Amélioration automatique de l'intelligibilité de la parole

Vincent Colotte, Yves Laprie

Amélioration automatique de l'intelligibilité de la parole

Vincent Colotte et Yves Laprie

LORIA
BP 239 - Campus scientifique - 54506 Vandœuvre-lès-Nancy, FRANCE
Mél: Vincent.Colotte@loria.fr , Yves.Laprie@loria.fr

Résumé

This paper presents a speech signal transformation which slows down speech signals selectively and enhances some important acoustic cues. This transformation can be used not only for hearing aids but also for second language acquisition by facilitating oral comprehension.
The strategy used to control slowing down exploits a spectral variation function which locates rapid spectral changes. The enhancement simply consists of amplifying stop bursts and unvoiced fricatives. These acoustic cues are detected automatically through the examination of energy criteria.
This approach was evaluated in the context of second language acquisition. Experiments show that the oral comprehension is improved.

1. Introduction

Le papier présente une méthode de transformation de la parole permettant de ralentir sélectivement son débit et de renforcer les évènements transitoires, pour améliorer l'intelligibilité. En effet, le ralentissement peut améliorer la compréhension orale lors de l'apprentissage de langues étrangères ou compenser la perte de sélectivité temporelle chez les personnes âgées notamment. Le ralentissement peut être appliqué globalement ou sélectivement sur certaines parties du signal (i.e. sur certains indices acoustiques). La dernière solution a l'avantage de ne pas allonger exagérément la durée du signal. De plus, il est possible de combiner le ralentissement par un renforcement des évènements transitoires dans le but d'améliorer la perception de certains indices acoustiques. De même que le ralentissement et le renforcement favorisent la discrimination des sons entendus, ces techniques rendent le signal plus robuste aux dégradations possibles (bruit, canal de télécommunications...).
Différentes méthodes en synthèse de la parole peuvent être mises en œuvre pour effectuer le ralentissement et le renforcement. Nous travaillons dans le contexte de la modification d'un signal de parole avec TD-PSOLA (*Time Domain Pitch Synchronous Overlap-Add*) [MC90] qui est une méthode bien connue en synthèse de la parole pour son faible coût calculatoire. PSOLA repose sur une décomposition du signal temporel en fenêtres recouvrantes synchronisées sur la fréquence fondamentale. Si cette méthode nécessite peu de calculs, elle nécessite en revanche la connaissance des marques de pitch indiquant le centre des fenêtres. L'algorithme de marquage du pitch que nous proposons exploite les résultats de l'algorithme d'extraction de la fréquence fondamentale. Nous effectuons la propagation des marques de période en période à l'aide d'un algorithme de programmation dynamique ; ainsi le marquage est optimal sur l'ensemble du signal. De plus, cette méthode a l'avantage d'être automatique et indépendante de l'algorithme d'extraction de la fréquence fondamentale (pour plus de détails, voir [LC98]).
La stratégie de contrôle du débit dépend du niveau auquel on veut modifier la parole : au niveau prosodique le ralentissement permet de renforcer l'intelligibilité d'un groupe de mots ou d'une partie de la phrase alors qu'au niveau phonétique le ralentissement porte sur des sons précis pour en améliorer l'intelligibilité, et ainsi améliorer l'intelligibilité de la phrase entière. A. Nakamura et al. [NSI+96] ont travaillé de manière à permettre aux personnes âgées de mieux comprendre les bulletins d'information télévisés. L'amélioration de l'intelligibilité est basée sur la modification des "groupes de souffle", et intervient donc au niveau prosodique. Il a été montré, pour les bulletins d'information en japonais, que la mélodie (en l'occurrence la fréquence fondamentale) suit une ligne de déclinaison sur les groupes de souffle et que l'information principale se situe surtout dans les parties à pitch élevé, c'est-à-dire au début des groupes. C'est pourquoi le signal est ralenti pendant la première partie des groupes de souffle puis accéléré sur la fin. Les silences entre groupes de souffle sont eux aussi accélérés. Dans les deux cas, l'accélération est motivée par le souci de ne pas trop allonger la durée du signal, l'opération s'effectuant en temps réel. V. Hazan et A. Simpson [HS98], quant à eux, travaillent au niveau phonétique ; ils recherchent, sur le signal, les régions à forte densité d'indices acoustiques puis renforcent ces régions, principalement par filtrage, pour améliorer l'intelligibilité. Ces régions indiquent les transitions du signal, c'est-à-dire les endroits où les caractéristiques acoustiques du signal changent très rapidement. Ces régions correspondent aux *landmarks* introduites par S.A. Liu [Liu96] pour repérer les régions de transition (ouverture/fermeture de la glotte, début/fin d'un burst, d'une nasale...). S.A. Liu assimile les changements de caractéristiques acoustiques aux variations de l'énergie du signal sur plusieurs bandes de fréquence : à partir de connaissances expertes, elle associe un type de variation acoustique à un type d'évènement articulatoire.
Dans la première partie de ce papier, nous présentons

