Synthèse de commentaires sportifs: intégration d'une annotation prosodique à deux niveaux dans un synthétiseur HMM

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

[1]Cental, ICTEAM (Université catholique de Louvain), Belgique
[2]TCTS Lab (Université de Mons), Belgique
<sandrine.brognaux@uclouvain.be, benjamin.picart@umons.ac.be, thomas.drugman@umons.ac.be>

Abstract

This paper proposes a new prosody annotation protocol specific to live sports commentaries. Two levels of annotation are defined with HMM-based speech synthesis in view. Local labels are assigned to all syllables and refer to accentual phenomena. Global labels classify sequences of words into five distinct sub-genres, defined in terms of valence and arousal. Our analysis shows that the labels are both related to a specific function and characterized by a distinct acoustic realization. The consideration of these constraints should allow for an automatic prediction of the labels both from the text or from the speech signal. The integration of this new annotation protocol within HMM-based speech synthesis shows promising results.

Mots-clés : *prosodie, annotation, commentaires sportifs, synthèse vocale, HMM*

1. Introduction

La synthèse de la parole fait partie intégrante de notre vie de tous les jours et s'impose dans de nombreuses applications (GPS, jeux, cartes électroniques, etc.). Cependant, le fossé qui demeure entre voix synthétique et voix humaine empêche encore sa commercialisation à plus grande échelle. Ce manque de naturel est notamment dû à l'incapacité des systèmes actuels à gérer l'expressivité, qui n'est pas ou peu modélisée (Campbell, 2006). La génération d'une prosodie expressive est d'une importance cruciale lors de la synthèse de commentaires sportifs. Leur niveau élevé d'expressivité remplit en effet de nombreuses fonctions (exprimer l'excitation, la frustration, attirer l'attention, etc.) (Trouvain, 2011 ; Kern, 2010) qui doivent être reproduites dans la voix de synthèse.

Plusieurs études se sont concentrées sur l'analyse prosodique des commentaires sportifs (basketball, football et rugby (Audrit, Pršir, Auchlin, & Goldman, 2012), courses de chevaux (Trouvain & Barry, 2000), football (Obin, Dellwo, Lacheret, & Rodet, 2010) et football américain

(Trouvain, 2011 ; Kern, 2010). Ce phonostyle se distingue des autres situations de communication par une réalisation prosodique très spécifique (Obin et al., 2010). Il se caractérise par des schémas accentuels expressifs et la plupart des études soulignent la distinction possible entre différents styles de parole au sein du commentaire sportif (Kern, 2010) (voir Figure 1). Ces styles de paroles seront dénommés « sous-genres » dans la suite de cet article et désignent ce que d'autres études appellent des « prosodies » (Odgen, 2001) ou « ambiances de discours » (Goldman, 2012). Tandis que *l'élaboration* correspond à un style de parole relativement neutre, la *parole dramatique* est caractérisée par un haut degré d'excitation, qui augmente durant la phase de *construction du suspense* et atteint son paroxysme lors de la *présentation du temps fort*. Ces sous-genres sont caractérisés par une fonction spécifique et une réalisation acoustique relativement stable. Audrit et al. (2012) montrent, par exemple, que les moments d'excitation intense sont réalisés avec une fréquence fondamentale (F0) significativement plus élevée. L'analyse de commentaires de courses de chevaux (Trouvain & Barry, 2000) offre des résultats similaires pour la frénésie des phases de fin de course. L'analyse de séquences survenant juste après un goal dans les commentaires de football indique également que la réalisation acoustique dépend du fait que le goal soit ou non pour l'équipe supportée par le commentateur (Trouvain, 2011). Ces deux réalisations acoustiques distinctes permettent ainsi à l'auditeur de décoder le message plus rapidement. Contrairement à la classification de Kern (2010), la distinction repose ici sur la valence et non sur le degré d'excitation. Globalement, la majorité des études tendent à suggérer que l'annotation des commentaires sportifs nécessite, en plus de l'information accentuelle locale, un niveau d'annotation plus global attribuant un sous-genre spécifique aux différents segments de parole.

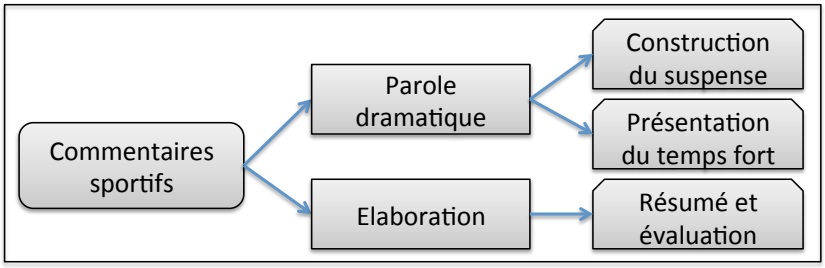

Figure 1 : *Styles de parole dans les commentaires sportifs (Kern, 2010)*

Différents modèles d'accentuation ont été développés pour la synthèse vocale et pourraient être exploités afin de définir une annotation locale de la parole expressive. ToBI (Silverman et al., 1992) propose un protocole d'annotation de la prosodie détaillé. Il faut cependant noter

