Reconnaissance vocale basée sur les phonèmes voisés

Matthieu Duvinage, Jean-Yves Parfait

Laboratoire TCTS, Université de Mons (UMONS) - MULTITEL
20, Place du Parc, 7000 Mons, Belgique - Parc Scientifique Initialis, 7000 Mons, Belgique
matthieu.duvinage@umons.ac.be - parfait@multitel.be

ABSTRACT

This paper describes a new approach for robust speech recognition. Inspired from mask theory, the Frame Dropping technique consists in using only highly voiced frames in the decoding process. The idea is to be more robust to noise by using more stable parts. This method was assessed on a database of isolated words in additive noise provided by the Signal Processing Information Base (NOISEX-92). The test showed that there is a deterioration in clean environment. However, the results are quite close to the baseline in adverse conditions. Moreover, it paves the way for global reduction of complexity.
Keywords: speech recognition, Frame Dropping, additive noise, multimodal, embedded systems

1. Introduction

La reconnaissance vocale a depuis quelques dizaines d'années été sources de recherches intensives. Les performances ont ainsi évoluées de manières éloquentes. Cependant, les excellentes performances mises en avant le sont dans le cas de signaux peu bruités. En effet, en présence de bruit, les résultats s'écroulent rapidement non seulement à cause de l'ajout d'un autre signal perturbateur mais également dû à l'effet Lombard. Celui-ci est connu pour notamment allonger la durée des phonèmes voisés (et particulièrement les voyelles) ainsi qu'augmenter leur Rapport Signal à Bruit local (RSB) en vue d'une meilleure intelligibilité [3, 4].

De plus, comme indiqué dans [9], étant donné la multiplication des systèmes embarqués dont les caractéristiques de faibles ressources en font des systèmes compliqués à appréhender, on est en droit de se demander s'il n'est pas possible de réduire le nombre de fenêtres à traiter afin d'obtenir un système globalement moins complexe, et ce, en n'utilisant que de l'information hautement fiable. Ceci permettrait également de ne pas utiliser l'information peu fiable pouvant impliquer des mauvais choix locaux du chemin de Viterbi, c'est-à-dire avec une probabilité très faible pour le phonème correct. Ceci peut encore être pire dans le sens où cette décision locale peut faire en sorte que l'algorithme ne sera plus capable de retrouver le chemin correct des suites des méthodes d'élagage.

Ces dernières remarques mettent le doigt sur la difficulté du problème : trouver de l'information hautement fiable. Dans ce cadre, les phonèmes voisés s'avèrent être une bonne option de par leurs caractéristiques de renforcement en milieu bruité. De plus, ils sont caractérisés par une plus grande stabilité et une allure périodique plus facilement détectable dans un environnement fortement bruité, au contraire des phonèmes non-voisés beaucoup plus proche du bruit.

En outre, une approche multimodale est envisagée. La reconnaissance vocale sur des systèmes embarqués confrontés à des milieux fortement bruités peut tenir compte de la possibilité de l'utilisateur d'interagir. L'évaluation du système se fera donc sur base des N-meilleures solutions parmi lesquelles l'utilisateur peut choisir la réponse correcte.

C'est donc dans cette optique que le comportement d'un système de reconnaissance vocale basé sur les fenêtres fortement voisées (technique de Frame Dropping) est étudié. Ce système est inspiré des méthodes de masques décrites dans [6, 7, 8] qui utilisent soit le rapport signal à bruit (RSB) soit le voisement pour réduire l'espace de recherche. Ce papier propose une nouvelle approche différente des méthodes existantes d'utilisation du voisement basées respectivement sur l'incorporation dans le vecteur de caractéristiques et dans le modèle de Markov [10, 5]. Le présent papier s'organise donc en plusieurs sections. La section 2 explique le choix de l'algorithme de voisement ainsi que son principe. La section 3 décrit brièvement comment ce système est implémenté dans un système existant. La section 4 illustre les résultats obtenus. Les conclusions et travaux futurs sont finalement exposés.

2. Algorithme d'estimation de voisement : YIN

Comme il est voulu de n'utiliser que des fenêtres correspondants à des fenêtres hautement fiables, donc hautement voisées selon notre contexte, il est nécessaire d'avoir un estimateur avec un très faible taux de faux positifs en définissant la classe voisée comme positive. Le choix de l'algorithme YIN est guidé par sa robustesse au bruit étudiée préalablement. Comme l'estimateur YIN fournit une valeur continue et non un choix, le choix est réalisé en calibrant le seuil de décision pour des taux de faux positifs de 1%, 5% et 10% par une analyse en courbe de ROC. Ceux-ci ont été obtenus sur une base de données indépendante segmentée en phonèmes à la main par une experte linguiste. Cette base était destinée à la synthèse vocale et a une couverture phonétique bien répartie.