la mise en œuvre et la stratégie de contrôle du débit à partir du calcul d'une fonction d'évaluation des variations spectrales. Dans une deuxième partie, nous décrivons la stratégie de renforcement qui repose sur la détection de bursts et de fricatives sourdes. Nous conclurons sur les résultats obtenus et les perspectives envisagées.

2. Ralentissement

2.1. Mise en œuvre du marquage du pitch et du ralentissement

TD-PSOLA est une technique qui permet de modifier facilement le débit de la parole et le contour de la F0 ; son principal avantage est son faible coût calculatoire. Contrairement à la synthèse à partir du texte où le problème est de concaténer de courtes unités de parole, notre objectif est de modifier la totalité de la phrase. TD-PSOLA est facile à mettre en œuvre du moment que la localisation des marques du pitch, qui décomposent le signal en fenêtres recouvrantes synchronisées sur la fréquence fondamentale, est réalisée. Comme il existe un grand nombre d'algorithmes de détection de la fréquence fondamentale qui deviennent de plus en plus robustes, il semble intéressant d'exploiter les résultats de l'extraction du pitch pour déterminer les marques du pitch. Une fois les marques du pitch connues, la mise en œuvre du ralentissement devient simple.

Déterminer les marques du pitch à partir de la connaissance des résultats de la fréquence fondamentale consiste à sélectionner des extrema du signal séparés d'un intervalle correspondant à la période du pitch locale. Les marques sont choisies parmi les extrema locaux (minima ou maxima suivant le meilleur coût global). La programmation dynamique permet de propager les marques sur l'ensemble de la phrase de manière optimale. Le marquage du pitch s'effectue en deux étapes.
La première étape consiste à construire un ensemble de candidats à partir des extrema locaux. Les candidats sont recherchés sur des fenêtres espacées régulièrement. Pour être assuré que toutes les périodes de pitch ont un candidat au moins, la fenêtre de recherche doit être plus petite que la plus petite période de pitch sur toute la phrase à modifier.
La deuxième étape consiste à choisir, parmi les candidats, un ensemble, d'extrema séparés par une période de pitch. Soit $C = [c(i)] = c(1) \dots c(i) \dots c(N)$ l'ensemble des candidats où $c(i)$ est l'instant du i^{eme} extremum. Nous devons déterminer une fonction de sélection j donnée par $J = [j(k)] = j(1) \dots j(k) \dots j(K)$ avec $K < N$ et telle que $j(k) < j(k+1)$ (pour préserver l'ordre chronologique). Une solution pour le marquage du pitch est alors $\overline{C} = [c(j(k))] = c(j(1)) \dots c(j(k)) \dots c(j(K))$. Pour trouver la meilleure fonction de sélection nous devons définir un critère qui exprime la qualité des marques du pitch. Nous avons choisi le critère local suivant pour deux marques consécutives :

$$d(c(i),c(l)) = \begin{array}{l} |(c(l) - c(i)) - pitchPeriod(c(i))| \\ -\alpha \times |amplitude(c(i))| \end{array} \quad (1)$$

Le premier terme exprime le fait que la distance entre les marques l et i est approximativement égale à la période de pitch ; le second terme favorise les forts extrema et α représente le compromis entre les deux termes. La recherche de l'ensemble des marques revient donc à trouver K_{opt} et j_{opt} qui minimisent $D = \sum_{k=1}^{K-1} d(c(j(k)),c(j(k+1)))$. Ce problème est résolu par programmation dynamique (voir [LC98]) et les résultats obtenus sont très bons. D'un point de vue pratique, le marquage du pitch est effectué pour les minima et les maxima séparément et la meilleure solution est conservée.
Modifier le débit de la parole revient à dupliquer les fenêtres centrées aux marques du pitch sans changer le pitch. Pour connaître les fenêtres à dupliquer nous utilisons des marques virtuelles qui représentent les positions de marques pour le taux du débit cible. Soit $c_a(i)$ la i^{eme} marque d'analyse (trouvée par l'algorithme précédemment exposé), et $c_v(i)$ la i^{eme} marque virtuelle, $c_v(i+1)$ est donnée par $c_v(i+1) = c_v(i) + r(i+1) \times (c_a(i+1) - c_a(i))$ où $r(i+1)$ est le taux de modification du débit appliqué entre l'instant $c_a(i)$ et $c_a(i+1)$.

