Une pénalité floue fondée phonologiquement pour améliorer la sélection d'unité

David Guennec Damien Lolive
IRISA - ENSSAT/Université de Rennes 1, 6 rue de Kerampont, 22305 Lannion Cedex, France
{david.guennec, damien.lolive}@irisa.fr

RÉSUMÉ

Les systèmes de synthèse par corpus reposent, sauf de rares exceptions, sur des coûts cibles et des coûts de concaténation pour sélectionner la meilleure séquence d'unités. Le rôle du coût de concaténation est de s'assurer que l'assemblage de deux segments de parole ne causera l'apparition d'aucun artefact acoustique. Pour cette tâche, des distances acoustiques (MFCC, F0) sont généralement utilisées, mais dans de nombreux cas cela ne suffit pas. Dans cet article, nous introduisons une pénalité héritée du domaine de la couverture de corpus dans le coût de concaténation afin de bloquer certaines concaténations en fonction de la classe phonologique des diphones à concaténer. En outre, une seconde version faisant appel à une fonction floue est proposée pour relâcher la pénalité en fonction du positionnement du coût de concaténation par rapport à sa distribution. Une évaluation objective montre que la pénalité est efficace et amène à un meilleur classement des séquences d'unités candidates au cours de la sélection. Une évaluation subjective révèle une performance supérieure de l'approche floue.

ABSTRACT

A Phonologically Motivated Penalty To Improve Unit Selection

Unit selection speech synthesis systems rely, except for rare cases, on target and concaténation costs for selecting the best unit sequence. The role of the concatenation cost is to insure that joining two voice segments will not cause any acoustic artefact to appear. For this task, acoustic distances (MFCC, F0) are typically used but in many cases, this is not enough. In this paper, we introduce a penalty in the concaténation cost, inherited from the field of corpus covering, in order to block some concaténations based on their phonological class. Moreover, a derived fuzzy version is proposed to relax the penalty based on the concaténation quality with respect to the cost distribution. An objective evaluation showed that the penalty is effective to better rank candidate unit sequences during selection. The subjective evaluation we conducted reveals a superior performance of the fuzzy approach.

MOTS-CLÉS : Coût de Concaténation, Synthèse Par Corpus, Sélection d'Unité.

KEYWORDS: Concaténation cost, corpus-based TTS, Unit Selection.

1 Introduction

Au cours des dernières années, la recherche en synthèse de la parole à partir du texte s'est essentiellement portée sur deux techniques. L'approche statistique paramétrique (SPSS pour *Statistical Parametric Speech Synthesis*) est la plus récente et a été l'objet de nombreux travaux universitaires ces dernières années. Elle comprend principalement la synthèse par HMM et plus récemment par

DNN (Black *et al.*, 2007; Yamagishi *et al.*, 2008; Hashimoto *et al.*, 2015). Cette méthode offre un contrôle avancé sur le signal et produit une synthèse très intelligible, mais la voix générée manque de naturel. La méthode historique, la Synthèse Par Corpus (SPC), est un raffinement de la synthèse par concaténation (Sagisaka, 1988; Hunt & Black, 1996; Taylor *et al.*, 1998; Breen & Jackson, 1998; Clark *et al.*, 2007). La SPC permet la création de synthèse de haute qualité, dont le naturel et la qualité prosodique restent inégalés par les autres méthodes grâce à l'utilisation de parole naturelle pour réaliser la synthèse. La plupart des systèmes industriels fonctionnent grâce à cette méthode qui a cependant des inconvénients, telle la difficulté à contrôler la prosodie et le risque d'artefacts de concaténation pénalisant l'intelligibilité.