Par définition, un signal périodique de période T_0 est invariant pour chaque décalage temporel T_0 : $x_t - x_{t+T_0} = 0, \forall t$. Etant donné que la période est la valeur T_0 minimale qui vérifie cette équation, si la période est inconnue, la fonction de différence suivante est nulle pour cette période :

$$d_t(\tau = T_0) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2 = 0$$

où W est la moitié de la taille de la fenêtre utilisée dans l'extraction des caractéristiques. Si le signal n'est pas parfaitement périodique, il suffit de déterminer pour quelle période d_t est minimum. Afin de normaliser les valeurs comme proposé dans [2], l'estimateur final pour la fenêtre n est :

$$d^n_{YIN}(\tau) = \begin{cases} 1 & \text{si } \tau = 0 \\ \dfrac{d_t(\tau)}{(\frac{1}{\tau}) \sum_{j=1}^{\tau} d_t(j)} & \text{autrement} \end{cases}$$

3. Frame Dropping

Dans cette section, l'astuce mathématique qui a permis d'utiliser un système de reconnaissance vocale actuel basé sur un algorithme de Viterbi existant pour implémenter le Frame Dropping est décrite.

Afin de ne pas tout réimplémenter, le Frame Dropping est implémenté d'une manière similaire à [1]. D'un point de vue conceptuel, il faut supprimer la fenêtre non fiable et donc, il faut qu'elle n'ait plus de poids dans la décision du chemin de Viterbi, c'est-à-dire de ne plus tenir compte ni de la vraisemblance de cette fenêtre, ni de l'incrément de temps lié à la présence de celle-ci.
D'un point de vue pratique, cette méthode a été implémentée en mettant toutes les valeurs de vraisemblance à une constante pour la fenêtre correspondante. Pour chaque état j du modèle de Markov et à un temps t, la vraisemblance de chaque chemin du Viterbi est calculée en multipliant les probabilités de transition a_{ij} entre les états et la probabilité d'émission b_j selon ce chemin. La vraisemblance partielle $\sigma_{j,t+1}$ peut s'exprimer :

$$\sigma_{j,t+1} = \max_i [\sigma_{i,t} a_{ij}][b_j(o_t)] \quad (1)$$

Grâce à cette définition, comme proposé dans [1], l'astuce est donnée par :

$$\sigma_{j,t+1} = \max_i [\sigma_{i,t} a_{ij}][b_j(o_t)]^{\gamma_t} \quad (2)$$

où γ_t permet de supprimer mathématiquement la fenêtre. Dans notre cas, γ_t est une variable binaire ($\gamma_t = 1$ si la fenêtre est voisée, $\gamma_t = 0$ dans le cas contraire). De plus, vu que les probabilités de transition a_{ij} sont unitaires, l'influence d'une fenêtre non-voisée est annulée lors de l'algorithme de Viterbi.

4. Evaluation des performances

Dans cette section, le système originel est d'abord décrit. Ensuite, la base de données utilisée est définie ainsi que la méthode d'évaluation. Finalement, les performances obtenues sont discutées.

Le système originel est basé sur un système hybride ANN/HMM utilisant pour chaque fenêtre de 30 ms (décalage de 10 ms) des coefficients jRASTA-PLP et fournissant les N-meilleures solutions. Les bruits utilisés viennent de la base NOISEX-92 (obtenu de Signal Processing Information Base[1]) : un bruit stationnaire de voiture à bande étroite, un bruit non-stationnaire de parole à large bande et un autre bruit non-stationnaire courant enregistré dans une usine. Ces bruits ont été ajoutés à la base de données dans des rapports signaux à bruit moyen de 5 à 15 dB.

