Synthèse vocale par sélection linguistiquement orientée d'unités non-uniformes : LiONS

Vincent Colotte et Richard Beaufort

LORIA
Campus scientifique BP239 59506 Vandœuvre-lès-Nancy, France
MULTITEL
Parc initialis, avenue Copernic - 7000 Mons, Belgique
Mél : colotte@loria.fr - beaufort@multitel.be

ABSTRACT

This paper presents a new Non-Uniform Units selection-based Text-To-Speech synthesizer. Nowadays, systems use prosodic models that do not allow the prosody to vary as far as we should hope, involving a listening comfort degradation. Our system has the advantage to avoid the use of prosodic model. Speech units selection builds its features set exclusively from the linguistic information generated by the natural language analysis. We also present an original method to automatically weight these features. Last but not least, selected units are not restrained by a predetermined prosody. Our linguistic features proved their efficiency : we obtain a various prosody and the units concatenation is performed without resorting to heavy signal processing.

1. INTRODUCTION

De nos jours, les systèmes de synthèse de la parole à partir du texte reposent sur une architecture séquentielle et modulaire, classiquement divisée en trois blocs majeurs : traitement de la langue, sélection des unités, et traitement numérique du signal.

Tout système de synthèse basé sur cette architecture a besoin d'une base de données vocales contenant les différentes unités de parole à utiliser. Les premiers systèmes à utiliser des bases de données vocales en synthèse n'employaient qu'un seul exemplaire de chaque unité (généralement des diphones). L'idée sous-jacente était de régénérer la fréquence fondamentale, la prosodie et la durée souhaitées (lors de la première étape) dans la dernière étape. Malheureusement, ces modifications acoustiques apportées aux unités de manière à obtenir les caractéristiques prosodiques demandées entraînent une détérioration de la qualité et du naturel de la parole de synthèse : la voix dénote un caractère plus ou moins métallique.

Pour conférer à la parole de synthèse un caractère plus naturel, proche de celui de la parole humaine, les chercheurs [13, 3, 2, 8] ont voulu mettre en œuvre le principe de *choose the best to modify the least* [1] : la recherche de l'unité souhaitée est réalisée sur un corpus qui contient non plus un seul, mais plusieurs représentants de chaque unité, de sorte que les modifications acoustiques à apporter à l'unité sélectionnée soient réduites au strict minimum.

Au niveau de la sélection d'unités, cela s'est traduit par la recherche des unités qui d'une part correspondent au mieux aux unités décrites par l'analyse de la langue, et d'autre part se concatènent au mieux de manière à éviter autant que possible les modifications du signal. Cette méthode de recherche a de ce fait impliqué le calcul d'un double coût - coût cible et coût de concaténation -, dont naquît la notion d'unité non-uniforme.

Par *non-uniforme*, on entend que l'unité peut varier sur deux plans : en longueur (diphone, phone, semi-phone, syllabe, ou mot...) et au niveau de ses réalisations acoustiques. La variation en terme de réalisation acoustique signifie qu'une même unité peut (et même doit) être présente plusieurs fois dans le corpus, chaque instance de l'unité se différenciant des autres au niveau acoustique. Les unités ne sont donc plus neutralisées ; elles conservent les variations obtenues au moment de l'élocution.

Toutes les unités sont identifiées par un nombre fini de caractéristiques. Ces caractéristiques doivent être pertinentes, c'est-à-dire représentatives des variations acoustiques qui peuvent apparaître dans une unité. Tous les systèmes mis au point emploient, en proportions variables, des caractéristiques linguistiques, acoustiques et symboliques. Les caractéristiques linguistiques sont directement issues de l'analyse de la langue, tandis que les caractéristiques acoustiques et symboliques sont déterminées à l'aide de modèles prosodiques. Parmi les caractéristiques acoustiques, on notera la fréquence fondamentale et la durée. La caractéristique symbolique récurrente, quant à elle, est le ton.

Pondération. L'évaluation de l'importance des différentes caractéristiques intervenant dans l'un ou l'autre coût est un point crucial pour l'élaboration d'un tel système. En effet, il a rapidement paru évident que toutes les caractéristiques ne devaient pas être mises sur un pied d'égalité, certaines influençant plus que d'autres la qualité du résultat obtenu. Des recherches ont dès lors été réalisées afin de trouver la pondération idéale à appliquer lors du processus de sélection. Parmi les systèmes mis au point, aucun ne propose une pondération automatique de la totalité de ses caractéristiques. Les pondérations sont toujours revues manuellement à une étape ou une autre du processus.