Cette méthode fait intervenir la notion d'unité, laquelle est une liste de segments contigüs (des diphones généralement) dans un corpus de parole correspondant à une partie de la séquence cible de segments à synthétiser. Afin de discriminer les segments provenant du corpus qui correspondent aux besoins exprimés par l'intermédiaire de la séquence cible, la méthode habituelle est de classer les unités en évaluant le degré de ressemblance avec la séquence cible (coût cible) et le risque de créer un artefact lors de la concaténation des unités (coût de concaténation) via des fonctions des coûts. Le coût de concaténation repose principalement sur des caractéristiques acoustiques (MFCC, F0) (Stylianou & Syrdal, 2001; Tihelka *et al.*, 2014) pour évaluer le niveau de ressemblance spectrale entre deux stimuli vocaux sur et autour du point de concaténation. Ces coûts de concaténation sont toutefois loin d'être parfaits et de nombreux artefacts apparaissent à la fois dans les systèmes commerciaux et de recherche, même après un traitement post-concaténation. Plusieurs analyses ont montré que ces artefacts se produisent plus souvent sur certains phonèmes que sur d'autres (Yi, 1998; Cadic *et al.*, 2009). La concaténation sur une voyelle est par exemple plus risquée que sur une fricative sourde. Cette observation est à l'origine d'une méthode de construction de script d'enregistrement dans (Cadic *et al.*, 2009) où la couverture de «sandwichs vocaliques» vise à favoriser les concaténations sur des diphonèmes jugés peu risqués. Dans cet article, nous proposons d'intégrer ces contraintes directement dans la fonction de coût, sans l'aide d'un corpus construit avec des sandwichs vocaliques. Nous intégrons ainsi une pénalité en fonction de la classe de phonèmes dans la fonction de coût lors de la sélection d'unité. Deux versions sont proposées : d'abord en utilisant une pénalité fixe puis une fonction floue visant à rendre la pénalisation des unités plus flexible. La version floue a été utilisée dans notre système pour le Blizzard Challenge 2015 (Alain *et al.*, 2015), bien qu'elle n'aie pas été évaluée à l'époque.

L'article est organisé comme suit. Dans la section 2, le système de synthèse est présenté. La section 3 propose l'utilisation de contraintes phonologiques et l'introduction d'une pénalité dans la fonction de coût, de manière à obtenir un meilleur classement des chemins lors de la sélection. La section 4, décrit le corpus utilisé pour l'évaluation de cette nouvelle méthode. Enfin la section 5 présente les résultats des expériences menées pour évaluer les approches proposées de manière objective et perceptive.

2 Le système de synthèse de l'IRISA

Le système de synthèse de l'IRISA (Guennec & Lolive, 2014), utilisé pour les expériences présentées dans ce document, est fondé sur une approche de type sélection d'unité, réalisée via un algorithme optimal de recherche de plus court chemin dans un graphe (ici un algorithme A*). Dans les systèmes

de SPC, la fonction optimisée est habituellement écrite comme suit (Hunt & Black, 1996) :

$$U^* = \underset{U=u_1,\ldots,u_N}{\arg\min} \left(W_{tc} \sum_{n=1}^{N} C_t(u_n) + W_{cc} \sum_{n=2}^{N} C_c(u_{n-1}, u_n)\right) \quad (1)$$

où U^* est la meilleure séquence d'unités selon la fonction de coût et u_n est l'unité candidate que l'on essaie de faire correspondre à la n$^{\text{ème}}$ unité cible dans la séquence candidate U. $C_t(u_n)$ est le coût cible et $C_c(u_{n-1}, u_n)$ est le coût de concaténation. W_{tc} et W_{cc} sont les pondérations associées aux deux sous-coûts (Alías *et al.*, 2011). Ces poids sont calculés à l'aide des distributions de coût observées dans le corpus et visent à compenser les ordres de grandeur des sous-coûts comme dans (Blouin *et al.*, 2002). Le coût de concaténation est composé de distances euclidiennes sur les MFCC (hors coefficients dérivés Δ et $\Delta\Delta$), l'amplitude et la F0 :

$$C_c(u_{n-1}, u_n) = C_{mfcc}(u_{n-1}, u_n) + C_{amp}(u_{n-1}, u_n) + C_{F0}(u_{n-1}, u_n) \quad (2)$$

où $C_{mfcc}(u_{n-1}, u_n)$, $C_{amp}(u_{n-1}, u_n)$ et $C_{F0}(u_{n-1}, u_n)$ sont les trois sous-coûts de MFCC, amplitude et F0.