que la complexité du système rend difficile la prédiction automatique des labels à partir du texte. Mertens (1987) a également proposé un modèle spécifique au français qui permet de discrétiser le continuum prosodique en niveaux de hauteurs, attribués aux syllabes. Ce système présente cependant des limites similaires à ToBI. En effet, plusieurs tons différents peuvent être associés à une même fonction, ce qui complique leur prédiction à partir du texte. En ce qui concerne l'annotation globale en sous-genres, elle pourrait se baser sur Kern (2010) et Trouvain (2011).

L'objectif de cet article est de présenter un protocole d'annotation prosodique spécifique aux commentaires sportifs (et plus particulièrement au basketball) sur base de deux niveaux d'annotation. Une annotation locale est associée au niveau syllabique et concerne les phénomènes accentuels. Une annotation globale classifie les groupes de mots en différents sous-genres. Ce protocole d'annotation a été développé en vue de l'intégrer dans la synthèse vocale par modèles de Markov cachés (HMMs). Les labels locaux sont inclus dans l'information contextuelle fournie au système tandis que l'annotation globale est exploitée afin d'entraîner des modèles distincts pour chaque sous-genre. L'utilisation de cette annotation pour la synthèse vocale implique quelques contraintes. Les labels locaux et globaux doivent être associés à une fonction expressive particulière. Si l'on suppose qu'une analyse sémantique du texte est disponible, il sera alors aisé de prédire les labels à partir du texte. D'autre part, afin d'éviter de moyenner les différents effets acoustiques, chaque label doit également être associé à des caractéristiques acoustiques spécifiques.

L'article s'organise comme suit. La Section 2 présente le corpus de commentaires sportifs utilisé dans le cadre de cette étude. La Section 3 propose une description détaillée de notre protocole d'annotation prosodique. Une analyse acoustique des deux niveaux d'annotation (local et gobal) est présentée en Section 4 et leur intégration en synthèse HMM est évaluée en Section 5. Enfin, la Section 6 conclut l'article.

2. Corpus de commentaires sportifs

Cette étude repose sur un corpus de commentaires de deux matchs de basketball, prononcés par un commentateur professionnel francophone et enregistrés en chambre anéchoïque. Il a été demandé au commentateur de regarder le match et de le commenter, sans script. Les deux matchs mettent en scène l'équipe belge des Spirou et se terminent sur des scores très serrés, ce qui engendre un niveau d'excitation élevé. Le corpus a une durée totale de 162 minutes, silences compris. Le problème principal des corpus de commentaires sportifs est généralement le niveau important de bruit de fond qui ne permet pas une analyse précise des caractéristiques acoustiques (Trouvain, 2011). Au contraire,

notre corpus offre l'avantage d'être spontané et de bonne qualité acoustique, rendant ainsi possible son utilisation en synthèse vocale.

Le corpus a été transcrit orthographiquement et phonétiquement par EasyAlign (Goldman, 2011), avec vérification manuelle. La transcription phonétique a été alignée avec le son avec Train & Align (Brognaux, Roekhaut, Drugman, & Beaufort, 2012), la fonction de bootstrap permettant d'atteindre des taux d'alignement proches de 85 % à 20 ms près. Enfin, d'autres informations linguistiques (syllabes, catégories grammaticales, groupes rythmiques, etc.) ont été générés par eLite (Beaufort & Ruelle, 2006).

