speaking people, mainly naive listeners, participated in this evaluation. During the test, they could listen to the pair of sentences as many times as wanted in the order they preferred. They were nonetheless advised to first listen to the two sentences in a row so as to estimate approximately their relative position. However, they were not allowed to come back to previous sentences after validating their decisions.

4.1.3. Results

MOS scores are computed for all evaluations in this paper to provide comparable results in a coherent evaluation framework. The actual MOS scores are, however, less informative in this first evaluation. Therefore, our analysis relies on the preference percentages which are computed as the listener's relative preference for a synthesis method compared to another. Figure 2 shows the preference scores for the three criteria. The light grey segment corresponds to the proportion of cases in which both methods are assigned the same MOS score. It can be observed that *Loc* is preferred for the rendering of the expressivity. Interestingly, it is also shown to improve the segmental quality, while achieving an intelligibility level that is similar to the baseline (i.e. *Base*).

The analysis of the MOS scores further confirms that *Loc* allows to slightly increase the expressivity compared to *Base*. This means that local labels were properly learned during training of *Loc* and that specific accentual Probability Density Functions (PDFs) were properly estimated. At synthesis time, the model was thus able to predict more precise accentual realizations. On the contrary, since only a manually-checked phonetic transcription was provided for the *Base* training, all acoustic realizations that should have corresponded to local labels were merged into more global PDFs.

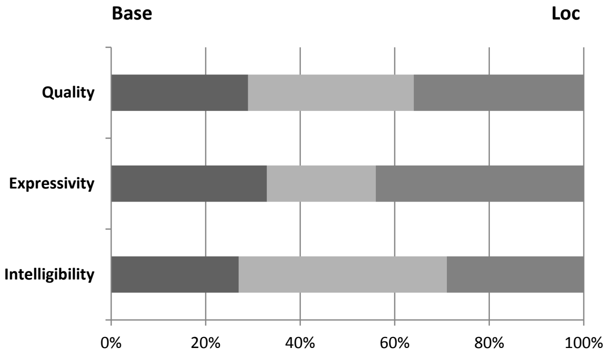

Figure 2: *Preference scores for each criterion and each synthesis method, i.e. with (Loc) and without (Base) the integration of the local labels.*

4.2. Integration of the Global Labels

This section studies the integration of the single global annotation layer into HMM-based speech synthesis. It should be noted that speaking styles defined by global annotation layer are not uniformly distributed throughout the corpus. The total duration of each speaking style of the database is shown in Table 2.

[1] http://hts.sp.nitech.ac.jp/

Table 2: *Total duration (in sec.) of the various speaking styles of our database, long silences (>1 sec.) being excluded.*

Excited	ExMax	ExRise	NegTension	Neutral
1032	475	485	582	2955

4.2.1. Method

Three distinct methods are investigated regarding the integration of the global labels in speech synthesis. The first method consists in training speaking style-dependent models on exclusive subsets of the whole corpus, specific to the global label they correspond to (see Table 2). They will be referred to as *full data models* in the remainder of the paper. At the end of this step, 5 different full data models, called *Glob1*, are obtained (i.e. Excited, ExMax, ExRise, NegTension and Neutral).

The two other methods exploit adaptation techniques. The second method relies on the fact that the Neutral style has the highest amount of speech data amongst the different speaking styles. Assuming that this amount of speech data is sufficient to obtain a strong Neutral *full data model*, voice adaptation techniques [18] can be applied to train more reliably the remaining models, for which less speech data are available. The *Glob1* Neutral *full data model* was then adapted using the Constrained Maximum Likelihood Linear Regression (CMLLR) transform [33] [34] in the framework of Hidden Semi Markov Model (HSMM) [35] with the adaptation sets of the four remaining speaking styles. It produces respectively Excited, ExMax, ExRise and NegTension HMM-based synthesizers. The linearly-transformed models were further optimized using MAP adaptation [18], providing the 4 adapted models called *Glob2*.