Dans cet article, le coût cible est mis à 0 dans l'équation 1. À la place, nous filtrons les unités candidates du corpus, en n'incluant dans le graphe que celles correspondant à un ensemble de caractéristiques linguistiques et phonétiques, que nous appelons les filtres de présélection (Donovan, 2001). Afin d'obtenir un nombre suffisant d'unités candidates u_n pour chaque unité cible t_n, que nous notons MIN_u (10 au minimum dans ce travail), les contraintes liées aux filtres peuvent être temporairement relâchées. Plus formellement, on considère que l'on dispose d'un n-uplet de J filtres modélisés par des fonctions indicatrices $f_j(u_n, t_n)$ ($j \in [0\,;J]$) valant 1 si u_n respecte la condition posée par le filtre j pour le diphone cible t_n et 0 sinon. Considérons l'ensemble des unités satisfaisant les I premiers filtres pour le diphone cible t_n :

$$O(I_n, t_n) = \left\{ u_n / \prod_{i=1}^{I_n \leq J} f_i(u_n, t_n) = 1 \right\}. \quad (3)$$

L'étape de présélection vise à rechercher, pour chaque diphone cible t_n, l'ensemble $O(I_n, t_n)$ de noeuds candidats, à intégrer dans le graphe de sélection, pour lequel I_n est maximal :

$$I_n = \arg\min card(O(I_n, t_n)) \geq MIN_u. \quad (4)$$

En conséquence, le relâchement des filtres est effectué en partant du dernier. Ainsi, l'ordre des filtres utilisés peut avoir un impact important lors de la sélection.

La principale raison de ne pas intégrer ces filtres à la fonction de coût elle-même est de réduire la taille du graphe d'unités candidates (donc de réduire le temps de sélection). Cependant, il ne faut pas perdre de vue le fait qu'ils font partie intégrante du coût de sélection. En effet, les filtres constituent un ensemble de fonctions binaires de coûts cible se fondant sur l'hypothèse suivante : si une unité ne respecte pas l'ensemble des filtres actifs, elle ne peut pas être utilisée pour la sélection. Les filtres de présélection utilisés dans ce travail (choisis empiriquement) sont les suivants :

1. Label du segment associé, diphonème ou autre (ne peut être relâché).
2. Est-ce un *Non Speech Sound* (ne peut être relâché) ?
3. Le phone est-il dans la dernière syllabe du groupe de souffle ?

4. Le phone est-il dans la dernière syllabe de la phrase ?
5. La syllabe courante est-elle en fin de mot ?

On pourrait faire valoir que plus de filtres permettrait une meilleure sélection, mais par expérience plus de raffinement dans les filtres ne s'avère pas donner de meilleurs résultats. En effet, la meilleure unité est essentiellement un compromis entre un bon coût cible et un bon coût de concaténation, ce qui implique que chaque coût doit disposer d'un panel de choix suffisant.