3. Annotation de la prosodie

3.1. Protocole d'annotation prosodique

Une annotation prosodique à deux niveaux est proposée. Une couche locale, associée au niveau syllabique, indique les phénomènes accentuels. Elle comprend un ensemble d'étiquettes qui peuvent être prédites à partir du texte (voir Tableau 1). Chaque étiquette remplit une fonction distincte et spécifique. Cinq étiquettes correspondent à des accents non-emphatiques (Di Cristo, 2000) et sont attribuées à des syllabes en fin de groupe intonatif. Elles sont caractérisées par un niveau de hauteur, H pour montant ou haut (continuatif) vs. L pour descendant ou bas (conclusif). Elles se distinguent également par le niveau de frontière qu'elles déterminent, similairement au système d'annotation ToBI (Silverman et al., 1992). Afin de faciliter l'annotation automatique des labels (à partir du texte ou de l'acoustique), ces deux niveaux se distinguent sur base de l'absence ou la présence d'un silence suivant la syllabe. Contrairement aux syllabes L et H, les syllabes LL et HH sont directement suivies par une pause silencieuse. Un label spécifique E est associé à la syllabe finale d'un nom de joueur dans une énumération, les énumérations étant particulièrement fréquentes dans les commentaires sportifs et étant potentiellement caractérisées par une réalisation acoustique particulière. Un accent de focus (F) est assigné aux syllabes marquées par un accent emphatique. Notons qu'une analyse approfondie des accents emphatiques de notre corpus a montré qu'il n'était pas utile de distinguer plusieurs types d'accents 'F' (Brognaux, Drugman, & Saerens, 2014). Des labels d'hésitation (He) et de voix rauque (C) permettent d'éviter la dégradation des modèles à l'entraînement. En effet, les hésitations sont caractérisées par un allongement de la syllabe tandis que les syllabes rauques sont réalisées par un niveau de pitch particulièrement bas (Drugman, Kane, & Goble, 2012). Si ces syllabes ne sont pas exclues, leurs caractéristiques prosodiques pourraient influencer la prosodie synthétisée. Toutes les syllabes restantes reçoivent une étiquette NA (non accentué).

Accents		Non accentué	Autre
Non emphatique	Emphatique		
H HH L LL E	F	NA	He C

Tableau 1 : *Liste des étiquettes prosodiques locales*

La couche d'annotation globale, inspirée par (Kern, 2010) et (Trouvain, 2011), assigne une étiquette par groupe de mots, contrairement aux étiquettes locales qui sont attribuées à chaque syllabe. Ces labels globaux classifient les segments de parole en sous-genres, sur base d'une analyse dimensionnelle des émotions (Mehrabian & Russel, 1974). La valence et le niveau d'excitation permettent de définir cinq sous-genres (voir Figure 2). Tandis que les labels 'Excité' et 'Excitation max' correspondent à une excitation positive, 'Tension négative' est lié à un sentiment de frustration, notamment lors d'un tir manqué.

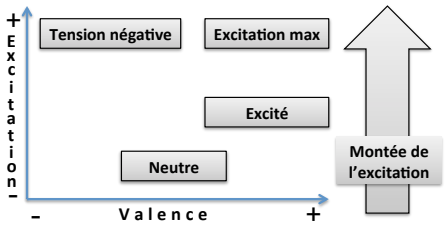

Figure 2 : *Etiquettes prosodiques globales sur une échelle dimensionnelle*

Le corpus a été annoté manuellement à l'aide de cette annotation à deux niveaux. Seule une étiquette locale a été assignée à chaque syllabe, éventuellement associée à He ou C. Si un accent emphatique tombe sur une frontière finale de groupe intonatif, le label 'F' l'emporte sur les accents non emphatiques. L'annotation résulte de deux ou trois écoutes de chaque séquence de 4-5 mots et a été soumise à une seconde vérification par le même annotateur. Cette annotation a été réalisée dans le logiciel Praat (Boersma & Weenink, 2009), avec affichage des informations acoustiques (F0, intensité et formants). Il était demandé à l'annotateur de se fier à son impression auditive uniquement, sauf si les indications acoustiques contredisaient ce jugement. En ce qui concerne les labels locaux, ceci a permis d'éviter l'annotation d'accents qui ne présentaient pas la réalisation acoustique spécifique à ce type d'accent, ce qui dégraderait la qualité de la synthèse. En effet, il a été montré que les annotateurs humains ont tendance à sur-détecter les proéminences sur des syllabes où un accent est généralement attendu, ou à les sous-détecter sur des mots clitiques, tels que les déterminants (Goldman, Auchlin, Roekhaut, Simon, & Avanzi, 2010). Nous avons dès lors tenté d'éviter de tels problèmes, l'objectif de notre protocole d'annotation étant d'offrir une description de la réalisation acoustique et non de proposer une

représentation perceptive de la prosodie.

3.2. Taux d'accord inter-annotateurs

Vingt pourcents du corpus ont été annotés par un second expert, l'annotation initiale étant cachée. L'annotation locale atteint ainsi un score kappa (Cohen, 1960) de 0,66 avec un accord observé de 80,22 %. Ce taux est comparable à l'accord inter-annotateurs atteint par ToBI (Silverman et al., 1992). En ce qui concerne l'annotation globale, un score kappa inférieur est obtenu : 0,38 avec un accord observé de 54,33 %. Tout comme pour l'annotation locale, le score a été calculé au niveau de la syllabe. Il nous faut cependant souligner que la matrice de confusion montre des confusions logiques entre les sous-genres (voir Tableau 2). Sans surprise, *Neutre* et *Excité* (ainsi que *Excitation max* et *Excité*) ont tendance à être confondus, la distinction entre ces labels correspondant à une discrétisation du continuum du niveau d'excitation.