A potential drawback of the second method is that the speech data used to train the Neutral *full data model* may not be large enough. To alleviate this issue, Yamagishi proposed in [36] to adapt a so-called *average-voice model* to a particular target speaker. The average-voice model is computed once and for all over a database containing many different speakers. This technique proved to be efficient when few speech data is available. The average-voice model was here computed on the entire database, regardless of the speaking styles. This model was then adapted following the same procedure as for *Glob2*. We finally obtained 5 different adapted sub-synthesizers called *Glob3* (i.e. Excited, ExMax, ExRise, NegTension and Neutral).

4.2.2. Evaluation Protocol

A second MOS test is conducted in order to elect which of the three methods is the most suited for the integration of global labels. For this evaluation, participants were asked to listen to three versions of the same sentence, synthesized by the models corresponding to the three aforementioned methods. Each sentence was scored according to two criteria: intelligibility and segmental quality. The same experimental protocol as in Section 4.1.2 was applied. However, contrarily to that section, the expressivity was not assessed here. Indeed our first informal experiments showed that *Glob* models exhibit some intelligibility and segmental quality issues. These had to be addressed before focusing on a good rendering of the expressivity.

The test consists of 15 triplets. Sentences were randomly chosen amongst the synthesis set of the database. Conversely to Section 4.1.2, each sentence only contains one speaking style. 12 native French-speaking testers, mainly naive listeners, participated in this evaluation.

4.2.3. Results

For each synthesis technique and each speaking style, Figures 3 and 4 display respectively the averaged intelligibility and segmental quality MOS scores, together with their 95% confidence intervals (CI). It clearly turns out that *Glob3*, i.e. the adapted average models, provides the highest results, both in terms of intelligibility and segmental quality of the generated speech. The *full data models*, conversely, achieve the lowest scores in most speaking styles. As a reminder, a score of 3 or 4 on the MOS scale means respectively "Fair" or "Good".

This preference for the adapted average models can be explained by the fact that they are computed using all the training sets for each speaking style, thus providing a robust model which is then adapted to each speaking style. It can however be noted that Neutral, Excited and NegTension voices are better rendered than ExMax and ExRise ones. This is mainly due to the fact that Neutral, Excited and NegTension have more speech data, leading to a better average-voice model adaptation compared to ExMax and ExRise.

It should also be noted that all the synthesizers achieved the same performance for the Neutral speaking style. This can be understood by the fact that this latter style is the only one having a comfortable amount of speech data for a reliable estimation of the model, independently of the training method.

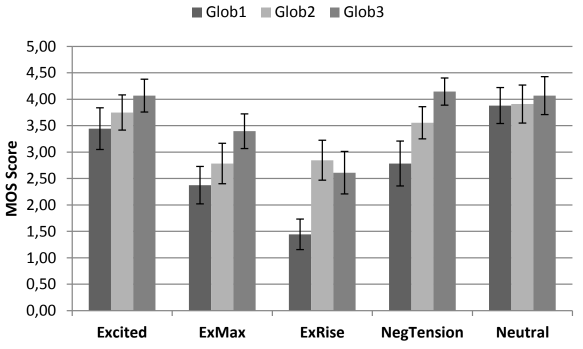

Figure 3: *Averaged* ***intelligibility*** *MOS scores for each synthesis method and each speaking style, with their 95% CI.*

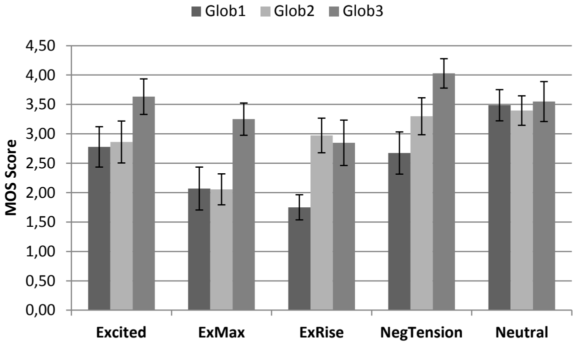

Figure 4: *Averaged* ***segmental quality*** *MOS scores for each synthesis method and each speaking style, with their 95% CI.*

Preference scores corroborate these findings. Regarding intelligibility, *Glob3* is respectively preferred in 69.4% and 51.1% of the cases compared to *Glob1* and *Glob2* (which are assigned 13.3% and 25.6% of the preferences). *Glob3* is chosen in 73.9%

and 64.4% of the cases over respectively *Glob1* and *Glob2* for its segmental quality.