2.2. Stratégie de contrôle du ralentissement

L'approche la plus pertinente pour une aide aux malentendants ou pour l'apprentissage des langues étrangères semble se situer au niveau phonétique contrairement à [NSI+96] qui opère au niveau prosodique. En effet la méthode proposée par A. Nakamura et al. [NSI+96] ne semble pas transposable au français. Bien que la ligne de déclinaison existe aussi en français, l'importance de l'information apportée par les groupes de souffle ne suit pas la déclinaison de la fréquence fondamentale comme cela semble être le cas dans les bulletins d'information télévisés japonais. De plus, les modifications au niveau phonétique permettent de focaliser le ralentissement sur des évènements particuliers : seulement un type de son, difficile à percevoir, peut être transformé. La stratégie de ralentissement doit donner lieu à un algorithme automatique et doit pouvoir être piloté par la détection de certains indices acoustiques.

Deux directions sont alors possibles. Soit le signal est segmenté selon les frontières phonétiques de la phrase énoncée - il s'agit donc d'une segmentation en sons -, soit le signal est marqué aux instants[1] à forte densité d'indices acoustiques, c'est-à-dire les régions où les caractéristiques acoustiques du signal varient fortement et rapidement.
Une segmentation phonétique stricte peut être obtenue implicitement lors de la reconnaissance automatique de la parole ou à partir d'une transcription phonétique manuelle : ces deux solutions ont été écartées, la première pour sa lourdeur en calcul et sa précision, et la dernière, bien qu'envisageable dans des exercices de compréhension orale lors d'apprentissage de langue, car nous voulons mettre en œuvre une méthode totalement automatique.
La seconde approche repose sur la localisation des ré-

1. Une région, ici, est définie par un intervalle centré sur un instant. La région est localisée par cet instant et les modifications sont effectuées sur son voisinage.

gions à forte densité d'indices acoustiques. Contrairement à Liu [Liu96] qui utilise un critère sur l'énergie et une connaissance experte sur les caractéristiques des évènements à localisés, nous utilisons une méthode qui évalue les variations acoustiques de la parole. Cette méthode, appelée *Spectral Variation Function*, proposée G. Flammia et al. [FDAL92] et F. Brugnara et al. [BMGO92] utilise une analyse mel-cepstre. Un coefficient, reflétant le taux de variation du spectre, est calculé pour chaque fenêtre (de 20 ms toutes les 10 ms) par rapport aux fenêtres voisines. Le recherche des maxima locaux de la fonction SVF nous donne les instants de forte variation des caractéristiques acoustiques. Nous avons retenu cette méthode qui indique les régions à forte variation acoustique parce qu'elle permet de localiser 82% des frontières de sons placées par un expert. Les 18% restant sont soit des marques mal placées (à plus de 20 ms) soit des insertions. Ce deuxième type d'erreurs n'est pas trop gênant voire utile ; en effet, étant donné le but à atteindre, à savoir ralentir le signal pour une meilleur intelligibilité, le fait de prendre plus de marques et de ralentir la parole au voisinage de ces marques ne risque pas de diminuer l'intelligibilité, d'autant que les marques sont proches et donnent lieu de fait à un seul ralentissement sur une région englobant ces marques. Le choix de la valeur du facteur de ralentissement est arbitraire, mais une valeur trop forte (supérieure à 3) dénature le son par rapport à l'articulation habituelle (en particulier, les bursts peuvent être artificiellement transformés en fricatives). Nous avons donc retenu une valeur de 1.8 à 2 pour ce facteur. Même avec cette valeur de ralentissement plutôt élevée, l'allongement moyen global (sur toute la phrase) est seulement de 1.3.

3. Renforcement

La recherche des segments à renforcer s'est principalement focalisée sur les occlusives et les fricatives. D'après V. Hazan et A. Simpson [HS98], le renforcement de ce type de phonème permet d'améliorer l'intelligibilité en parole spontanée. Nous avons décidé de nous focaliser sur les bursts et fricatives sourdes car ces sons peuvent être localisés avec un fort degré de robustesse et leur renforcement améliore la perception de la structure temporelle de la parole.
Nous allons expliquer comment les fricatives et les bursts sont détectés à partir d'un critère basé sur l'énergie. Différencier une fricative d'un autre son peut être facilement réalisé à partir de l'énergie. En première approximation, l'énergie d'une fricative est principalement localisée en haute fréquence. Nous avons donc choisi de calculer le rapport de l'énergie dans la bande $3600 - 6000$ Hz sur l'énergie dans la bande $600 - 1000$ Hz, les autres fréquences étant considérées comme moins pertinentes. Les fricatives sourdes correspondent à une valeur élevée de ce rapport.
Les occlusives se caractérisent par l'absence d'énergie pendant l'occlusion : cela se répercute par une faible moyenne d'énergie au centre du segment de l'occlusion par rapport à l'énergie aux bords. En présence d'une occlusion, la moyenne au centre est plus faible que la moyenne globale. Le second indice, qui différencie un burst d'un début de voisement, par exemple, est la dérivée de l'énergie. Un burst présente un pic dû à la forte variation du spectre au moment de l'explosion. Un seuil est utilisé pour sélectionner les pics significatifs (50% du maximum de l'amplitude de la dérivée). Fig. 1 résume la stratégie utilisée pour détecter les bursts et les fricatives.

