Extraction et représentation des caractéristiques prosodiques pour la reconnaissance de la langue du locuteur (Leena Mary , B. Yegnanarayana,2008)

Boutheina WATHEK Armel Sitou AFANOU

Master SETI
Université Paris Sud

Reconnaissance Vocale et Indexation Multilingue, Fevrier 2013

Outline

Outline

Introduction.

Qu'est-ce que la prosodie ?

Caractéristiques

- Intonation
- Rythme
- Stress

Outline

Extraction de la prosodie

Extraction de la prosodie

Approches précédentes

- L'approche basée sur l'ASR

Approche présentée dans l'article

- Segmentation du signal en unités syllabiques
- Extraction de la fréquence fondamentale F0 et de l'énergie

Extraction de la prosodie

Extraction de la prosodie

Approches précédentes

- L'approche basée sur l'ASR

Approche présentée dans l'article

- Segmentation du signal en unités syllabiques
- Extraction de la fréquence fondamentale F0 et de l'énergie

Outline

La segmentation syllabique

Les VOPs

VOPs (Vowels Onset Point)

- Les points d'apparition de voyelle sont des événements importants dans la production de la parole.
- Le point d'apparition de voyelle fait référence à l'instant où le début de la voyelle a lieu dans une syllabe
- La région entre deux VOP successifs est ensuite considérée comme une région syllabique.

Outline

La segmentation syllabique

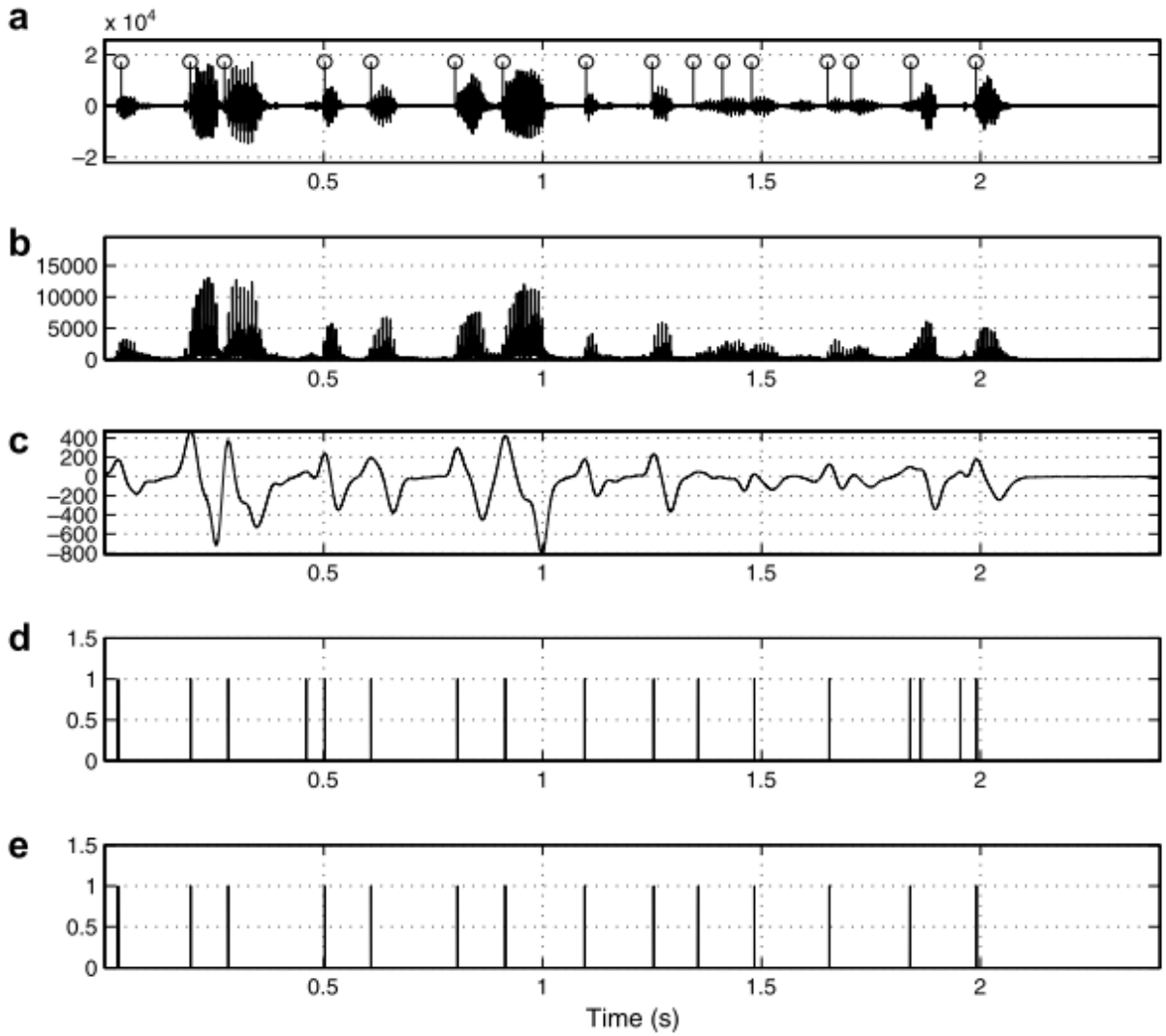
En pratique

Extraction des VOPs

- Utilisation de l'enveloppe de Hilbert : $h_e = \sqrt{r^2(n) + r_h^2(n)}$ 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)
- Détection des pics
- Elimination des pics parasites

