Constitution d'un corpus textuel basée sur la divergence de Kullback-Leibler pour la synthèse par corpus

Aleksandra Krul[1]*, Géraldine Damnati*[1]*,Thierry Moudenc*[1]*, François Yvon*[2]

[1] France Télécom Division R&D, TECH/SSTP
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
{aleksandra.krul,thierry.moudenc,geraldine.damnati}@francetelecom.com

[2] GET/ENST et CNRS/LTCI
46 rue Barrault
75624 Paris Cedex 13
yvon@enst.fr

ABSTRACT

This paper presents a text design method for Text-To-Speech synthesis application. The aim of this method is to build a corpus whose unit distribution is close to a target distribution. As text selection is a NP-hard set covering problem, a greedy algorithm is used. We propose the Kullback-Leibler divergence to compute the score of each candidate sentence. The proposed criterion gives the possibility to control the unit distribution at each step of the algorithm. Finally, we present the first results and we compare the proposed criterion with two standard criteria.

1. INTRODUCTION

La plupart des systèmes de synthèse vocale à partir du texte reposent sur une technique de concaténation d'unités acoustiques pré-enregistrées, la synthèse par corpus étant la plus utilisée. Cette approche repose sur l'utilisation d'une base d'unités acoustiques élémentaires qui résulte de la lecture d'un corpus textuel soigneusement choisi. La qualité du corpus textuel conditionne donc la qualité de la synthèse.

La conception du corpus textuel peut être vue comme un problème de recouvrement d'un ensemble. Chaque phrase du corpus est un ensemble d'unités. L'ensemble cible C contient des unités à couvrir. Le problème consiste à trouver un ensemble de phrases de cardinal minimum dont l'union forme C. Étant donné que le problème est NP-difficile, il n'y a pas d'algorithme exact applicable, d'où le recours à des méthodes heuristiques. L'algorithme glouton est une méthode appropriée pour résoudre ce problème. Il consiste à construire itérativement une solution en ajoutant à chaque pas un élément choisi parmi les autres selon un critère.

Dans le cas de la synthèse de parole, les critères habituellement utilisés découlent de l'objectif principal qui est d'obtenir la couverture des unités. La méthode gloutonne consiste à choisir incrémentalement à partir d'un grand corpus un sous-ensemble de phrases qui atteint la couverture souhaitée. Celle-ci est le pourcentage des unités existantes dans le corpus de départ et présentes dans le corpus construit. Selon les approches, les unités à couvrir sont des diphones, des diphones en contexte, des triphones ou des syllabes. Pour chaque phrase candidate un score est calculé qui permet de choisir la phrase la plus utile, c'est-à-dire celle qui augmente le plus la couverture. La phrase sélectionnée est ensuite retirée du corpus de départ et les unités de celle-ci sont alors enlevées de l'ensemble d'unités à couvrir. De nombreux travaux [8, 2, 6, 3] ont eu recours à l'algorithme glouton pour la constitution du corpus textuel.

D'autres méthodes, inspirées de la méthode gloutonne, ont été proposées notamment la méthode gloutonne inversée (ou cracheuse) [6] et la méthode d'échange par paires [7]. L'algorithme cracheur, à l'inverse de l'algorithme glouton, démarre avec une couverture totale c'est-à-dire celle du corpus de départ. Les phrases sont supprimées une à une jusqu'à ce que la suppression d'une phrase fasse perdre des unités à la couverture totale. Quant à la méthode d'échange par paires, elle vise à améliorer la couverture plutôt qu'à la construire soit en augmentant le nombre d'unités couvertes, soit en diminuant le taux de couverture selon un seuil minimal.

L'efficacité de ces trois méthodes dépend du critère choisi pour calculer le score de chaque phrase candidate. Indépendamment de la méthode utilisée, les critères sont relatifs au nombre d'unités distinctes de la phrase et au nombre d'unités de la couverture. Afin de contrôler la longueur des phrases sélectionnées, le nombre total d'unités dans la phrase est également pris en compte. Dans [6] plusieurs critères ont été présentés et évalués, comme, les critères basés sur le nombre d'unités utiles à la couverture dans les phrases candidates, ou encore la présence d'unités rares dans la phrase.