This shows that, as the amount of speech data is unevenly distributed amongst the different speaking styles, adapting a robustly trained average-voice model with an efficient technique such as CMLLR allows generating various speaking styles of reasonable intelligibility and segmental quality.

4.2.4. Comparison with the Baseline

Similarly to the integration of local labels (Section 4.1.3), the integration of global labels was compared to the baseline through a MOS test. For this comparison, the best integration technique, i.e. the adaptation of the average model, was used. The baseline is the same as in Section 4.1.3, which means that it disregards both annotation layers. 20 native French-speaking people, mainly naive listeners, participated in this evaluation. Conversely to what was expected, no statistically significant differences were observed between both methods. This integration achieves indeed comparable or even slightly lower scores in terms of intelligibility, expressivity and segmental quality.

4.3. Integration of both Local and Global Labels

This section is devoted to the integration of the two-layer annotation (both local and global labels) into HMM-based synthesis.

4.3.1. Method

We showed in Section 4.1.3 that the integration of local labels results in an enhanced expressivity, while it provides slightly higher intelligibility and segmental quality performance. Regarding the integration of global labels, the use of adaptation techniques, from an average model, was shown to provide the best results (see Section 4.2.3). However, this annotation level seemed to achieve no improvement regarding expressivity in comparison with a baseline model.

We investigate, in this last section, whether the integration of both local and global labels achieves higher scores. These combined models are referred to as *Loc+Glob3*.

4.3.2. Evaluation Protocol

A third MOS test is conducted. For this evaluation, participants were asked to listen to four versions of the same sentence synthesized by the following models (randomly shuffled): (i) the baseline model (*Base*); (ii) the model integrating local labels (*Loc*); (iii) the model adapted from the average-voice model (*Glob3*); (iv) the model adapted from the average-voice model integrating local labels (*Loc+Glob3*). Here again, intelligibility, expressivity and segmental quality are evaluated. The same experimental protocol as in Section 4.1.2 was applied.

The test consists of 10 quadruplets. Sentences were randomly chosen amongst the synthesis sets of each speaking style of the database. Similarly to Section 4.2.2, each sentence only contains one speaking style. 20 native French-speaking people, mainly naive listeners, participated in this evaluation.

4.3.3. Results

Table 3 shows the preference scores for the four methods. It should be noted that the scores obtained by two reverse pairs are not summing to 100%. This is due to the fact that cases when both methods composing the considered pair are found to be equivalent are also taken into account. Regarding intelligibility, Table 3 shows for example that *Base* is preferred to *Loc* in 31% of the cases, while *Loc* is preferred to *Base* in 35.5% of the cases. The remaining percentage, i.e. 33.5%, corresponds then to the cases where both *Base* and *Loc* are equivalently preferred.

As in Section 4.1.3, it is observed that the integration of the local labels carries out an improvement in the rendering of the expressivity and provides comparable or slightly better intelligibility and segmental quality of the generated speech. On the other hand, the integration of the global labels only (using the adapted average models) does not improve any of the analyzed criteria compared to the baseline, which corroborates the results obtained in Section 4.2.4. Regarding the integration of both prosody annotation levels, an insightful observation is that *Loc+Glob3* is preferred in 41% of the cases in terms of expressivity against *Loc*, which is assigned 35% of the preferences. The segmental quality degrades, however, from *Loc* to *Loc+Glob3* as they are respectively preferred in 39.5% and 31.5% of the time. Nonetheless, both methods achieve similar intelligibility performance.