La première pondération proposée, notamment utilisée par le système CHATR [2, 8, 6], implique de former un réseau entre tous les sons du corpus. Le réseau étant mis en place, une phase d'apprentissage commence. Son objectif est d'améliorer la similarité acoustique entre une phrase de référence et le signal donné par le système. Cette amélioration s'obtient en ajustant les poids attribués aux caractéristiques, par itérations successives ou par régression linéaire. Deux inconvénients sont inhérents à cette méthode : d'une part la charge de calcul, bien que celui-ci

soit réalisé *off-line*, et d'autre part le nombre restreint de caractéristiques que le calcul permet de pondérer. Généralement, un ajustement manuel de certains poids reste nécessaire. Pour alléger la charge de calcul, certains auteurs [1, 4] opèrent un regroupement (*clustering*) des sons en ne gardant qu'un représentant de ceux-ci, le *centroïde*, sur lequel sont réalisés les calculs de la sélection.

Une autre pondération repose sur la représentation du corpus en arbre phonétique et phonologique [5, 14]. Lors de la sélection, Breen et Jackson [5] recherchent des unités candidates présentant un contexte identique à celui de l'unité cible. Les caractéristiques qu'ils emploient, acoustiques et linguistiques, ne sont cependant pas pondérées automatiquement. Taylor et Black [14], de leur côté, ont une approche descendante tâchant de retrouver la plus longue unité possible parmi les phrases du corpus. Ils commencent au niveau de la phrase complète, espérant la retrouver telle quelle. Si celle-ci n'est pas trouvée, la recherche descend au niveau inférieur, celui des groupes de mots. Le système descend ainsi de niveau en niveau jusqu'à ce qu'une unité, la plus longue possible, soit trouvée dans le corpus. Cette méthode présente malheureusement le risque d'obtenir une synthèse par mots, ce qui a pour conséquence d'introduire dans le signal de franches coupures prosodiques, gênantes pour l'auditeur.

Modèle prosodique. L'état de l'art en synthèse de la parole par sélection d'unités non-uniformes met en lumière deux inconvénients des systèmes actuels : l'obligation de recourir à un modèle prosodique dépendant de la langue, et la trop grande rigidité de la prosodie générée.

Le processus d'élaboration ou d'acquisition d'un modèle prosodique est une étape obligée, parce que les caractéristiques acoustiques et symboliques sont utilisées lors de la sélection. En outre, le modèle prosodique est dépendant de la langue, ce qui signifie qu'il est à réadapter ou à réentraîner pour chaque langue à traiter. Enfin, la modélisation prosodique est une tâche ardue, difficile à appréhender dans son ensemble même par les linguistes, pourtant habitués à étudier le domaine.

Les modèles prosodiques actuels tâchent de modéliser les phrases déclaratives, impératives et interrogatives. Les méthodes employées se répartissent entre systèmes experts et modèles appris. Dans les deux cas, cependant, les variations prosodiques sont restreintes au minimum, se limitant à une prosodie standard. L'avantage de ces modèles est qu'ils proposent une prosodie "toujours" correcte. L'inconvénient, par contre, est qu'un même patron prosodique se répète inlassablement de phrase en phrase, ce qui installe une certaine monotonie à l'écoute. De là découle une diminution incontestable de la satisfaction de l'auditeur en présence de parole de synthèse.

Constat. Ainsi l'objectif visé par les systèmes à sélection d'unités non-uniformes est de conférer à la parole de synthèse un caractère plus naturel, plus proche de celui de la parole humaine que ne l'était la parole générée par les systèmes antérieurs. De par l'utilisation d'unités non-uniformes, c'est-à-dire non-neutralisées et de longueur variable, l'objectif est atteint. Cependant, les restrictions dues à l'utilisation de modèles prosodiques limitent les variations des patrons prosodiques au strict minimum. Il en découle une diminution incontestable de la satisfaction de l'auditeur en présence d'une parole de synthèse quelque peu monotone.

Les seuls travaux essayant de s'affranchir du modèle prosodique sont ceux de Prudon [12, 11], qui n'utilisent que trois caractéristiques linguistiques pour la sélection des unités : le nom du phonème, sa position dans le mot et sa position dans la syllabe. Malheureusement, les unités sélectionnées à l'aide de ces critères entraînent des discontinuités acoustiques qui nécessitent un recours à du traitement du signal. Le côté naturel de la parole générée par le système de Prudon est de ce fait dégradé. Ainsi, ces résultats semblent corroborer les dires de Breen, qui affirmait : "if a system existed which contained a complete inventory of sounds, then an unordered list of adequate features would be sufficient to select the desired sound. Unfortunately this is not the case" [5].