En fonction de l'objectif à atteindre, le score de chaque phrase candidate peut être calculé de différente manière. Pour atteindre la couverture, le calcul du score le plus simple peut consister à normaliser le nombre d'unités nouvelles d'une phrase par le nombre total d'unités contenues dans cette phrase. Si, en plus d'obtenir une couverture, l'objectif est de favoriser les événements rares alors le calcul du score fait appel aux fréquences des unités observées dans le corpus initial. Dans le but d'obtenir une grande variabilité au niveau phonétique, [3] propose de calculer le score de chaque unité (diphone) de la phrase candidate en fonction de ses contextes phonétiques gauche et droit. Les unités retenues sont celles qui augmentent la variabilité phonétique du corpus. Cette approche permet d'obtenir une meilleure variabilité de triphones dans la base acoustique. L'objectif qui consiste à atteindre la couverture et celui qui vise la variabilité des unités étant partiellement antagonistes, un compromis entre les deux est nécessaire. Cela conduit à des scores différents.

Une des difficultés de la constitution du corpus textuel se trouve dans le choix d'une couverture optimale des unités au sens d'un objectif applicatif. Dans le cas de la synthèse générale, la distribution des unités souhaitée est celle

qui limite la redondance des unités fréquentes et maximise la présence des unités rares. La couverture totale doit être atteinte au moins pour les unités élémentaires. De plus, une représentation suffisante des unités doit être assurée pour anticiper les différents contextes d'apparition possible de ces unités. Pour la synthèse dédiée à des domaines restreints, la distribution idéale des unités est celle qui reflète un contexte applicatif particulier. Dans ce cas, la contrainte sur les unités rares peut être moins forte. En revanche, les unités les plus fréquentes relativement au domaine doivent être bien représentées. La base peut également être plus petite et plus spécifique au domaine visé.

Pour les critères qui cherchent à obtenir une couverture, la distribution des unités dans le corpus final est difficile à maîtriser. C'est pourquoi nous proposons un critère qui vise à contrôler globalement la distribution des unités dans le corpus construit à chaque étape du processus.

Dans cet article nous proposons une méthode gloutonne de constitution de corpus textuel qui repose sur la divergence de Kullback-Leibler. Cette approche vise à construire un corpus dont la distribution des unités tend vers une distribution *a priori*. Le critère utilisé évalue l'utilité d'une phrase en fonction de toutes les unités du corpus construit. Ce critère permet également de maîtriser globalement la distribution des unités dans le corpus. Pour cette étude, la distribution visée est uniforme et l'unité considérée est le diphone. Dans la section 2 nous introduisons la mesure de Kullback-Leibler et nous détaillons notre approche. Enfin, nous présentons les premiers résultats obtenus avec cette méthode et nous la comparons aux méthodes standard qui visent la couverture des unités.

2. APPROCHE ALTERNATIVE

2.1. *La mesure de Kullback-Leibler*

Avant de détailler notre approche, nous introduisons ici la divergence de Kullback-Leibler (KL). C'est une mesure de similarité entre deux distributions de probabilité P et Q. Elle se calcule de la façon suivante :

$$D(P \parallel Q) = \sum_{i=1}^{t} p_i \log \frac{p_i}{q_i} \qquad (1)$$

La divergence est d'autant plus petite que les distributions sont proches. La divergence de KL est toujours positive, elle est nulle si et seulement si les deux distributions sont identiques [4].

2.2. *Sélection de phrases basée sur la divergence de KL*

Nous proposons d'utiliser la divergence de KL pour calculer le score d'une phrase dans le processus de sélection de corpus. Le but étant de construire une distribution *a priori*.

Algorithme L'algorithme utilisé est de type glouton. À chaque itération la phrase retenue est celle qui minimise la divergence de KL à la distribution cible. Notons par $S = \{s_1, s_2, \ldots, s_l\}$ le corpus à partir duquel les phrases sont sélectionnées. Nous représentons celles-ci par $S' = \{s'_1, s'_2, \ldots, s'_m\}$, où $m \leq l$.

La distribution *a priori* des probabilités est donnée par Q.