Table 3: *Integration of both Global and Local Labels - Preference scores (in [%]) for each method and each criterion.*

		Base	*Loc*	*Glob3*	*Loc+* *Glob3*
Intelligibility	*Base*	0	31	33	29.5
	Loc	35.5	0	37.5	29.5
	Glob3	29	29	0	26.5
	Loc+Glob3	40	27.5	38	0
Expressivity	*Base*	0	37.5	40	32.5
	Loc	45.5	0	46	35
	Glob3	35.5	35	0	29.5
	Loc+Glob3	51	41	47.5	0
Quality	*Base*	0	39.5	40	44
	Loc	38.5	0	46.5	39.5
	Glob3	27.5	33.5	0	35
	Loc+Glob3	36	31.5	45.5	0

5. Conclusion

In this paper, we proposed the integration of a two-layer prosody annotation specific to live sports commentaries into HMM-based speech synthesis. The local annotation relates to accentual phenomena while the global layer classifies the speech segments into distinct speaking styles. Our study was divided into three parts.

First, the improvement carried out by local labels was quantified by comparing: (i) a baseline model, in which a manually-checked phonetic transcription was the only contextual information provided and (ii) a model integrating local labels. Subjective tests revealed that, compared to the baseline, the integration of local labels results in an enhanced expressivity, while providing slightly higher intelligibility and segmental quality scores.

Secondly, the integration of global labels (i.e. speaking styles) was evaluated. Three methods were investigated: (i) a speaking style-dependent training and the adaptation of (ii) the neutral model or (iii) the average-voice model to each speaking style. It was shown that adaptation techniques, and the adaptation from an average-voice model in particular, outperform style-dependent models both in terms of intelligibility and segmental quality. However, the comparison with the baseline, i.e. the model disregarding global labels, showed that, contrary to what was expected, the integration of global labels does not enhance expressivity and slightly degrades the segmental quality.

A last experiment allowed evaluating the effects achieved by the combination of both prosody annotation layers on the generated speech. Interestingly, the complete integration of the two-layer annotation, compared to the model integrating local labels only, led to an even better rendering of expressivity, while achieving similar intelligibility scores. However, it slightly degrades the segmental quality. Our future work should thus focus on the improvement of speaking style adaptation techniques in order to increase the segmental quality of the generated speech.

Audio examples related to this study are available online at http://tcts.fpms.ac.be/~picart/.

6. Ackowledgements

Authors are supported by FNRS. The project is partly funded by the Walloon Region Wist 3 SPORTIC. Authors are grateful to S. Audrit for her implication in the recording of the corpus.

7. References

[1] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, “Constructing emotional speech synthesizers with limited speech database,” in *International Conference on Spoken Language Processing (ICSLP)*, 2004, pp. 1185–1188.

[2] J. Yamagishi, K. Onishi, T. Musuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis,” *IECE Transactions on Information and Systems*, vol. E88-D(3), pp. 502–509, 2005.

[3] T. Takahashi, T. Fujii, M. Nishi, H. Banno, T. Irino, and H. Kawahara, “Voice and emotional expression transformation based on statistics of vowel parameters in an emotional speech database,” in *Interspeech*, 2005, pp. 537–540.

[4] L. Qin, Z.-H. Ling, Y.-J. Wu, B.-F. Zhang, and R.-H. Wang, “HMM-based emotional speech synthesis using average emotion models,” in *ICSLP*, 2006, pp. 233–240.

[5] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Comm.*, vol. 51(11), pp. 1039–1064, 2009.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Eurospeech*, 1999, pp. 2347–2350.

[7] I. Fonagy, *L’accent en français contemporain*. Ottawa: Marcel Didier Ltée, 1979, ch. L’accent français : Accent probabilitaire, pp. 123–232.

[8] K. Hirose, K. Sato, and N. Minematsu, “Emotional speech synthesis with corpus-based generation of f0 contours using generation process model,” in *Speech Prosody*, 2004, pp. 417–420.

[9] M. E. Beckman and J. B. Pierrehumbert, “Japanese prosodic phrasing and intonation synthesis,” in *Twenty-Fourth Annual Meeting of ACL*, 1986, p. 173180.

[10] N. Braunschweiler, M. J. Gales, and S. Buchholz, “Lightly supervised recognition for automatic alignment of large coherent speech recordings,” in *Interspeech*, 2010, pp. 2222–2225.