3 Proposition

3.1 Spécification du coût de concaténation

L'analyse des phrases contenant des artefacts de concaténation montre que certains phonèmes, en particulier les voyelles et semi-voyelles, sont plus susceptibles d'engendrer des ruptures spectrales que d'autres (occlusives et fricatives par exemple) (Yi, 1998). Les phonèmes voisés, présentant une énergie acoustique élevée ou fortement dépendants du contexte sont généralement soumis à plus de distorsions. Sur la base de cette constatation, (Cadic *et al.*, 2009) propose un critère de couverture de corpus visant à optimiser la couverture d'unités dites « sandwich ». Un sandwich vocalique est une séquence de phonèmes où un ou plusieurs noyaux syllabiques sont entourés de deux phonèmes considérés comme peu susceptibles de provoquer des artefacts lors de la concaténation (*i.e.* résistantes aux artefacts de concaténation). En ce qui concerne les coûts de concaténation, l'utilisation de la classe phonétique pour contraindre les phonèmes considérés comme problématiques pour la concaténation n'est pas une idée nouvelle (Donovan, 2001; Yi, 1998). Cependant, dans ces travaux, les coûts sont trop contraignants, essayant de trouver une unité parfaite qui n'existe que rarement. En outre, généralement, quand cette unité n'est pas trouvée, aucune unité moins ambitieuse n'est recherchée.

3.2 Une pénalité floue fondée phonologiquement

Dans notre approche, nous définissons deux méthodes de pénalisation fondées sur trois groupes phonétiques :

V (voyelle) : Voyelles, sur lesquelles une concaténation est difficilement acceptable.

A (acceptable) : Semi-voyelles, liquides, nasales, fricatives voisées et schwa. Ces phonèmes sont vus comme des points de concaténation acceptables, du moins s'il n'y a pas de meilleur choix, mais restent dangereux.

R (résistant) : les phonèmes restants (consonnes non voisées, plosives voisées), considérés comme de bons points de concaténation.

Un point clé de la méthode est de limiter le nombre de classes (seulement 3 sous-ensembles de phonèmes ici), ceci afin de ne pas ajouter trop de contraintes dans la fonction de coût. Le but de la pénalité n'est pas d'agir comme un coût à part entière, mais simplement d'introduire des connaissances qui ne sont pas capturées par le coût de concaténation pour affiner le classement des unités. Il convient de noter que les classes proposées ici sont basées sur le bon sens et il peut être nécessaire de les adapter en fonction de la langue, voire d'effectuer une étude plus poussée.

La première méthode pour appliquer la pénalité, appelée *pho-class*, consiste à attribuer une pénalité fixe $p(v)$ dépendante de la classe du phonème qui débute l'unité v : 0 pour les phonèmes de R, une pénalité légèrement supérieure à la valeur la plus élevée de C_c observée dans le corpus pour tous les phonèmes dans A. Dans notre système, pour des coûts de concaténations dans l'intervalle $[0\,;6]$, la pénalité vaut 7 pour la classe A. On attribue une pénalité bien plus importante aux voyelles (V), assez grande pour empêcher toute compensation par d'autres coûts dans la séquence candidate (une pénalité de valeur 100 est appliquée dans notre cas). Une concaténation sur les voyelles ne peut donc intervenir que si l'on n'a pas d'autre choix. Dans ce cas, un nouveau coût de concaténation C'_c est formulé comme suit :

$$C'_c(u,v) = C_c(u,v) + K(u,v) \tag{5}$$

avec $K(u,v) = p(v)$.

La deuxième méthode, appelée *fuzzy-pho-class*, consiste à moduler la pénalité dans certains cas. Ainsi, nous introduisons une fonction floue de pondération multipliant chaque pénalité par un poids compris entre 0 et 1, comme le montre la figure 1. Elle décrit le degré de satisfaction de l'unité candidate à l'égard de la qualité de concaténation. En faisant l'hypothèse que les distributions des coûts de MFCC, d'amplitude de F0 définies dans (2) suivent des lois normales, nous définissons deux seuils pour chaque sous-coût. Ces distributions sont estimées à l'aide du corpus de parole en calculant le sous-coût de concaténation pour la F0, l'amplitude et les MFCC en utilisant toutes les unités présentes dans le corpus. Par exemple, les deux seuils T^1_{F0} et T^2_{F0} pour le sous-coût de F0 peuvent être définis comme suit :