L'estimation des probabilités des unités dans le corpus construit doit se faire sur l'ensemble des unités sélectionnées à chaque itération de l'algorithme. n_i est le nombre d'occurrences de l'unité i dans le corpus candidat qui englobe le corpus déjà construit à l'itération précédente et la phrase candidate. N est la somme du nombre total d'unités déjà sélectionnées et du nombre total d'unités de la phrase candidate. Dans le corpus construit pas à pas la probabilité pour chaque unité est définie par $p_i = \frac{n_i}{N}$, c'est-à dire par sa fréquence d'apparition. Pour les unités qui ne sont pas encore représentées dans le corpus en construction nous considérons que $p_i = 0$. En utilisant la convention $0 \log \frac{0}{q} = 0$, ces unités ne contribuent pas au calcul de la divergence. Le score de chaque phrase est :

$$D(P \parallel Q) = \sum_{i, n_i \neq 0} \frac{n_i}{N} \big(\log \frac{n_i}{N} - \log q_i \big) \qquad (2)$$

À chaque étape, l'algorithme 1 ajoute à l'ensemble de phrases sélectionnées la phrase qui minimise la divergence de KL. La nouvelle distribution des unités ainsi obtenue se rapproche de la distribution visée. L'algorithme s'arrête après avoir inclus au maximum L phrases, où L est une borne fixée à l'avance.

Algorithme 1 Sélection de phrases basée sur la divergence de KL

Définir une distribution cible Q
$S'_0 = \emptyset$
Pour $j = 1$ jusqu'à L **Faire**
  $D_{min} \leftarrow +\infty$
  Pour Chaque phrase $s_k \in S \setminus S'_{j-1}$ **Faire**
    $A_{jk} = S'_{j-1} \cup \{s_k\}$
    Estimer la distribution des probabilités P_{jk} sur l'ensemble A_{jk}
    Calculer $D(P_{jk} \parallel Q)$
    Si $D(P_{jk} \parallel Q) < D$ **Alors**
      $D_{min} \leftarrow D(P_{jk} \parallel Q)$
      $s_{best} \leftarrow s_k$
    FinSi
  FinPour
  $S'_j \leftarrow S'_{j-1} \cup \{s_{best}\}$
FinPour

Distribution uniforme Nous utilisons comme distribution cible la distribution uniforme où toutes les unités sont équiprobables. Pour cette étude l'unité considérée est le diphone. La distribution des diphones dans le corpus à partir duquel les phrases sont sélectionnées est exponentielle et suit la loi de Zipf [1]. Certaines unités sont très fréquentes, alors que de nombreuses unités sont très peu représentées. En visant la distribution uniforme, le critère tend à mettre au même niveau les unités fréquentes et les unités rares présentes dans le corpus d'origine. Ceci se ramène au choix de la distribution d'entropie maximale.

Couverture En visant la distribution uniforme nous introduisons indirectement l'objectif d'atteindre la couverture totale de diphones. En effet, le critère proposé va chercher à prendre toutes les unités distinctes. Cependant, dans la mesure où l'algorithme sélectionne les phrases entières, la distribution obtenue hérite naturellement de la distribution de départ. Nous pouvons ainsi nous attendre à ce

que la méthode proposée ne permette pas d'atteindre rapidement la couverture. Afin d'obtenir la couverture nous avons développé une variante de l'algorithme 1 en ajoutant une contrainte sur l'ensemble des phrases candidates. Tant que la couverture n'est pas atteinte, la phrase retenue est celle qui minimise le critère de KL parmi les phrases qui apportent de nouveaux diphones distincts. L'algorithme reprend son fonctionnement normal une fois la couverture atteinte.

3. RÉSULTATS EXPÉRIMENTAUX

3.1. Données

Le corpus sur lequel nous avons travaillé contient 7337 phrases issues principalement des articles du journal Le Monde. Il contient également une centaine de phrases utilisées pour des services vocaux. Il a été constitué pour l'enregistrement de la base acoustique. L'objectif de constitution de ce corpus était de couvrir 100% de diphones distincts, 90% de diphones en contextes et 80% de triphones observés dans le corpus général de Le Monde. La longueur maximale de phrases est de 27 mots. Il y a 1170 de diphones distincts dans ce corpus. L'intêret d'extraire des phrases de ce corpus, *a priori* déjà équilibré, est d'observer le comportement des algorithmes dans une optique de réduction de corpus. Dans la mesure où nous n'observons pour l'instant que les diphones distincts, la taille de ce corpus est suffisante pour cette étude.

3.2. Comparaison des trois critères

Nous comparons notre critère avec deux critères standard différents. Le premier score est basé sur le nombre de diphones nouveaux présents dans la phrase candidate normalisé par la longueur de la phrase [6]. Le second score est dérivé du premier, mais favorise la sélection des unités rares en pondérant la contribution de chaque unité nouvelle par l'inverse de sa fréquence dans le corpus d'origine [5].