[11] F. Eyben, S. Bucholz, and N. Braunschweiler, “Unsupervised clustering of emotion and voice styles for expressive TTS,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[12] S. Audrit, T. Psir, A. Auchlin, and J.-P. Goldman, “Sport in the media: A contrasted study of three sport live media reports with semi-automatic tools,” in *Speech Prosody*, 2012.

[13] J. Trouvain and W. Barry, “The prosody of excitement in horse race commentaries,” in *ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, 2000, pp. 86–91.

[14] N. Obin, V. Dellwo, A. Lacheret, and X. Rodet, “Expectations for discourse genre identification,” in *Interspeech*, 2010.

[15] J. Trouvain, “Between excitement and triumph - live football commentaries in radio vs. tv,” in *17th International Congress of Phonetic Sciences (ICPhS XVII)*, 2011.

[16] F. Kern, *Prosody in Interaction*. John Benjamins, 2010, ch. Speaking Dramatically. The Prosody of Live Radio Commentary of Football Matches, pp. 217–237.

[17] S. Brognaux, B. Picart, and T. Drugman, “A new prosody annotation protocol for live sports commentaries,” in *Interspeech*, 2013.

[18] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “Robust speaker-adaptive hmm-based text-to-speech synthesis,” *IEEE Audio, Speech, & Language Processing*, vol. 17(6), pp. 1208–1230, 2009.

[19] J. Yamagishi, T. Masuko, and T. Kobayashi, “Hmm-based expressive speech synthesis - towards tts with arbitrary speaking styles and emotions,” in *Proc. of SWIM*, 2004.

[20] T. Nose, M. Tachibana, and T. Kobayash, “Hmm-based style control for expressive speech synthesis with arbitrary speakers voice using model adaptation,” *IEICE Transactions on Information and Systems*, vol. 92(3), pp. 489–497, 2009.

[21] J. Yamagishi, “Average-voice-based speech synthesis,” Ph.D. dissertation, Tokyo Institute of Technology, 2006.

[22] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, “Statistical parametric speech synthesis based on speaker and language factorization,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20(6), pp. 1713–1724, 2012.

[23] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, “Discrete/continuous modelling of speaking style in hmm-based speech synthesis: Design and evaluation,” in *Interspeech*, 2011.

[24] J.-P. Goldman, “Easyalign: an automatic phonetic alignment tool under Praat,” in *Interspeech*, 2011, pp. 3233–3236.

[25] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, “Train&Align: A new online tool for automatic phonetic alignments,” in *IEEE SLT Workshop*, 2012.

[26] V. Colotte and R. Beaufort, “Linguistic features weighting for a text-to-speech system without prosody model,” in *Interspeech*, 2005, pp. 2549–2552.

[27] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “ToBI: A standard for labeling english prosody,” in *ICSLP*, 1992, pp. 867–870.

[28] A. Di Cristo, “Vers une modélisation de l’accentuation du français: deuxième partie,” *Journal of French Studies*, vol. 10, pp. 27–44, 2000.

[29] T. Drugman, J. Kane, and C. Gobl, “Modeling the creaky excitation for parametric speech synthesis,” in *Interspeech*, 2012.

[30] A. Mehrabian and J. A. Russel, *An Approach to Environmental Psychology*. MIT Press, 1974.

[31] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20(1), pp. 37–46, 1960.

[32] T. Drugman and T. Dutoit, “The deterministic plus stochastic model of the residual signal and its applications,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20(3), pp. 968–981, 2012.

[33] V. Digalakis, D. Rtischev, and L. Neumeyer, “Speaker adaptation using constrainted reestimation of gaussian mixtures,” *IEEE Transactions on Speech and Audio Processing*, vol. 3(5), pp. 357–366, 1995.

[34] M. Gales, “Maximum likelihood linear transformations for hmm-base speech recognition,” *Computer Speech and Language*, vol. 12(2), pp. 75–98, 1998.

[35] J. Ferguson, “Variable duration models for speech,” in *Symp. on Application of Hidden Markov Models to Text and Speech*, 1980.

[36] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training,” *IEICE Transactions Information and Systems*, vol. 90(2), pp. 533–543, 2007.