C'est sur la base de ce constat que nous proposons maintenant un système de synthèse, libéré de tout modèle prosodique et de tout traitement du signal, dont les variations mélodiques peuvent à juste titre être qualifiées de naturelles. Nous présenterons le système dans son intégralité : la méthode originale de pondération des caractéristiques linguistiques, ainsi que la procédure de sélection, affranchie de toute référence à un modèle prosodique.

2. LE SYSTÈME LIONS

Nous avons pu mettre en lumière qu'une description linguistique suffisamment fine de la phrase devrait permettre à elle seule d'en régir la prosodie, sans pour autant la contraindre. Tout l'art réside dès lors dans un choix réfléchi des caractéristiques à utiliser.

Notre nouvelle méthode de synthèse se divise classiquement en une phase d'entraînement et une phase d'application. Dans les deux phases, le même moteur d'analyse linguistique est utilisé pour l'extraction des caractéristiques linguistiques, de manière à conférer une certaine homogénéité au système. Il s'agit du module de traitement du langage naturel d'*elite*, système de synthèse de Multitel, signifiant *« Enhanced LInguistically-based TExt-to-speech synthesizer »* et utilisant la technique MBROLA pour la sélection et la génération du signal de parole).

Avant tout, il est nécessaire d'établir la liste des caractéristiques linguistiques pertinentes pour la phase de sélection. Cette liste étant faite, l'entraînement consiste en un étiquetage et une segmentation des corpus, ainsi qu'en une pondération des caractéristiques linguistiques.

La phase d'application est la phase de synthèse à proprement parler. Elle est opérée sur une phrase textuelle présentée à l'entrée du système de synthèse.

2.1. Entraînement

Choix des caractéristiques. L'utilisation exclusive de caractéristiques linguistiques pour la sélection oblige à ajouter à l'ensemble de caractéristiques utilisées telles que les phonèmes environnant la cible, la syllabation, le nombre de syllabes du mot, la place d mot dans la phrase...un certain nombre d'informations qui vont influer indirectement sur la prosodie. Le moteur d'analyse utilisé doit donc être suffisamment puissant pour déterminer les informations supplémentaires nécessaires. Parmi celles-ci, on notera :

– les accents primaires et secondaires du mot, qui sont strictement linguistiques et peuvent être extraits de lexiques de phonétisation,
– les groupes de souffle :
  – qui englobent plusieurs catégories et permettent de déterminer implicitement les lieux où les accents de groupe peuvent apparaître, et
  – qui permettent d'adapter la syllabation du texte.

Préparation des corpus. Les corpus écrit et oral sont préparés séparément.

A l'aide du moteur d'analyse de la langue, chaque phrase du corpus écrit est annotée comme suit : nombre de mots et place des mots dans la phrase, syllabation et phonétisation des mots, synthèse en terme de critères articulatoires des contextes phonémiques pour chaque phonème.

Les phrases du corpus oral sont segmentées en phonèmes et en diphones. On crée alors une base de données dans laquelle chaque phonème du corpus est recensé. Pour chacun d'eux, on calcule — et on ajoute dans la base — les caractéristiques acoustiques qui seront utiles au coût de *concaténation* : fréquence fondamentale, coefficients LPC et intensité.

Pondération des caractéristiques linguistiques. La méthode mise au point offre l'avantage d'être complètement automatique.

Du fait de leurs différences articulatoires, les phonèmes se comportent différemment les uns des autres dans un même contexte d'élocution. De ce fait, une seule pondération des caractéristiques linguistiques ne serait pas pertinente ; il est préférable de pondérer les caractéristiques pour chaque phonème indépendamment.

Pour un phonème donné, on prélève sur le corpus oral l'ensemble de ses réalisations acoustiques. A l'aide de l'algorithme K-Means, on répartit celles-ci en sous-ensembles. Un sous-ensemble regroupe les réalisations acoustiques considérées comme similaires. L'indice de similarité utilisé est la distance perceptuelle de Kullback-Liebler [9]. Le nombre optimal de sous-ensembles est calculé automatiquement sur le principe de la maximisation du rapport des variances. L'initialisation est fixée à plusieurs sous-ensembles, dans lesquels les réalisations acoustiques sont réparties en fonction de leur durée.

Les sous-ensembles étant constitués, la pondération des caractéristiques linguistiques peut commencer. L'objectif de la pondération est de déterminer dans quelle mesure chaque caractéristique permet de distinguer plusieurs sous-ensembles, chaque sous-ensemble étant vu comme une classe à choisir, une décision à prendre. La méthode la plus appropriée, de ce fait, est de construire un *arbre de décision*.