Nous examinons l'évolution de la couverture en fonction de l'itération sur la figure 1 et en fonction du nombre total d'unités sélectionnées sur la figure 2.

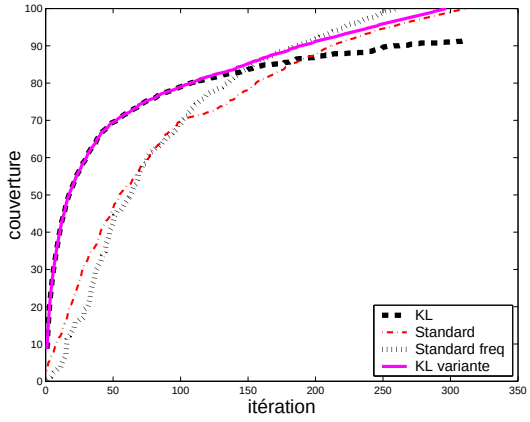

FIG. 1: Couverture

La couverture de diphones est atteinte très rapidement par les méthodes standard et par la variante de notre algorithme. En revanche, elle n'est atteinte qu'à la fin du processus par l'algorithme 1. Ce phénomène a été observé par [5]. Nous l'expliquons par le fait que la méthode KL préfère sélectionner les phrases qui rendent la distribution des unités dans le corpus plus plate plutôt que les phrases avec des unités nouvelles qui déséquilibrent la distribution. Ainsi, les phrases qui contiennent les unités non-couvertes ne sont pas sélectionnées parce qu'elles ne diminuent pas la divergence de KL. Au début, la méthode proposée choisit les phrases longues qui apportent beaucoup d'unités nouvelles. À la fin du processus, l'algorithme sélectionne des phrases courtes et qui ne contiennent pas d'unités nouvelles. Grâce à la modification de l'algorithme la couverture est atteinte assez rapidement. Il est intéressant d'observer que les couvertures obtenues par notre algorithme et sa variante sont identiques au début du processus.

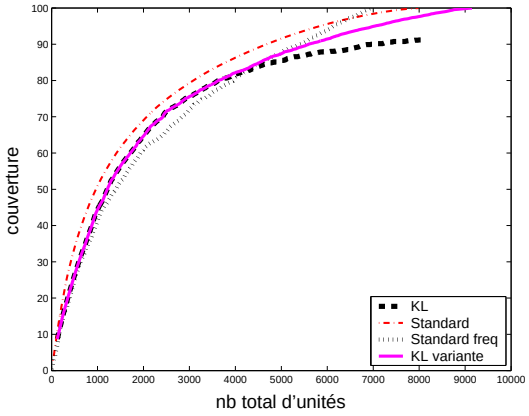

FIG. 2: Couverture

Pour la variante de l'algorithme basé sur la divergence de KL la couverture est atteinte avec un nombre d'unités le plus élevé. Ceci est dû au fait qu'aucun contrôle sur la longueur de phrases n'est effectué.

Pour examiner le comportement de l'algorithme 1 nous l'avons éxécuté jusqu'à ce qu'il n'y ait plus de phrases à choisir : les 7337 ont été sélectionnées. De même, nous avons lancé le processus de sélection de toutes les phrases avec les deux autres algorithmes et calculé ensuite la divergence de KL. La figure 3 illustre cette mesure.

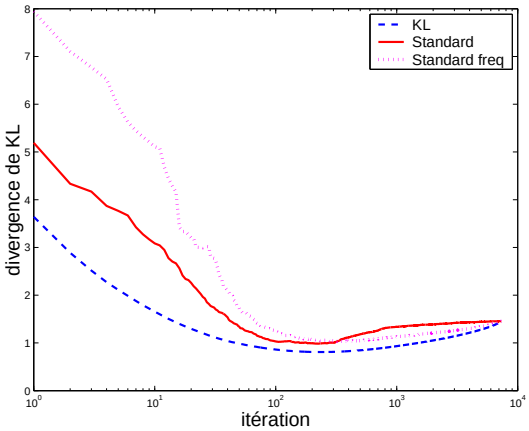

FIG. 3: Divergence de KL

L'allure des courbes de la divergence de KL est similaire pour les trois approches. La distribution des probabilités

des unités qui se rapproche le plus de la distribution uniforme est bien celle qui a été construite avec notre méthode. Pour les trois distributions obtenues la divergence de KL diminue rapidement au début du processus, après avoir atteint un minimum elle commence à croître. Toute nouvelle phrase ajoutée à partir de ce minimum augmente la divergence à la distribution uniforme des unités. Ceci et lié au fait que des phrases entières sont sélectionnées et que la distribution des unités commence à suivre la loi de Zipf.