	Neutre	Excité	ExMax	TNeg	ExRise
Neutre	1047	147	0	77	17
Excité	869	499	144	53	28
ExMax	14	20	152	24	0
TNeg	122	68	17	247	3
ExRise	97	124	58	9	305

Tableau 2 : *Matrice de confusion de l'accord inter-annotateurs pour l'annotation prosodique globale du corpus en sous-genres (TNeg = Tension négative, ExMax = Excitation max, ExRise = Montée de l'excitation)*

4. Analyse acoustique des labels prosodiques

Cette section présente une analyse acoustique des labels prosodiques. L'objectif est de déterminer dans quelle mesure ces labels sont caractérisés par des réalisations acoustiques distinctes, qui pourront être ensuite reflétées dans la voix de synthèse.

4.1. Analyse de l'annotation locale

Le Tableau 3 montre les valeurs acoustiques moyennes pour quatre paramètres prosodiques discriminants. Les mesures de durée sont extraites avec Prosogramme (Mertens, 2004). L'intensité perceptive totale (Peeters, 2004) et la fréquence fondamentale (basée sur l'algorithme de SRH (Drugman & Alwan, 2011)) sont également analysées. Les valeurs dynamiques sont définies ici comme la différence entre la valeur du premier et du dernier segment de 10 ms de la syllabe.

Une première observation concerne le fait que la durée du noyau permet d'effectuer la distinction entre les frontières intonatives suivies (LL et HH) ou non (L et H) par un silence. La durée des accents de focus (F) reste quant à elle relativement faible. D'autre part, le pitch moyen pour les syllabes F est plus élevé que pour toutes les autres syl-

	Durée noyau vocalique (sec)	F0 moyenne (Hz)	Dyn F0 (Hz)	Dyn intensité (dB)
H	0,07	223	10,1	-2,2
HH	0,14	222	9,2	-26,2
E	0,11	184	19,5	-16,5
L	0,10	221	-18,1	-2,7
LL	0,12	202	-42,3	-26,0
F	0,07	245	44,5	5,8
He	0,18	185	-11,6	-3,3
C	0,03	170	-6,5	13,6
NA	0,05	227	-2,7	9,2

Tableau 3 : *Valeurs acoustiques moyennes des étiquettes prosodiques locales*

labes. Ceci confirme les résultats de nombreuses études montrant que les accents emphatiques sont davantage réalisés en termes de pitch que d'allongement, contrairement aux accents de fin de groupe intonatif (Lacheret-Dujour & Beaugendre, 1999). Le dynamisme de la F0 permet quant à lui d'effectuer la distinction entre les accents non emphatiques montants/hauts (ou continuatifs) et descendants/bas (ou conclusifs). Les accents emphatiques de notre corpus sont caractérisés par un pitch relativement montant. Enfin, le dynamisme de l'intensité établit une distinction entre les accents emphatiques et non emphatiques, ces derniers étant caractérisés par une diminution de l'intensité due à leur position en fin de groupe intonatif. Tout comme HH, les syllabes E présentent un pitch montant, un allongement du noyau et une diminution de l'intensité. Elles sont cependant réalisées avec un niveau de pitch nettement inférieur.

Une étude plus approfondie des accents emphatiques de notre corpus a mis en lumière quelques éléments intéressants. Contrairement aux accents non emphatiques, les accents emphatiques ont tendance à tomber sur des nombres (dans 27% des cas contre 9% pour les accents non emphatiques). Ce constat est clairement spécifique aux commentaires sportifs dans lesquels les nombres correspondent souvent à des scores et jouent un rôle significatif dans la fonction expressive. Nous observons également que les accents emphatiques tombent généralement sur la première syllabe du mot (comme indiqué par Séguinot (1976)), contrairement aux accents non emphatiques. Enfin, ils sont également souvent précédés par un silence.

Une dernière phase de notre étude s'est concentrée sur la comparaison entre les labels locaux et les annotations produites par des outils de détection automatique des proéminences. Prosoprom (Goldman, Avanzi, Lacheret-Dujour, Simon, & Auchlin, 2007) annote automatiquement chaque syllabe comme proéminente ou non. Cette annotation repose sur une série de règles apprises sur un corpus annoté en français et prend en compte différents types de caractéristiques prosodiques (F0,