La base de données est composée de 1611 enregistrements qui couvrent un vocabulaire de 154 mots français dits de manière isolée. Le bruit fut ajouté à des degrés de rapport signal à bruit différents, et donc, la base de données ne présente pas d'effet Lombard. Ceci représente une limitation de l'étude présentée mais voulu afin d'isoler les effets. L'étude de l'effet Lombard pour cette approche se fera prochainement. De plus, aucune détection d'activité de voix ni de techniques d'élagage n'ont été utilisées.
En ce qui concerne la méthode d'évaluation, celle-ci se base sur la possible interaction de l'utilisateur avec le système. Le système peut donc fournir une liste des N-meilleures réponses à la requête et l'utilisateur choisira la bonne solution. Il faut donc que la solution apparaisse dans la liste, d'où l'importance de l'ordre et la méthode adaptée d'évaluation : la précision à N. Celle-ci se calcule par le rapport du nombre de bonnes réponses fournies sur le nombre d'essais, considérant qu'une bonne réponse est définie par la présence du mot demandé en requête dans les N-meilleurs éléments retournés par le système (dans notre cas, la liste des N-meilleures solutions est de 10).

Les résultats obtenus par la méthode développée sont globalement moins bons que le système originel mais des tendances peuvent être dégagées. Tout d'abord, comme montré sur la figure 1, pour un taux de faux positifs plus faible, c'est-à-dire un nombre plus bas de fenêtres utilisées comme montré à la table 1, la performance diminue. Dans un environnement sans bruit, comme les phonèmes non-voisés représentent une information pertinente vu leur taux de reconnaissance élevés, il est logique de voir les performances diminuer. Néanmoins, comme escompté, dans un environnement fortement bruité, l'écart entre les différents cas devient plus faible par l'utilisation d'information plus fiable. Ceci tend à dire qu'en milieu bruité, il vaut mieux utiliser l'information pertinente. Cependant, l'approche proposée ne conclut pas que le masque basé sur le voisement suffit ou est le bon choix.
Par exemple, pour un bruit de parole à un RSB de 5 dB, les performances s'écroulent. Cependant, les résultats des différentes approches sont quasi-identiques et restent au dessus de la chance (précision à 10 de $\frac{10}{154} = 6.49$ %). En ce qui concerne le bruit de voiture, celui n'affecte pas vraiment les performances et n'a pas été présenté. Ceci s'explique peut-être par la grande robustesse des coefficients jRasta-PLP à ce genre de bruits stationnaires et en basses fréquences. De plus, quand N augmente, l'écart se réduit entre la nouvelle approche et le système originel sauf pour