Il est à noter que le minimum est atteint à différentes étapes des algorithmes. Le tableau 1 présente l'état des corpus en construction à l'itération j pour laquelle la divergence de KL est minimale. N_j est le nombre total d'unités à l'itération j.

TAB. 1: État des corpus en construction pour la divergence de KL minimum.

	itération j	N_j	couverture
KL	239	6414	88,46
Standard	218	4852	90
Standard freq	371	10317	100

Pour le critère standard la divergence remonte assez rapidement et ce avec un nombre total d'unités assez faible. Ceci montre que les phrases sélectionnées par ce critère présentent plus de redondance. Le fait de sélectionner en priorité les unités peu fréquentes (Standard freq) retarde la remontée de la courbe. Malgré le fait qu'à ce stade du processus la couverture est atteinte pour cette méthode, la distribution des unités n'est pas celle qui se rapproche le plus de l'uniforme.

3.3. Remarques

Étant donné que, le corpus sur lequel nous avons travaillé résulte déjà d'une sélection, nous avons utilisé un second corpus. Celui-ci contient des phrases choisies aléatoirement dans le corpus général du journal Le Monde. À partir de 20000 phrases, 10263 phrases dont la longueur maximale est de 20 mots ont été retenues. Ce corpus contient 1117 diphones distincts.

De façon générale, le comportement des trois critères est similaire. Nous observons, par ailleurs, que pour l'obtention de la couverture le nombre d'itérations est plus élevé. De même, la couverture est atteinte avec un nombre total d'unités plus élevé.

4. CONCLUSION ET PERSPECTIVES

Nous avons présenté une méthode alternative pour la construction d'un corpus textuel basée sur la divergence de Kulllback-Leibler. La critère proposé offre la possibilité de contrôler globalement la distribution des unités dans le corpus final. Son objectif principal est de répartir les unités de façon à ce que leur distribution se rapproche d'une distribution cible. Pour cette étude la distribution uniforme des unités est visée. Des distributions adaptées à des domaines spécifiques peuvent être envisagées. L'avantage de cette méthode est qu'elle permet de viser différentes distributions. L'adaptation du corpus textuel à un contexte applicatif peut être facilement réalisée.

De plus, pour atteindre la couverture des unités nous avons implémenté une variante de l'algorithme. Grâce à la modification de l'algorithme la couverture est atteinte et la distribution des unités reste la plus proche de la distribution souhaitée.

Toutefois, pour valider la méthode proposée, des tests sur des corpus de taille plus importante doivent être effectués. Nous envisageons également de travailler sur d'autres types d'unités, par exemple sur des diphones en contexte ou encore sur des triphones.

Enfin, des évaluations des critères sont envisagées d'un point de vue applicatif en comparant la qualité de la synthèse obtenue avec ces critères.

RÉFÉRENCES

[1] R.H. Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, 2001.

[2] A. W. Black and K. A. Lenzo. Optimal Data Selection for Unit Selection Synthesis. In 4^{rd} *ESCA Workshop on Speech Synthesis*, Scotland, 2001.

[3] B. Boozkurt, O. Ozturk, and T. Dutoit. Text design for TTS speech corpus building using a modified greedy selection. In 8^{th} *European Conference on Speech Communication and Technology (Eurospeech)*, pages 277–280, Geneva, Switzerland, September 2003.

[4] T.M Cover and J.A Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[5] Y. Feng. Selection of text script for text-to-speech synthsesis. In 5^{th} *IASTED International Conference SIGNAL AND IMAGE PROCESSSING*, Honolulu, Hawaii, USA, August 2003.

[6] H. François. *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*. PhD thesis, Université de Rennes 1, 2002.

[7] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu. A Design Method of Speech Corpus for Text-to-Speech Synthesis Taking Account of Prosody. In 6^{th} *International Conference on Spoken Language Processing (ICSLP)*, pages 277–280, Beijing, China, September 2000.

[8] J.P.H van Santen and A. L. Buchsbaum. Methods for Optimal Text Selection. In 5^{th} *European Conference on Speech Communication and Technology (Eurospeech)*, pages 553–556, Rhodes, Greece, September 1997.