durée, silences, etc.). 42,1% des syllabes accentuées dans notre corpus (H, HH, L, LL, E et F) sont considérées comme proéminentes par Prosoprom, contre 10,9% pour les syllabes non accentuées. Les syllabes 'F' atteignent le pourcentage le plus élevé (60,6%). Prosoprom a été récemment enrichi afin de proposer une annotation graduelle du niveau de proéminence (Goldman, Avanzi, Auchlin, & Simon, 2012) (PromGrad), qui définit cinq niveaux de proéminences s'échelonnant de 0 à 4. Le Tableau 4 montre la valeur moyenne de proéminence assignée à chaque label de notre corpus. Les deux niveaux de frontières pour les accents non-emphatiques sont clairement distingués par l'annotation automatique. PromGrad détermine le niveau de proéminence d'une syllabe sur base d'un calcul complexe combinant différents paramètres prosodiques (le pitch moyen, la durée, le mouvement de pitch et la durée de la pause subséquente). Si un accent est réalisé par peu de caractéristiques prosodiques différentes, il est donc peu probable qu'il obtienne un niveau de proéminence élevé. Nous observons ce cas ici pour les accents emphatiques qui sont moins caractérisés par l'allongement et ne sont généralement pas suivis par une pause. Un algorithme prenant en compte la présence d'une pause précédant la syllabe pourrait permettre de régler ce problème. Le lien existant entre notre annotation et PromGrad (Goldman et al., 2012) suggère, dans une certaine mesure, qu'il serait possible de prédire automatiquement les étiquettes locales à partir de la réalisation acoustique du corpus.

H	HH	E	L	LL	F	He	C	NA
0,9	3,2	2,7	0,7	2,8	1,5	1,4	0,3	0,2

Tableau 4 : *Niveau de proéminence moyen assigné par PromGrad (Goldman et al., 2012)*

4.2. Analyse de l'annotation globale

Une analyse de la réalisation prosodique de chaque sous-genre se trouve dans le Tableau 5. La proportion de montées correspond au pourcentage de syllabes dont le pitch est montant. Le débit d'articulation est calculé comme le nombre de syllabes divisé par la durée de l'articulation (durée du sous-genre en excluant les silences). Enfin, la proportion de silences correspond à la durée totale des silences dans chaque sous-genre divisée par la durée totale de chaque sous-genre. Les silences initiaux et finaux ne sont pas pris en compte puisqu'ils pourraient être arbitrairement attribués au sous-genre précédent ou suivant.

Une première observation intéressante concerne la valeur de pitch plus élevée du sous-genre *Excitation max*. Ce constat corrobore les résultats d'autres études attestant des valeurs de F0 plus élevées pendant et après les tirs au but (Audrit et al., 2012 ; Trouvain, 2011) qui reçoivent généralement, dans notre corpus, une étiquette *Excitation max*. Comme

	F0 moyenne (Hz)	Intensité moyenne (dB)	Proportion de montées (%)	Débit d'articulation (syll/sec)	Proportion de silences (%)
Neutre	212	29,7	2,6	5,6	46
Excité	227	33,4	2,3	6	32,8
ExMax	261	41,1	1,5	5,2	18
TNeg	215	32,6	2,7	5,6	32,7
ExRise	215	29,4	15,5	5,5	29,5

Tableau 5 : *Valeurs acoustiques moyennes des labels globaux*

prévu, le sous-genre *Excité* est réalisé avec une fréquence de pitch entre *Neutre* et *Excitation max*. Des observations similaires peuvent être effectuées pour l'intensité, avec des valeurs plus élevées pour *Excitation max* et des valeurs intermédiaires pour *Excité*. Il faut cependant souligner que le sous-genre *Tension négative* est caractérisé par un niveau de pitch relativement faible, malgré son niveau d'excitation important sur l'échelle dimensionnelle. Ceci confirme les hypothèses de Trouvain (2011) qui avance la possibilité que l'expression de la déception soit associée à une valeur de pitch réduite. Le pourcentage de montées montre une claire distinction entre la *Montée de l'excitation* et les autres sous-genres. Contrairement à ce qui pourrait être attendu, le débit d'articulation reste relativement stable à travers les différents styles. Ce phénomène a d'ailleurs déjà été souligné par d'autres études sur les commentaires sportifs (Trouvain & Barry, 2000 ; Kern, 2010). Cependant, on observe une réduction nette de la proportion de silences en fonction du niveau d'excitation, ce qui indique une montée du débit de parole en incluant les silences.

Nous avons également analysé la probabilité de transition d'un sous-genre vers un autre. La *Montée de l'excitation* a ainsi tendance à être suivie par *Excitation max* (0,38) ou *Tension négative* (0,38), en fonction du succès ou de l'échec de l'action et de l'équipe impliquée. La fin d'une séquence *Excitation max* ou *Tension négative* indique généralement la fin d'une action et est souvent suivie par un label *Neutre*. Cette matrice de transition, ainsi que la durée moyenne des segments de chaque sous-genre, pourrait être prise en compte lors de la sélection des labels globaux lors de la synthèse de nouvelles phrases.