$$\begin{aligned} T^1_{F0} &= \mu_{F0} - \sigma_{F0} \\ T^2_{F0} &= \mu_{F0} + \sigma_{F0} \end{aligned} \tag{6}$$

où μ_{F0} et σ_{F0} désignent l'espérance et l'écart-type de $C_{F0}(u,v)$. Formellement, la fonction floue pour le sous-coût de F0 est définie comme suit :

$$f_{F0}(u,v) = \begin{cases} 0 & \text{si } C_c(u,v) < T^1_{F0}, \\ 1 & \text{si } C_c(u,v) > T^2_{F0}, \\ 1 - \frac{(T^2_{F0} - C_c(u,v))}{(T^2_{F0} - T^1_{F0})} & \text{sinon.} \end{cases} \tag{7}$$

Le choix de cet intervalle de tolérance est motivé par l'observation de la répartition des coûts réels sur le corpus de voix. Pour être complet, le choix des seuils devrait être différencié selon le type de sous-coût et optimisé séparément (ce qui n'est pas le cas ici). Enfin, la pénalité est modifiée de la façon suivante :

$$K(u,v) = (f_{mfcc}(u,v) + f_{amp}(u,v) + f_{F0}(u,v)) * p(v)$$

où $f_{mfcc}(u,v)$, $f_{amp}(u,v)$ et $f_{F0}(u,v)$ correspondent aux fonctions floues de la forme décrite dans la figure 1 pour les MFCC, l'amplitude et la F0 respectivement (la figure prend l'exemple de la F0, mais les autres fonctions sont identiques). Avec ces fonctions, l'idée principale est de diminuer la pénalité lorsque l'unité a une valeur de sous-coût de concaténation qui est statistiquement parmi les meilleures. Si le coût de concaténation est au-dessus du seuil le plus élevé alors la pénalité complète doit être appliquée car l'unité considérée est alors parmi les pires possibles. Entre les deux seuils, la pénalité est progressivement augmentée au fur et à mesure que le coût de concaténation augmente.

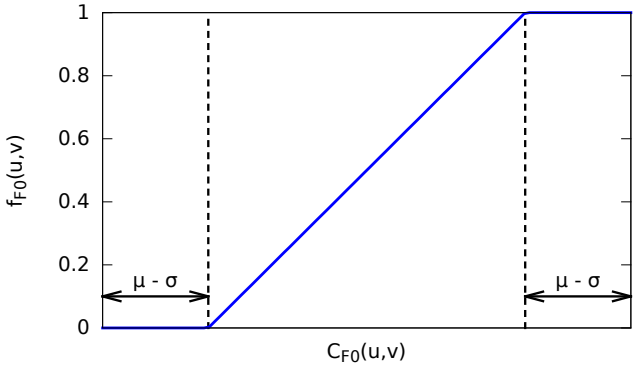

FIGURE 1 – La fonction floue $f_{F0}(u, v)$ (ici pour la F0) sur la distribution des sous-coûts $C_{F0}(u, v)$. Le poids 0 (resp. 1) est accordé aux unités qui ont un coût de concaténation parmi les 15 % les plus faibles (resp. les plus élevés). Entre ces seuils, le poids augmente de manière linéaire.

4 Description du corpus

Pour réaliser les expériences, deux corpus de parole ont été utilisés. Le premier est un corpus extrait d'un livre audio expressif, nommé ici *Audiobook*. Il contient 10h45 de parole et est échantillonné à 44,1 kHz avec un encodage sans perte en mono-canal. Le locuteur est un homme et la F0 moyenne des segments voisés est faible, à seulement 87Hz dans l'ensemble du corpus. Les données ont été automatiquement annotées en utilisant le procédé décrit dans (Boeffard *et al.*, 2012) et en utilisant le toolkit ROOTS (Chevelu *et al.*, 2014). Le corpus est composé de 3339 énoncés, avec 388251 phonèmes et 31491 NSS (*Non Speech Sound*). Sa couverture de diphonème est incomplète à 78 %, mais seuls des diphonèmes très rares/irréalisables en français en sont exclus.