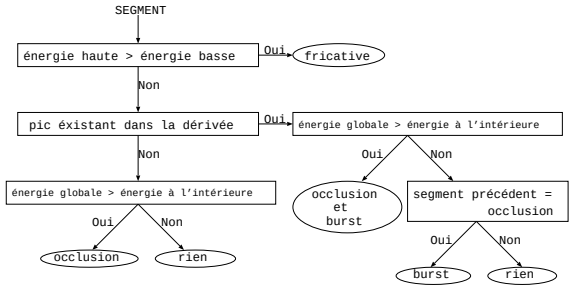


FIG. 1 – *Algorithme simplifié de détection des occlusives et des fricatives*

La stratégie de renforcement de l'énergie du signal est basée sur les expériences de V. Hazan et A. Simpson [HS98]. Le principe est d'amplifier progressivement les transitions des bursts et des fricatives (dans le domaine temporel) jusqu'à un niveau voulu et de revenir au niveau initial (voir Fig. 2).

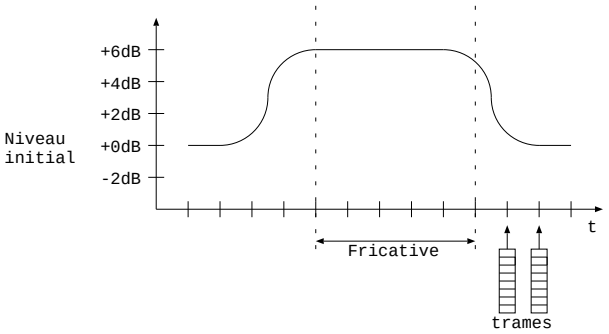


FIG. 2 – *Exemple de renforcement d'une fricative sourde*

4. Expérimentation

Comme mentionnées dans l'introduction, les deux applications possibles sont l'aide aux malentendants et la compréhension orale pour une seconde langue. Tout d'abord, nous avons testé l'amélioration de la compréhension orale (de phrases anglaises par des personnes françaises) car l'ajustement de la stratégie de contrôle du débit pouvait se faire facilement dans ce cas.
Considérant que nos transformations s'apparentent à des études de perception basées sur l'exagération d'indices acoustiques, nous avons décidé d'évaluer l'amélioration au niveau des mots plutôt que sur des unités plus spécifiques (de type VCV par exemple).
Le corpus de test est constitué de 50 phrases sélectionnées dans la base de données TIMIT.

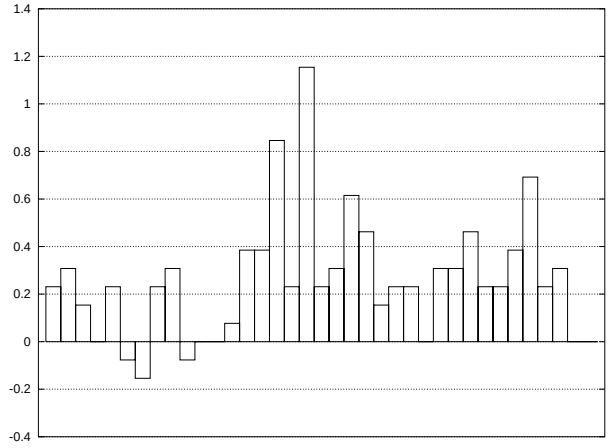
4.1. Evaluation des transformations

La première évaluation porte sur la pertinence du ralentissement sélectif et du renforcement. Chaque technique donne séparément de bons résultats et sont robustes aux erreurs. En effet, les erreurs commises par le marquage SVF sont principalement des insertions et aucun compromis n'est donc fait par rapport à l'in-

