Compte rendu de lecture d'article

Reconnaissance Vocale et Indexation Multilingue
Compte rendu de lecture d'article

Extraction et représentation des caractéristiques prosodiques pour la reconnaissance de la langue du locuteur

Boutheina WATHEK Armel Sitou AFANOU

15 février 2013

Résumé de l'article

La prosodie désigne la mélodie, le rythme et l'intensité de la voix. La fréquence fondamentale F0 et l'énergie, mesures de la mélodie et de l'intensité, sont les paramètres les plus couramment utilisés. Dans cet article une nouvelle approche a été proposée pour extraire et représenter ces caractéristiques à partir du signal de parole pour des applications telles que et la reconnaissance de la langue et du locuteur. En effet l'hypothèse sous-jacente est que la prosodie est liée à des unités linguistiques comme les syllabes,d'où la necessité d'une segmentation syllabique. Cela se réalise en détectant automatiquement les points d'apparition de voyelles (VOP) grâce à l'enveloppe de Hilbert de prédiction linéaire du signal de parole. La région entre deux VOP successifs est ensuite considérée comme une région syllabique. La connaissance des VOPs sert de référence pour extraire des paramètres prosodiques : fréquence fondamentale F0,la durée, et les variations d'énergie correspondant à chaque région syllabique. L'efficacité de la méthode "prosodique" est ensuite expérimentée avec succès dans le cas du NIST LRE 2003 et du NIST SRE 2003, respectivement pour la reconnaissance du langage, et du locuteur.

Techniques et résultats présentés

– **Détection des points d'apparition des voyelles (VOP)**

Pour l'extraction des VOPs ,on utilise une technique basée sur l'utilisation de l'enveloppe de Hilbert du signal résiduel de prédiction linéaire (LP). Cet enveloppe montre un changement significatif dans la transition d'une consonne à une voyelle..En pratique,l'enveloppe de Hilbert est définie comme : $h_e = \sqrt{r^2(n) + r_h^2(n)}$ où r(n) est le résiduel de prédiction linéaire du signal de parole, et $r_h(n)$ est la transformée de Hilbert de r(n)

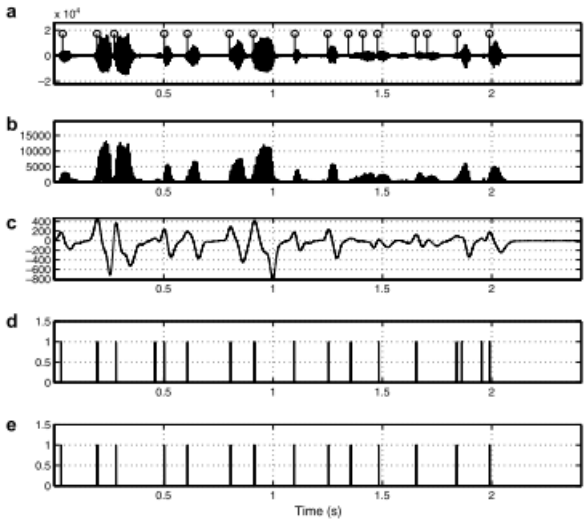

FIGURE 0.1: Détection des VOP

La figure présentée ci-dessus montre un signal de parole avec les VOP marqués à la main, l'enveloppe de Hilbert , une courbe de marquage de VOP, la sortie de l'algorithme de détection des pics, et les VOP trouvés. La courbe de marquage de VOP est obtenue à partir de l'enveloppe de Hilbert, multipliée par le filtre de Gabor, et en prenant la somme du produit pour chaque échantillon. A partir du tracé de la courbe de marquage de VOP, les sommets sont détectés à l'aide d'un algorithme de detection des pics, puis les pics parasites sont ensuite éliminés.