Le second corpus utilisé pour nos expériences est ici nommé *IVS*. Il a été enregistré à des fins de synthèse de parole avec pour but de l'intégrer à un système vocal interactif. Le script d'enregistrement a été construit spécifiquement pour couvrir tous les diphonèmes présents en français. Il comprend également un ensemble de mots très usités dans le domaine des télécommunications. Il dispose d'une voix féminine échantillonnée à 16kHz (codage sans perte, 1 canal) avec une F0 moyenne sur les segments voisés assez basse à 163Hz. Le corpus est composé de 7662 énoncés, 239260 phonèmes et 20424 NSS pour 7h05' de parole. Le corpus de test utilisé pour les expériences consiste en 100 phrases issues de styles très différents. Ce corpus est distinct d'*IVS* et d'*Audiobook*. La langue des corpus utilisés est le français.

5 Expériences

Premièrement, une analyse du comportement de nos trois méthodes (*baseline*, *pho-class* et *fuzzy-pho-class*) en termes de coût de concaténation est effectuée. Les unités résistantes (classe R) sont considérées à part. Ensuite, l'évaluation subjective que nous avons menée pour valider l'approche proposée est présentée avec ses résultats.

IVS	Unités résistantes (R)		Unités non résistantes (A, V)		Toutes les unités	
	μ (*std*)	Nb.	μ (std)	Nb.	μ (std)	Nb.
baseline	2.90 (0.69)	582	3.14 (0.70)	1249	3.06 (0.71)	1831
pho-class	3.28 (0.92)	1025	3.35 (0.88)	813	3.31 (0.90)	1838
fuzzy-pho-class	3.35 (0.92)	**1095**	**2.58** (0.42)	1169	**2.95** (0.80)	2264
Audiobook	Unités résistantes (R)		Unités non résistantes (A, V)		Toutes les unités	
	μ (std)	Nb.	μ (std)	Nb.	μ (std)	Nb.
baseline	2.44 (0.52)	606	2.90 (0.60)	1057	2.74 (0.61)	1663
pho-class	2.65 (0.71)	865	3.14 (0.78)	785	2.88 (0.78)	1650
fuzzy-pho-class	2.65 (0.64)	**907**	**2.47** (0.38)	1139	**2.55** (0.52)	2046

TABLE 1 – Coûts de concaténation sans pénalité suivant les trois stratégies (pénalités soustraites *a posteriori*). R, A et V désignent les classes introduites section 3.2, et Nb. le nombre de concaténations.

5.1 Analyse des coûts de concaténation

Premièrement, nous avons étudié l'évolution des coûts en utilisant les trois systèmes et en comparant (1) les coûts moyens de concaténation des unités résistantes seulement (classe R), (2) les unités non-résistantes seulement (classes A, V) et (3) les deux à la fois, à chaque fois en excluant de nos statistiques les phonèmes contigüs. Tous ces résultats sont présentés dans la table 1. Comme on peut le voir, le système *baseline* a des coûts moins élevés que le système *pho-class*, à la fois pour les unités résistantes et non résistantes. L'explication de ce fait est que *pho-class*, en pénalisant les unités non-resistantes, favorise les unités résistantes même si leur coût de concaténation est plus important. Le nombre de concaténations effectuées sur les unités résistantes (1025 pour *IVS*) est significativement plus élevé en comparaison du système *baseline* (582 pour *IVS*). En ce qui concerne *fuzzy-pho-class*, les résultats en terme de nombre de concaténations sont plus nuancés. En effet, étant donné que les faibles coûts de concaténation sur les unités non résistantes sont moins pénalisés, le système *fuzzy-pho-class* obtient le coût le plus faible pour les unités non résistantes. L'introduction de pénalités variables permet d'évaluer les unités avec plus de précision (et plus de pertinence) qu'avec *pho-class*, pour lequel toutes les unités pénalisées le sont de manière équivalente. En contrepartie, le nombre de concaténations augmente globalement pour le système *fuzzy-pho-class*, ce qui n'est pas un problème étant donné que celles-ci sont mieux choisies. Il est intéressant de mentionner que ces résultats se retrouvent sur les deux voix. On peut alors les considérer comme – raisonnablement – indépendants du type de voix.