telligibilité. De même, les erreurs commises lors de la détection des bursts et des fricatives sont principalement des omissions de faibles bursts ou fricatives, ce qui là encore ne détériore pas l'intelligibilité de la parole, contrairement à des insertions (dues à de fausses détections) qui auraient pu introduire des bursts ou des fricatives artificiels.
Les 50 phrases du corpus de test ont été transformées et évaluées. Nous n'avons trouvé que deux bursts artificiels, dont un masqué par une fricative voisine (ce qui ne change pas la perception du mot) et l'autre perçu comme un clic. Nous avons remarqué une erreur de marquage de pitch sur un début de voisement d'un burst non-voisé [d] qui décale la portée de l'amplification et modifie la perception. Comme attendu, les marques SVF apparaissent dans des régions qui contiennent des variations spectrales rapides (principalement aux transitions formantiques et en bord de nasales). En général, plusieurs marques sont détectées dans les régions correspondantes à de rapides transitions formantiques. Comme ces marques sont très proches les unes des autres, elles donnent lieu à un taux de ralentissement unique et ne perturbent pas la stratégie de contrôle du débit.

4.2. Evaluation perceptive

13 adultes français ont participé à deux sessions d'expérimentation d'une demi-heure. Dans la première, les 50 phrases sont les phrases originales. Dans la seconde session, 25 furent conservées et 25 autres furent modifiées par la transformation exposée précédemment. Le corpus de test a été aléatoirement mélangé et les personnes devaient compléter le ou les deux mots manquants dans la transcription des phrases qu'ils écoutaient. Nous avons considéré quatre niveaux de réponse : **0** aucune réponse n'a été donnée, **1** la réponse donnée n'a rien en commun avec le mot correct, **2** au moins la moitié de phonèmes sont corrects, **3** le mot a été bien reconnu. Tab. 1 donne la différence



TAB. 1 – *Différences des valeurs d'identifications pour les 37 mots cibles*

en moyenne (sur les 13 personnes) pour les 37 mots testés entre leur version modifiée et leur version originale. Un test de Student a montré que les valeurs d'identifications ont été améliorées significativement ($p < 0.02$). Les résultats ont montré par ailleurs que l'amélioration est uniformément distribuée entre les bursts et les fricatives et que le ralentissement des transitions ne change pas la perception des transitions, excepté pour un seul burst dont l'articulation a été modifié et qui a été ainsi perçu comme une fricative.

5. CONCLUSION

La principale force de notre approche est le fait qu'elle repose sur une stratégie simple : renforcement des occlusives et des fricatives sourdes fiables. De plus, les modifications portent sur les régions ou les phonèmes qui ont un effet significatif sur la compréhension : les transitions - régions contenant une forte concentration d'indices acoustiques - et les phonèmes de type occlusives et fricatives. La combinaison du ralentissement sélectif à partir des marques SVF et du renforcement acoustique des bursts et des fricatives améliore l'intelligibilité de la parole. Les résultats se trouvent sur *http://www.loria.fr/∼colotte* [2].
L'avantage de notre approche est qu'elle est totalement automatique ; ainsi, elle peut être facilement combinée avec un système de synthèse de parole visuel dans le but de compléter le signal acoustique avec les mouvements des lèvres pour exploiter l'effet McGurk [MM76].

RÉFÉRENCES

[BMGO92] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo. Improved connected digit recognition using spectral variation functions. In *Proc. of Int. Conf. on Spoken Language Processing 1992*, pages 627–630, Banff, Canada, 1992.

[FDAL92] G. Flammia, P. Dalsgaard, O. Andersen, and B. Lindberg. Segment based variable frame rate speech analysis and recognition using a spectral variation function. In *Proc. of Int. Conf. on Spoken Language Processing 1992*, pages 983–986, Banff, Canada, 1992.

[HS98] V. Hazan and A. Simpson. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24:211–226, 1998.

[LC98] Y. Laprie and V. Colotte. Automatic pitch marking for speech transformations via td-psola. In *IX European Signal Processing Conference*, Rhodes, Greece, 1998.

[Liu96] S.A. Liu. Landmark detection for distinctive feature-based speech recognition. *J. Acoust. Soc. Am.*, 100(5):3417–3430, November 1996.

[MC90] E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for a text-to-speech synthesis using diphones. *Speech Communication*, 9(5,6):453–467, 1990.

[MM76] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 246:745–746, 1976.

[NSI+96] A. Nakamura, N. Seiyama, A. Imai, T. Takagi, and E. Miyasaka. A new approach to compensate degeneration of speech intelligibility for elderly listeners. *IEEE Trans. on Broadcasting*, 42(3):285–293, September 1996.

2. Ils sont susceptibles d'être modifiés en fonction de nos tests perceptifs.