– **Extraction des caractéristiques prosodiques pour la reconnaissance de la langue**

L'expérimentation est réalisée sur un classifieur basé sur un réseau de neurones multi-couches entrainé sur une base de test de 500 éléments.Les vecteurs de tests sont formés de 21 éléments répartis sur les caractéristiques prosodiques de 3 syllabes consécutives (distance entre VOPs successifs, durée du voisement (DV), variation de F0 (DF0), la distance du sommet de F0 par rapport au VOP, déclinaison de l'amplitude (At), déclinaison de la durée (Dt) et la variation de l'énergie (DE))

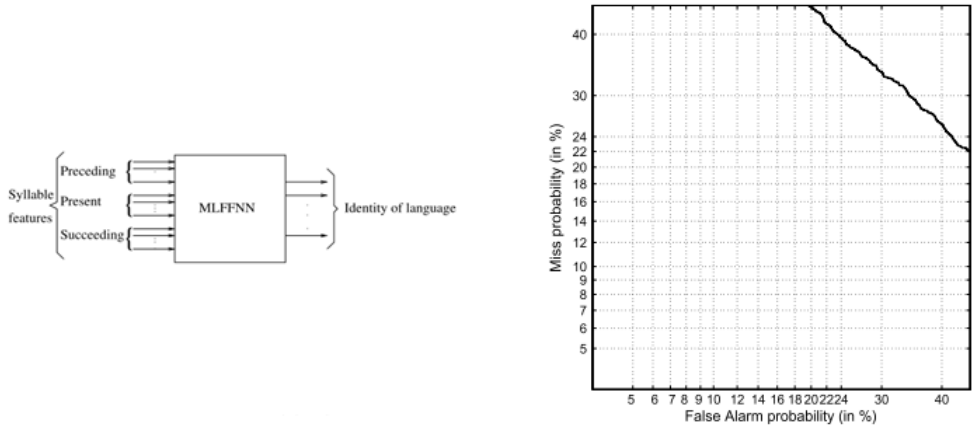

FIGURE 0.2: Expérimentation sur NIST 2003 LRE

Les résultats sont de 32% de taux d'erreur moyen, ce qui est proche des résultats de performances d'autres systèmes à base de prosodie

– **Extraction des caractéristiques prosodiques pour la reconnaissance du locuteur**
Pour expérimenter la reconnaissance du locuteur, les vecteurs de sept dimensions de caractéristiques prosodiques dérivés du signal de parole correspondant à deux locuteurs de sexe masculin contenu dans la base de données de NIST 2003 est comprimé (en utilisant un réseau neuronal auto associatif)(AANN)

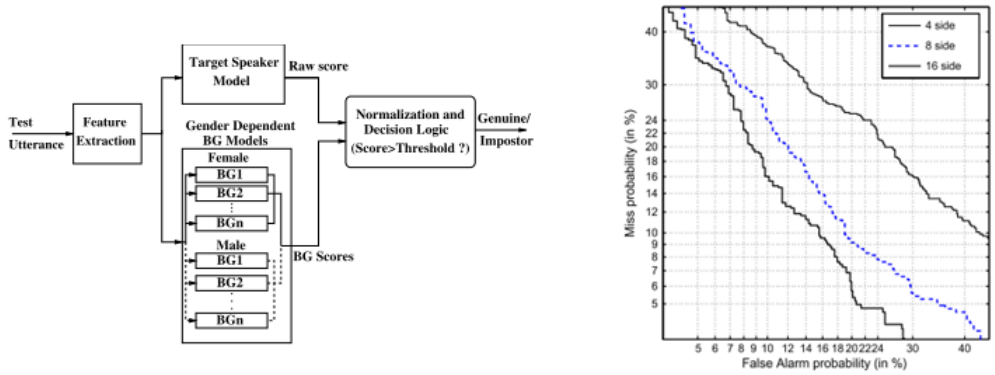

FIGURE 0.3: Expérimentation sur NIST 2003 SRE

Notre système de vérification du locuteur basé sur la prosodie donne lieu à un taux d'erreur (EER) de 12,4%, 15% et 23% pour des conversations à 16, à 8 et à 4. Les performances sont proches des résultats rapportés pour le NIST 2001.

Opinion éclairée

On peut cependant regretter que l'efficacité des caractéristiques prosodiques extraites (évaluée en utilisant l'approche proposée pour la reconnaissance des langues dans le cas d'NIST LRE 2003) est tributaire du nombre de données vocales disponibles pour l'apprentissage des réseaux de neurones. De même, les performances des caractéristiques prosodiques pour la reconnaissance du locuteur vérifiées par NIST SRE 2003, semble être significative en particulier pour les cas où les données vocales étaient disponible pour l'apprentissage des modèles