Enfin, l'évolution des caractéristiques acoustiques a également fait l'objet d'une analyse. L'objectif était ici de déterminer si un changement significatif pouvait être observé à l'intersection entre deux sous-genres, ce qui pourrait faciliter l'annotation automatique de nouveaux corpus sur base acoustique. ProsoDyn (Goldman, 2012) fournit une telle représentation du dynamisme prosodique d'un segment de parole. La valeur moyenne des caractéristiques prosodiques est calculée sur une fenêtre

glissante de 19 syllabes. Son application sur un fichier de parole représentatif de notre corpus peut être observé en Figure 3. L'évolution du pitch, en demi-tons, est fortement corrélée au niveau d'excitation.

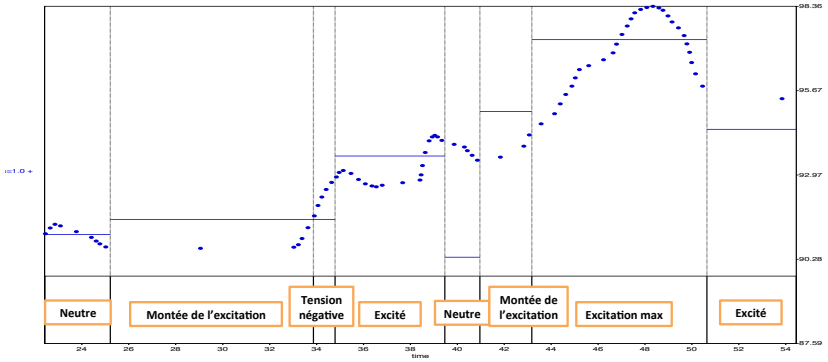

Figure 3 : *Représentation dynamique du pitch sur un segment du corpus avec ProsoDyn (Goldman, 2012)*

5. Application à la synthèse de la parole

Afin d'évaluer la validité de notre protocole d'annotation, plusieurs synthétiseurs HMM (Zen, Tokuda, & Black, 2009) ont été entraînés, en se basant sur l'implémentation du toolkit HTS (version 2.1), disponible publiquement (Zen et al., 2007). Pour chaque synthétiseur, 90% de la base de donnée correspondante a été utilisée pour l'entraînement et les 10% restants ont été utilisés pour la synthèse. Comme paramétrage du filtre, nous avons extrait les coefficients Mel Cepstraux Généralisés (MGC), traditionnellement utilisés en synthèse paramétrique. Pour la modélisation de l'excitation, le modèle déterministe plus stochastique (DSM (Drugman & Dutoit, 2012)) du signal résiduel a été utilisé afin d'améliorer le naturel de la voix générée. Des tests MOS ont été effectués afin de quantifier l'impact de l'intégration de la couche d'annotation locale et/ou globale. Dix à vingt locuteurs francophones, essentiellement des auditeurs naïfs, ont participé aux différents tests visant à évaluer la qualité des différents synthétiseurs.

Le premier synthétiseur est le système de référence (Base) et utilise comme information contextuelle une transcription phonétique manuellement vérifiée mais pas d'information prosodique. Dans un premier temps, seule l'annotation prosodique locale est utilisée afin d'entraîner un second synthétiseur (Loc). Elle permet d'enrichir l'information contextuelle exploitée par les modèles. Les résultats du test subjectif visant à comparer ces deux synthétiseurs peuvent être observés en Figure 4. Ils révèlent que l'intégration des étiquettes prosodiques locales permet d'améliorer l'expressivité tout en atteignant des niveaux d'in-

telligibilité et de qualité vocale légèrement supérieurs au synthétiseur de référence.

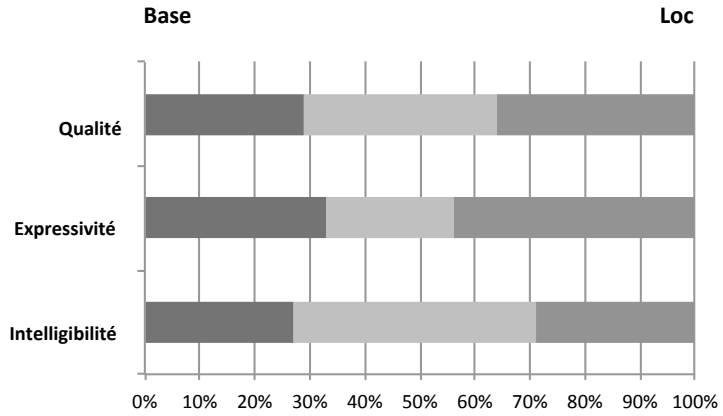

Figure 4 : *Scores de préférence pour chaque critère et les deux méthodes de synthèse, avec (Loc) et sans (Base) intégration des labels prosodiques locaux*

Nous nous sommes ensuite attelés à l'intégration des labels prosodiques globaux dans la synthèse. Trois possibilités ont été envisagées : (i) un entraînement distinct sur chaque sous-genre, (ii) l'adaptation d'un modèle neutre (entraîné sur le sous-genre *Neutre*) vers les autres sous-genres ou (iii) l'adaptation d'un modèle moyen entraîné sur l'entièreté du corpus vers les autres sous-genres. Les tests ont montré que les techniques d'adaptation, et plus particulièrement celles à partir d'un modèle moyen, permettent d'atteindre les meilleurs résultats en termes d'intelligibilité et de qualité comparé aux deux autres solutions. Cependant, la comparaison de ce modèle avec le synthétiseur de référence montre que, contrairement à ce qui était attendu, l'intégration des labels globaux ne permet pas d'améliorer le niveau d'expressivité et dégrade légèrement la qualité vocale.