[1] `http://spib.rice.edu/spib/select_noise.html`

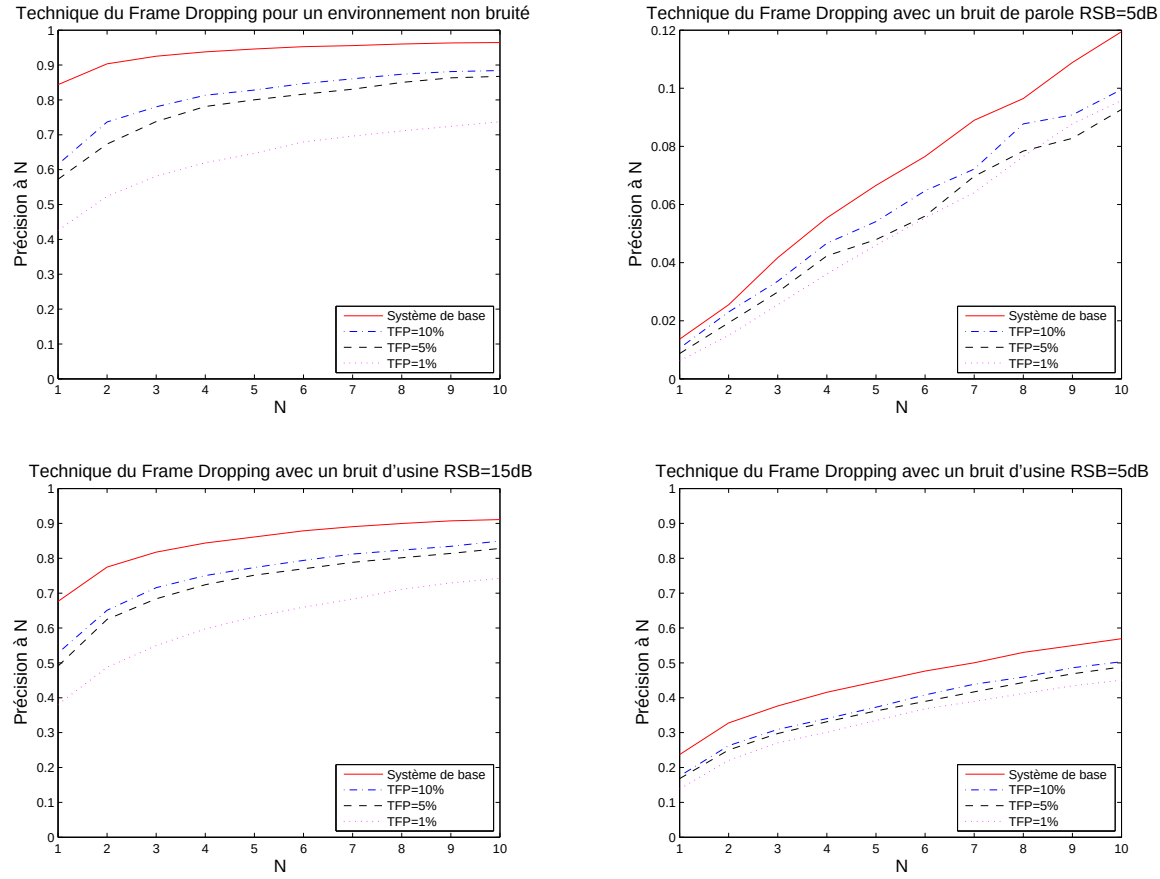

Fig. 1: Au dessus : à gauche, la technique du Frame Dropping amène une diminution substantielle de la précision dans un environnement non bruité. Néanmoins, à droite, la performance s'écroule pour un bruit de parole mais la technique proposée donne des résultats similaires. En dessous : on remarque que les courbes du Frame Dropping se rapprochent du système originel lorsqu'on se situe à des RSB plus bas.

deux cas montrés à la figure 1. Ceci tend à indiquer que la simplification du modèle HMM utilisé augmente la confusion entre mots (deux mots peuvent avoir un modèle plus proche de par le principe de l'approche) pour les N faibles mais permet d'extraire les éléments qui discernent le plus fortement ceux-ci d'où la diminution de l'écart pour les N élevés.
Enfin, la perte de performance peut-être compensée par l'approche multimodale. L'intervention de l'utilisateur peut permettre d'obtenir des résultats similaires à un système basé sur une précision à 1. En effet, on peut voir sur la figure 1 qu'il existe un nombre N_0 du Frame Dropping qui permet d'avoir des performances identiques au système classique ($N = 1$) sauf pour TFP=1% en milieu non bruité.

En ce qui concerne la précision résultante, la robustesse et la complexité du modèle, cette méthode a un bon comportement. En effet, même si la précision globale est affectée par la suppression de fenêtres, le modèle résultant est plus robuste à certaines variabilités dues à la coarticulation, potentiellement à un environnement bruité et aux mauvaises prononciations dans les transitions voisées/non-voisées.
L'effet de coarticulation est moins important car seule la partie la plus stable des phonèmes voisés est prise en compte, c'est-à-dire au milieu des phonèmes. Ceci pourrait laisser entrevoir des performances similaires entre système à contexte et sans contexte, impliquant la possibilité de ne pas devoir calculer les caractéristiques de toutes les fenêtres.
En ce qui concerne le bruit, bien qu'il n'est pas possible de conclure à l'heure actuelle, celui-ci serait moins dérangeant pour deux raisons. L'effet Lombard, bien que non présent ici, augmente l'intelligibilité des phonèmes voisés (en allongeant la longueur et en augmentant le RSB local pour la plupart des phonèmes voisés) en droite ligne avec la procédure proposée. Ce propos doit être nuancé par le désaccord entre modèles et données résultant du décalage en fréquence. D'autres part, le système n'utilise plus les phonèmes déjà semblables au bruit à la base qui provoquaient de nombreuses erreurs dans le réseau de neurones, et par conséquent dans la recherche du chemin de Viterbi. Ce concept permettrait d'être plus robuste aux techniques d'élagages agressifs.
De plus, cette approche supprime aussi les transitions entre éléments voisés et non-voisés non stable ce qui rend le système plus robuste à cette partie.

Finalement, une telle approche diminue la complexité du décodage en diminuant le nombre de fenêtres comme montré à la table 1. De surcroit, comme déjà dit, une étude plus profonde pourrait peut-être montrer que l'utilisation de contexte n'est plus nécessaire dans cette approche et d'où la possibilité de ne cal-

Tab. 1: Le nombre de fenêtres utilisées augmente avec le taux de faux positifs (TFP). Les enregistrements ont une durée fixe de deux secondes sans détection d'activité de voix.