Pour résumer, la pénalité semble se comporter comme prévu, permettant de favoriser des séquences d'unités avec un moindre coût sur les unités sensibles tout en maximisant les concaténations sur les unités jugées résistantes.

5.2 Évaluation subjective

Pour évaluer les améliorations apportées par les deux méthodes proposées, nous avons conduit deux évaluations subjectives de type MUSHRA impliquant respectivement 8 et 9 auditeurs pour les voix *IVS* et *Audiobook*. Chaque test comprenait 15 étapes, les testeurs écoutant pour chacune 3 échantillons de parole synthétisée de la même phrase, une pour chacun des trois systèmes *baseline*, *pho-class* et

	IVS	*Audiobook*
baseline	51.68 ± 3.51	49.83 ± 3.48
pho-class	56.72 ± 3.59	50.38 ± 3.48
fuzzy-pho-class	$\mathbf{57.34} \pm 3.50$	$\mathbf{53.06} \pm 3.42$

TABLE 2 – Résultats des tests d'écoute avec intervalles de confiance à 95%. L'approche *fuzzy-pho-class* obtient le meilleur score, suivi par *pho-class* puis finalement par *baseline*.

fuzzy-pho-class. Les testeurs notaient ensuite la qualité globale de chaque échantillon sur une échelle de 0 (mauvais) à 100 (excellent). Les conditions de test et le choix des échantillons sont conformes aux recommandations de l'UIT-T.

Les résultats de ces tests sont présentés dans la table 2. Pour les deux voix, l'approche *fuzzy-pho-class* obtient les meilleurs résultats, suivie du système *pho-class* avec des résultats intermédiaires. Il est intéressant de noter que dans le cas de la voix *IVS*, ces deux systèmes obtiennent des scores similaires. À l'inverse, la différence est plus importante pour la voix *Audiobook*. L'explication de ce phénomène réside certainement dans le fait que dans le cas d'une voix neutre (*IVS*), la faible variabilité des unités renforce les résultats de l'approche *pho-class*. À l'inverse, dans le cas d'une voix plus expressive – comme c'est le cas pour *Audiobook*– la variabilité des unités est bien plus importante et les contraintes sur les phonèmes à concaténer sont bien plus fortes. Dans ce cas, *fuzzy-pho-class* est plus efficace grâce à la flexibilité de l'approche floue, bien plus adaptable que la pénalité fixe octroyée par la méthode *pho-class*.

6 Conclusion

Dans cet article, nous avons proposé une nouvelle fonction de coût de concaténation introduisant une pénalité sur la base de contraintes phonologiques. Une seconde approche, nuançant cette pénalité en fonction de la répartition des coûts via une fonction d'appartenance floue, a également été présentée. La pénalité permet d'éviter des artefacts lors de la synthèse. La version floue permet en outre de garder suffisamment de variabilité lors de la sélection. Les expériences subjectives que nous avons menées montrent une meilleure performance de la version floue à la fois sur une voix neutre et une voix expressive. On montre ainsi que le coût de concaténation ne saisit pas toute l'information perceptive et que l'ajout de certaines préférences sur le type d'unités à concaténer améliore la qualité de la parole synthétisée. Des modèles flous plus avancés peuvent maintenant être étudiés, de manière à améliorer encore la méthode. Il est également nécessaire de mener d'autres travaux sur la classification des phonèmes dans les ensembles R, A et V. En effet, la classification sur laquelle ils reposent devrait être comparée à d'autres, peut-être plus adéquates. Par exemple, les liquides et les semi-voyelles se montrant souvent problématiques, on pourrait considérer leur ajout dans la classe V. Plus de classes pourraient également être ajoutées, en apportant toutefois un soin particulier à ne pas se montrer trop contraignant, trop de contraintes dégradant généralement la qualité de la synthèse. Nous souhaitons également mener une étude sur la dépendance de ces classes à la langue. En outre, l'efficacité de l'approche pourrait être évaluée sur des corpus construits à l'aide d'un script d'enregistrement optimisant le taux de couverture en sandwichs vocaliques (suivant la méthodologie de (Cadic *et al.*, 2009)).