Une dernière expérience nous a permis d'évaluer les effets obtenus lors de l'intégration des deux couches d'annotation pour la synthèse. Il est alors intéressant de constater que cette complète intégration, comparée à l'intégration des labels locaux uniquement, permet d'améliorer encore davantage l'expressivité, tout en atteignant des scores d'intelligibilité similaires. Cependant, on observe une légère dégradation de la qualité segmentale, probablement due aux techniques d'adaptation. Picart, Brognaux, et Drugman (2013) offrent une analyse plus complète des résultats obtenus lors de l'intégration des deux couches d'annotation en synthèse HMM.

6. Conclusion

Cet article a présenté un protocole d'annotation prosodique pour les commentaires sportifs en vue de la synthèse vocale par HMM. Deux niveaux d'annotation ont été définis afin de représenter les phénomènes accentuels et les différents sous-genres spécifiques aux commentaires sportifs. L'avantage est ici que les labels sont caractérisés par une fonction expressive distincte et par une réalisation prosodique relativement stable. Ce dernier point a été démontré sur un corpus de commentaires de basketball d'une durée totale de deux heures et devrait permettre une prédiction automatique des labels à partir du texte ou du signal de parole.

Les étiquettes locales ont ensuite été intégrées dans les informations contextuelles fournies au synthétiseur de parole tandis que les labels globaux ont permis l'entraînement de différents modèles pour chaque sous-genre. Des tests subjectifs montrent que l'intégration des labels prosodiques locaux comme information contextuelle permet d'améliorer l'expressivité, tout en conservant le niveau d'intelligibilité et de qualité. Enfin, bien que l'intégration de l'information prosodique globale seule n'améliore pas la voix de synthèse, son utilisation en complément aux informations prosodiques locales permet d'atteindre un niveau d'expressivité encore plus élevé en maintenant le niveau d'intelligibilité. On observe cependant une légère dégradation de la qualité vocale.

Remerciements

Les auteurs sont soutenus par le FNRS. Ce projet est partiellement financé par le projet Wist 3 SPORTIC de la Région Wallonne. Les auteurs remercient également S. Audrit pour son implication dans l'enregistrement du corpus et A. C. Simon et M. Avanzi pour leurs conseils avisés.

Références

Audrit, S., Pršir, T., Auchlin, A., & Goldman, J.-P. (2012). Sport in the media : A contrasted study of three sport live media reports with semi-automatic tools. In *Actes de Speech Prosody 2012* (pp. 127–130). Shanghai (Chine). Consulté sur http://sprosig.isle.illinois.edu/sp2012/uploadfiles/file/sp2012_submission_170.pdf

Beaufort, R., & Ruelle, A. (2006). eLite : système de synthèse de la parole à orientation linguistique. In *Actes des Journées d'Études sur la Parole 2006 (JEP)* (pp. 509–512). Dinard (France).

Boersma, P., & Weenink, D. (2009, May). *Praat : doing phonetics by computer (version 5.1.05) [Computer Program].* Consulté sur http://www.praat.org

Brognaux, S., Drugman, T., & Saerens, M. (2014). Synthesizing sports commentaries : One or several emphatic stresses ? In *Actes de Speech Prosody 2014.* Dublin (Irlande). Consulté sur http://tcts.fpms.ac.be/publications/

papers/2014/speechproso14_emphasis_sbtd.pdf

Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. (2012). Train&Align : A new online tool for automatic phonetic alignments. In *Actes du IEEE Workshop on Spoken Language Technologies (SLT).* Miami (USA). Consulté sur http://cental.fltr.ucl.ac.be/train_and_align/

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Acoustics Speech and Signal Processing*, *14(4)*, 1171-1179.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20(1)*, 37-46.

Di Cristo, A. (2000). Vers une modélisation de l'accentuation du français : deuxième partie. *Journal of French Studies*, *10*, 27-44.

Drugman, T., & Alwan, A. (2011). Joint robust voicing detection and pitch estimation based on residual harmonics. In *Actes de Interspeech 2011* (pp. 1973–1976). Florence (Italie). Consulté sur http://www.isca-speech.org/archive/interspeech_2011/i11_1973.html

Drugman, T., & Dutoit, T. (2012). The deterministic plus stochastic model of the residual signal and its applications. *IEEE Transactions on Audio, Speech and Language Processing*, *20(3)*, 968-981.