Frame Dropping	TFP=1 %	TFP=5 %	TFP=10 %
Rapport fenêtres utilisées/totales de parole	28 %	48.5 %	61.4 %
Rapport fenêtres utilisées/totales	11 %	19 %	24.25 %

culer les caractéristiques des fenêtres que pour les fenêtres suffisamment voisées. Comme proposé dans [2], l'algorithme YIN peut être optimisé au niveau de la complexité. En outre, d'après l'analyse des résultats, il pourrait être envisagé de ne plus utiliser de détection d'activité de voix. Tout ceci consiste en des pistes afin d'obtenir un système globalement moins complexe, c'est-à-dire ne compensant pas qu'uniquement le surcout de calcul de l'algorithme YIN.

5. Conclusions

Ce papier a étudié une nouvelle approche de Frame Dropping basée sur l'utilisation d'information hautement fiable. Dans ce papier, l'information fiable a été définie comme une fenêtre hautement voisée. Ce concept se base sur sa structure plus facilement reconnaissable dans des milieux fortement bruités et sur l'effet Lombard vis-à-vis des phonèmes voisés.
Les résultats ont montré que cette technique amène une dégradation dans les performances. Cependant, l'écart diminue par rapport au système originel avec un milieu de plus en plus bruité. Ceci tend à confirmer que l'utilisation d'un masque efficace avec de l'information hautement fiable permet d'amener le plus d'information utile dans cet environnement. Le Frame Dropping est aussi plus robuste face aux variabilités telles que la coarticulation et les transitions de phonèmes voisés/non-voisés.
De plus, par l'interaction de l'utilisateur, il est possible d'obtenir des résultats similaires à la méthode originelle à des RSB élevés.
En outre, cette approche possède un potentiel élevé de diminution de complexité, notamment par la possibilité d'utiliser des réseaux de neurones sans contexte, de ne pas calculer les coefficients pour toutes les fenêtres et de ne plus devoir utiliser une détection d'activité de voix poussée.
Finalement, cette approche serait logiquement plus robuste à un élagage agressif.

6. Travaux futurs

D'abord, en ce qui concerne les travaux futurs, il est évidemment indispensable d'étudier cette approche sur une base de données où l'effet Lombard est présent ainsi qu'un élagage agressif afin d'évaluer le réel potentiel au niveau performance. De plus, d'autres informations comme le RSB pourrait être incorporées dans la décision de fiabilité. Cette décision pourrait de surcroit être continue et non binaire comme le permet le cadre théorique développé.
Ensuite, une étude plus précise au niveau de la complexité du système obtenu en étudiant toutes les voies de simplifications évoquées semblent aussi être une voie de recherche pour le futur.
Finalement, vu le comportement au niveau du nombre de fenêtres utilisées, il est en droit de se demander si l'utilisation d'une détection de voisement ne serait pas utile dans une détection d'activité de voix.

Références

[1] Alexis Bernard and Abeer Alwan. Joint channel decoding - viterbi recognition for wireless applications. In *in Proceedings of Eurospeech*, pages 2703–6, 2001.

[2] Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *JASA : Journal of the Acoustical Society of America*, 111 :1917–1930, 2002.

[3] R. Hajislam, Y. Anglade, J.-C. Junqua, and J.-M. Pierrel. Etude acoustique du réflexe Lombard en vue de la reconnaissance de la parole produite en milieu bruité. *Journal de Physique IV*, 04(C5) :C5–485–C5–488, 1994.

[4] John H. L. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Commun.*, 20(1-2) :151–173, 1996.

[5] Peter Jančovič and Münevver Köküer. Incorporating the voicing information into hmm-based automatic speech recognition in noisy environments. *Speech Commun.*, 51(5) :438–451, 2009.

[6] Christopher Kermorvant, Christopher Kermorvant, Andrew Morris, and Andrew Morris. A comparison of two strategies for asr in additive noise : Missing data and spectral subtraction, 1999.

[7] D. O'Shaughnessy and H. Tolba. Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision. In *ICASSP '99 : Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*, pages 413–416, Washington, DC, USA, 1999. IEEE Computer Society.

[8] Michael L. Seltzer, Bhiksha Raj, and Richard M. Stern. Classifier-based mask estimation for missing feature methods of robust speech recognition. In *Proc. ICSLP*, pages 538–541, 2000.

[9] Tan Zheng-Hua and Lindberg Børge. *Automatic Speech Recognition on Mobile Devices and over Communication Networks (Advances in Pattern Recognition)*, chapter 1, pages 1–21. Springer, 2008.

[10] András Zolnay, Ralf Schlüter, Ralf Schl Uter, and Hermann Ney. Extraction methods of voicing feature for robust speech recognition. In *Proceedings of Eurospeech*, pages 497–500, 2003.