La génération d'un arbre de décision est basée sur le concept d'entropie. L'entropie est une mesure de l'agitation d'un système. Plus le système est agité, moins il présente de l'information. Ainsi, le calcul de l'entropie pour une liste de caractéristiques permet de classer celles-ci en fonction de l'information que chacune d'elles contient. Plus l'entropie d'une caractéristique est basse, plus elle est informative, et donc plus elle est pertinente.

Dans le cas qui nous occupe, l'entropie d'une caractéristique est calculée sous la forme du Rapport de Gain. Le Rapport de Gain permettra d'une part de déterminer l'ordre d'importance des caractéristiques, et servira d'autre part de pondérateur des caractéristiques lors du calcul du coût cible. Ce mode d'apprentissage des poids a l'avantage de rester cohérent avec les caractéristiques de la parole du locuteur. Cet avantage n'est pas présent lorsque l'on utilise un modèle prosodique appris sur des locuteurs différents pour obtenir une prosodie standard.

2.2. Application

Analyse linguistique. Pour une phrase présentée au système, l'analyse linguistique va générer les phonèmes correspondants ainsi que les caractéristiques linguistiques qui y sont associées. Nous qualifierons de *cible* toute paire {*phonème*, *caractéristiques*}.

Sélection des unités de parole. La sélection se divise en trois étapes : (1) pré-sélection d'unités phonémiques candidates et calcul du coût cible pour chaque candidat, (2) passage à une représentation diphonique, et (3) sélection des unités qui minimisent le double coût {*cible*, *concaténation*}.

Pré-sélection. Pour une cible donnée, tous les candidats doivent au moins avoir la même étiquette, à savoir le nom du phonème. Une pré-sélection plus drastique peut restreindre les candidats à ceux qui présenteraient certaines valeurs pour quelques caractéristiques prépondérantes.
Le calcul du coût cible de chaque unité candidate est réalisé à ce stade. Dans ce calcul, les caractéristiques sont pondérées à l'aide des poids déterminés lors de l'entraînement. Le coût cible CC d'un candidat j pour le phonème i correspond donc à une sommation pondérée de ses caractéristiques :

$$CC(cand_j, pho_i) = \sum_{k=1}^{N} W_k^i \times C_k^j \qquad (1)$$

où :
– $(cand_j, pho_i)$ est le candidat j pour le phonème i,
– k varie de 0 à N, le nombre de caractéristiques,
– W_k^i est le poids accordé par l'entraînement à la caractéristique k pour le phonème i, et
– C_k^j est la valeur de la caractéristique k pour le candidat j.

Représentation diphonique. A cette étape, les diphones candidats que l'on désire retenir sont uniquement ceux que l'on peut former à partir de phonèmes candidats adjacents dans le corpus. Cependant, si un diphone cible n'a pas de candidat, on crée des diphones candidats contenant le phonème cible en partie gauche ou en partie droite, selon le diphone dont on a besoin. Le coût cible de chaque diphone candidat est la somme des coûts des deux phonèmes candidats qui le constituent.

Sélection des unités. La sélection est opérée de manière classique, par résolution du treillis de possibilités à l'aide de l'algorithme de Viterbi. Le résultat de la sélection est le chemin, dans le treillis de diphones, qui minimise le double coût {*cible*, *concaténation*}. Nous venons de voir que le coût cible est pré-calculé au moment de la pré-sélection. Le coût de concaténation, par contre, est résolu lors du parcours du treillis de possibilités.
Le coût de concaténation a été défini comme la *distance acoustique qui existe entre les unités à concaténer*. Pour

calculer cette distance, le système a donc besoin de caractéristiques acoustiques, qu'il prélève aux frontières des unités à concaténer : fréquence fondamentale, spectre, énergie et durée. La distance, et donc le coût, est obtenue en sommant en autres sur :
– la différence au niveau de la fréquence fondamentale,
– la distance spectrale (de type *Kullback-Liebler*),
– la différence d'énergie,
– ...
La sommation est bien évidemment pondérée, mais la pondération, contrairement à celle du coût cible, n'est pas apprise automatiquement lors de l'entraînement : elle est déterminée manuellement, et favorise principalement la distance spectrale et la différence de fréquence fondamentale.

Il faut encore noter que le double coût {*cible*, *concaténation*} est lui-même pondéré, de sorte que le coût cible et le coût de concaténation n'ont pas le même poids dans le choix des meilleurs candidats. Actuellement, cette pondération est encore semi-manuelle : le système, dans sa globalité, repose donc encore sur deux molettes, dont le réglage est lié à un corpus de quelques phrases nous permettant d'évaluer la qualité de la synthèse.