Drugman, T., Kane, J., & Goble, C. (2012). Resonator-based creaky voice detection. In *Actes de Interspeech 2012* (pp. 1592–1595). Portland (USA).

Goldman, J.-P. (2011). EasyAlign : An automatic phonetic alignment tool under Praat. In *Actes de Interspeech 2011* (pp. 3233–3236). Florence (Italie). Consulté sur http://www.isca-speech.org/archive/interspeech_2011/i11_3233.html

Goldman, J.-P. (2012). Prosodyn : A graphical representation of macroprosody for phonostylistic ambiance change detection. In *Actes de Speech Prosody 2012* (pp. 75–78). Shanghai (Chine). Consulté sur http://sprosig.isle.illinois.edu/sp2012/uploadfiles/file/sp2012_submission_201.pdf

Goldman, J.-P., Auchlin, A., Roekhaut, S., Simon, A. C., & Avanzi, M. (2010). Prominence perception and accent detection in French. A Corpus-Based Account. In *Actes de Speech Prosody 2010.* Chicago (USA). Consulté sur http://speechprosody2010.illinois.edu/papers/100575.pdf

Goldman, J.-P., Avanzi, M., Auchlin, A., & Simon, A. C. (2012). A continuous prominence score based on acoustic features. In *Actes de Interspeech 2012* (pp. 2414–2417). Portland (USA). Consulté sur http://www.isca-speech.org/archive/interspeech_2012/i12_2414.html

Goldman, J.-P., Avanzi, M., Lacheret-Dujour, A., Simon, A. C., & Auchlin, A. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken French. In *Actes de Interspeech 2007* (pp. 98–101). Anvers (Belgique). Consulté sur http://www.isca-speech.org/archive/interspeech_2007/i07_0098.html

Kern, F. (2010). Prosody in interaction. In D. Barth-Weingarten, E. Reber, & M. Selting (Éds), (pp. 217–237). John Benjamins.

Lacheret-Dujour, A., & Beaugendre, F. (1999). *La prosodie du français*. Paris : CNRS Editions.

Mehrabian, A., & Russel, J. A. (1974). *An approach to environmental psychology*. MIT Press.

Mertens, P. (1987). *L'intonation du français. de la description linguistique à la reconnaissance automatique.* (Thèse de doctorat non publiée). Univ. Leuven (Belgique).

Mertens, P. (2004). The Prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In B. Bel & I. Marlien (Éds), *Actes de Speech Prosody 2004* (pp. 23–26). Nara (Japan). Consulté sur http://www.isca-speech.org/archive_open/sp2004/sp04_549.pdf

Obin, N., Dellwo, V., Lacheret, A., & Rodet, X. (2010). Expectations for discourse genre identification. In *Actes de Interspeech 2010.* Makuhari (Japan).

Odgen, R. (2001). We speak prosodies and we listen to them. In *Actes du Symposium on Prosody and Interaction* (pp. 1–4). Uppsala (Sweden). Consulté sur http://speechprosody2010.illinois.edu/papers/100575.pdf

Peeters, G. (2004). A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project.

Picart, B., Brognaux, S., & Drugman, T. (2013). HMM-based speech synthesis of live sports commentaries : Integration of a two-layer prosody annotation. In *Actes du 8th ISCA Speech Synthesis Workshop (SSW8)* (pp. 19–24). Barcelone. Consulté sur http://ssw8.talp.cat/papers/ssw8_OS1-4_Picart.pdf

Séguinot, A. (1976). L'accent d'insistance en français standard. *Studia Phonetica*, *12*, 1-58.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., … Hirschberg, J. (1992). ToBI : A standard for labeling English prosody. In *Actes de l'International Conference on Spoken Language Processing (ICSLP)* (pp. 867–870). Banff (Canada). Consulté sur http://www.isca-speech.org/archive/icslp_1992/i92_0867.html

Trouvain, J. (2011). Between excitement and triumph - live football commentaries in radio vs. TV. In *Actes du 17th International Congress of Phonetic Sciences (ICPhS XVII)* (pp. 2022–2025). Hong Kong.

Trouvain, J., & Barry, W. (2000). The prosody of excitement in horse race commentaries. In *Actes du ISCA Workshop on Speech and Emotion : A Conceptual Framework for Research* (pp. 86–91). Saarbrücken (Allemagne). Consulté sur http://www.isca-speech.org/archive_open/speech_emotion/spem_086.html

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., & Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Actes du Sixth ISCA Workshop on Speech Synthesis (SSW6)* (pp. 294–299). Bonn (Allemagne). Consulté sur https://www.cs.cmu.edu/~awb/papers/ssw6/ssw6_294.pdf

Zen, H., Tokuda, K., & Black, A. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51(11)*, 1039–1064.