Références

ALAIN P., CHEVELU J., GUENNEC D., LECORVÉ G. & LOLIVE D. (2015). The irisa text-to-speech system for the blizzard challenge 2015. In *The Blizzard Challenge*.

ALÍAS F., FORMIGA L. & LLORÁ X. (2011). Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms : A proof-of-concept. *Speech Communication*, **53**(5), 786–800.

BLACK A. W., ZEN H. & TOKUDA K. (2007). Statistical Parametric Speech Synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, **4**.

BLOUIN C., ROSEC O., BAGSHAW P. & D'ALESSANDRO C. (2002). Concatenation cost calculation and optimisation for unit selection in TTS. In *IEEE Workshop on Speech Synthesis*.

BOEFFARD O., CHARONNAT L., MAGUER S. L., LOLIVE D. & VIDAL G. (2012). Towards Fully Automatic Annotation of Audio Books for TTS. In *Proc. of LREC*, p. 975–980.

BREEN A. P. & JACKSON P. (1998). Non-uniform unit selection and the similarity metric within BT's Laureate TTS system. In *Proc. of the ESCA Workshop on Speech Synthesis*, p. 373–376.

CADIC D., BOIDIN C. & D'ALESSANDRO C. (2009). Vocalic sandwich, a unit designed for unit selection TTS. In *Tenth Conference of ISCA*, p. 2079–2082.

CHEVELU J., LECORVÉ G. & LOLIVE D. (2014). ROOTS : a toolkit for easy, fast and consistent processing of large sequential annotated data collections. In *Proc. of LREC*, p. 619–626.

CLARK R. A., RICHMOND K. & KING S. (2007). Multisyn : Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, **49**(4), 317–330.

DONOVAN R. E. (2001). A new distance measure for costing spectral discontinuities in concatenative speech synthesizers. In *ITRW*.

GUENNEC D. & LOLIVE D. (2014). Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System. In *Proc. of TSD*, p. 432–440.

HASHIMOTO K., OURA K., NANKAKU Y. & TOKUDA K. (2015). The Effect Of Neural Networks In Statistical Parametric Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 4455–4459, Melbourne.

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of ICASSP*, volume 1, p. 373–376.

SAGISAKA Y. (1988). Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. of ICASSP*, p. 679–682.

STYLIANOU Y. & SYRDAL A. K. (2001). Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. of ICASSP*, **2**, 837–840.

TAYLOR P., BLACK A. W. & CALEY R. (1998). The architecture of the Festival speech synthesis system. In *Proc. of the ESCA Workshop in Speech Synthesis*, p. 147–151.

TIHELKA D., MATOUŠEK J. & HANZLÍČEK Z. (2014). Modelling F0 Dynamics in Unit Selection Based Speech Synthesis. In *International Conference on Text, Speech and Dialogue*, p. 457–464.

YAMAGISHI J., LING Z. & KING S. (2008). Robustness of HMM-based speech synthesis. In *Ninth Annual Conference of the International Speech Communication Association*.

YI J. (1998). *Natural-sounding speech synthesis using variable-length units*. Rapport interne, Massachusetts Institute of Technology.