Concaténation des unités de parole. Les critères linguistiques utilisés dans la sélection ont ici montré leur pertinence : les unités choisies se concatènent sans discontinuité. Aucun traitement du signal, hormis la concaténation en elle-même, n'est de ce fait nécessaire. La suite de diphones sélectionnés est concaténée acoustiquement, à l'aide d'une technique de type *OverLap and Add* [7, 10] : les valeurs de *pitch* sont utilisées pour améliorer l'accolement des diphones.

3. CONCLUSION

Notre système de synthèse a libéré le système de sélection d'unités de tout modèle prosodique, qu'il soit acoustique ou symbolique, de manière à autoriser plus de variations dans la prosodie des phrases générées. Dans un même temps, le système a conservé l'avantage du faible recours au traitement du signal aux frontières des unités.
Pour ce faire, les caractéristiques de la sélection ont été choisies exclusivement parmi les informations linguistiques de l'analyse de la langue. Aux critères linguistiques classiques, tels que les phonèmes de contexte, ont été ajoutées de nouvelles informations, comme les groupes de souffle, propices à décrire non pas la prosodie des unités, mais le comportement prosodique potentiel de celles-ci. Cet apport prendra plus d'ampleur lors de la mise en place d'outils pour une construction automatisée du corpus. Pour l'instant, la base de donné actuelle est constituée de 1h15 de parole composée de 800 phrases environ tirées de nouvelles télévisées.
Les paramètres linguistiques choisis, mais également leur méthode de pondération, ont prouvé leur efficacité. D'une part, les unités sélectionnées à l'aide de ces critères se concatènent sans discontinuité acoustique, ce qui permet d'éviter de recourir à du traitement du signal, qui aurait dégradé le côté naturel de la parole générée. D'autre part et surtout, la prosodie des phrases générées peut à juste titre être qualifiée de naturelle, au vu de ses possibilités de variations.
Ceci présente indéniablement une nouveauté importante dans le domaine de la synthèse par sélection d'unités non-uniformes, au profit du confort d'écoute de l'auditeur confronté à de la parole de synthèse. Les travaux futurs mettront l'accent d'une part sur la finalisation de la pondération pour éviter le traitement manuel, bien que léger mais tout de même encore nécessaire et d'autre part sur la construction automatisée du corpus afin d'optimiser la présence des variations et des caractéristiques linguistiques pour une couverture optimale.

RÉFÉRENCES

[1] M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, and S. Sandri. Choose the best to modify the least : A new generation concatenative synthesis system. In *Eurospeech'99*, pages 2291–2294, Budapest, Hungary, 1999.

[2] A. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *Eurospeech'95*, volume I, pages 581–584, Madrid, Spain, 1995.

[3] A. Black and P. Taylor. Chatr , version 0.8, a generic speech synthesis system. In *COLING'94*, volume II, pages 983–986, Kyoto, Japan, 1994.

[4] A. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. In *Eurospeech'97*, pages 601–604, Rhodes, Greece, September 22–25 1997.

[5] A. P. Breen and P. Jackson. Non-uniform unit selection and the similarity metric within bt's laureate tts system. In *ESCA/COCOSDA 3rd Workshop on Speech Synthesis*, pages 201–206, Jenolan Caves, Australia, November 26–29 1998.

[6] N. Campbell and A. Black. *Prosody and Selection of source Units for Concatenative Synthesis*, pages 279–292. New York, Springer-Verlag, 1996.

[7] F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP*, pages 2015–2018, 1986.

[8] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP'96*, pages 373–376, Atlanta, Georgia, 1996.

[9] S. Kullback and R.A. Leibler. *On information and sufficiency*, volume 22 of *Annals of Mathematical Statistics*, pages 79–86. Institute of Mathematical Statistics, University of Connecticut, 1951.

[10] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16 :175–205, 1995.

[11] R. Prudon. *Synthèse de la parole multilocuteur par sélection d'unités acoustiques*. PhD thesis, LIMSI - Université Paris XI, 2003.

[12] R. Prudon and C. d'Alessandro. A selection/concatenation tts synthesis system : Databases developement, system design, comparative evaluation. In *ISCA/IEEE 4th Tutorial and Research Workshop on Speech Synthesis*, pages 201–206, Pitlochry, Schotland, August 29 – September 1 2001.

[13] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *ICASSP'88*, New York City, April 11 – 14 1988.

[14] P. Taylor and A. Black. Speech synthesis by phonological structure matching. In *Eurospeech'99*, pages 1–25, Budapest, Hungary, 1999.