La segmentation syllabique

En pratique - Schéma

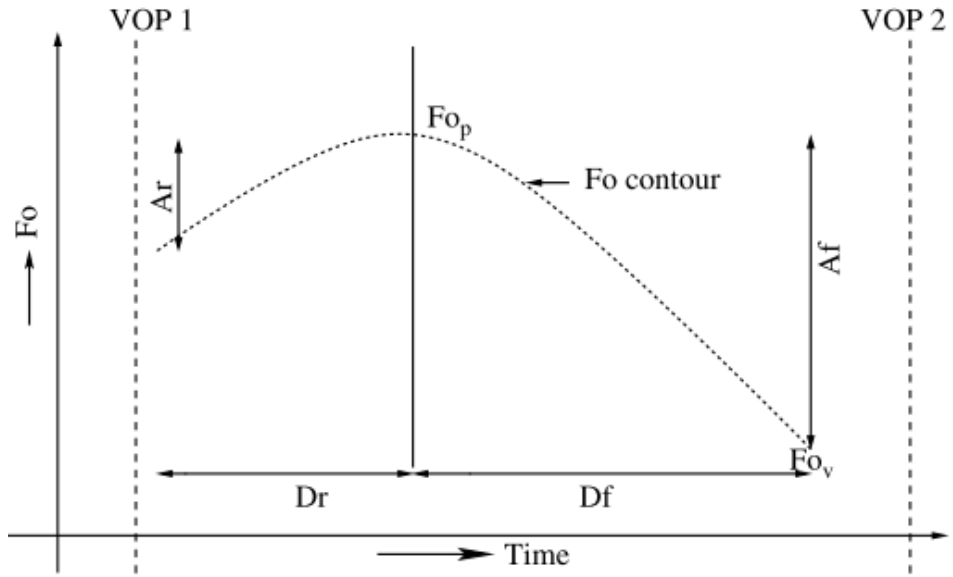


Extraction de la prosodie

Fréquence fondamentale F0

Extraction des paramètres

- Fréquence fondamentale



Extraction de la prosodie

Paramètres encodant la prosodie

- distance entre VOPs successifs,
- durée du voisement (Dv),
- variation de F0 (DF0), $DF0 = F_{0p} - F_{0v}$
- distance du sommet de F0 par rapport au VOP,
- déclinaison de l'amplitude (At), $A_t = \frac{|A_r| - |A_f|}{|A_r| + |A_f|}$
- déclinaison de la durée (Dt), $D_t = \frac{|Dr| - |D_f|}{|D_r| + |D_f|}$
- la variation de l'énergie (dE)

Caractéristiques prosodiques pour la reconnaisance de la langue : Techniques & Résultats

Techniques

Technique utilisée

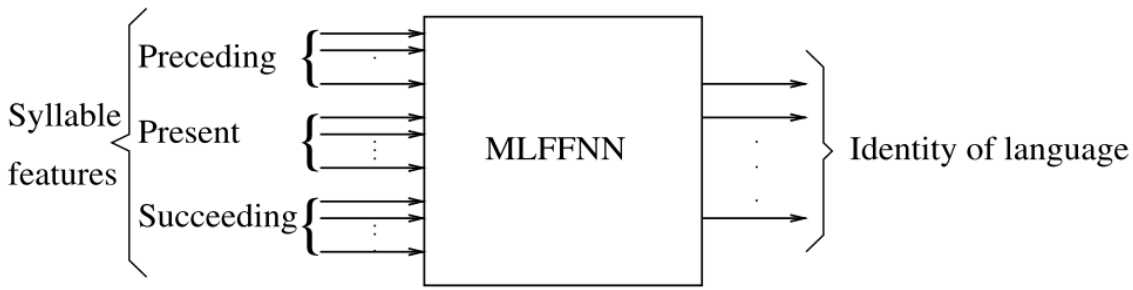
- Les vecteurs de tests sont formés de 21 éléments répartis sur les caractéristiques prosodiques de 3 syllabes consécutives (avec les 7 paramètres définis précédemment)
- La méthode est ensuite testée sur la base de données du système de reconnaissance de langue NIST 2003, le but étant de reconnaitre sur un échantillon de conversation téléphonique la langue parmi la liste de langues cible Arabe, anglais, farsi, français, Allemand, hindi, japonais, coréen, mandarin, espagnol ,Tamoul et vietnamien

Caractéristiques prosodiques pour la reconnaisance de la langue : Techniques & Résultats

Techniques

Technique utilisée

- L'expérimentation est réalisée sur un classifieur basé sur un réseau de neurones multi-couches entrainé sur une base de test de 500 éléments

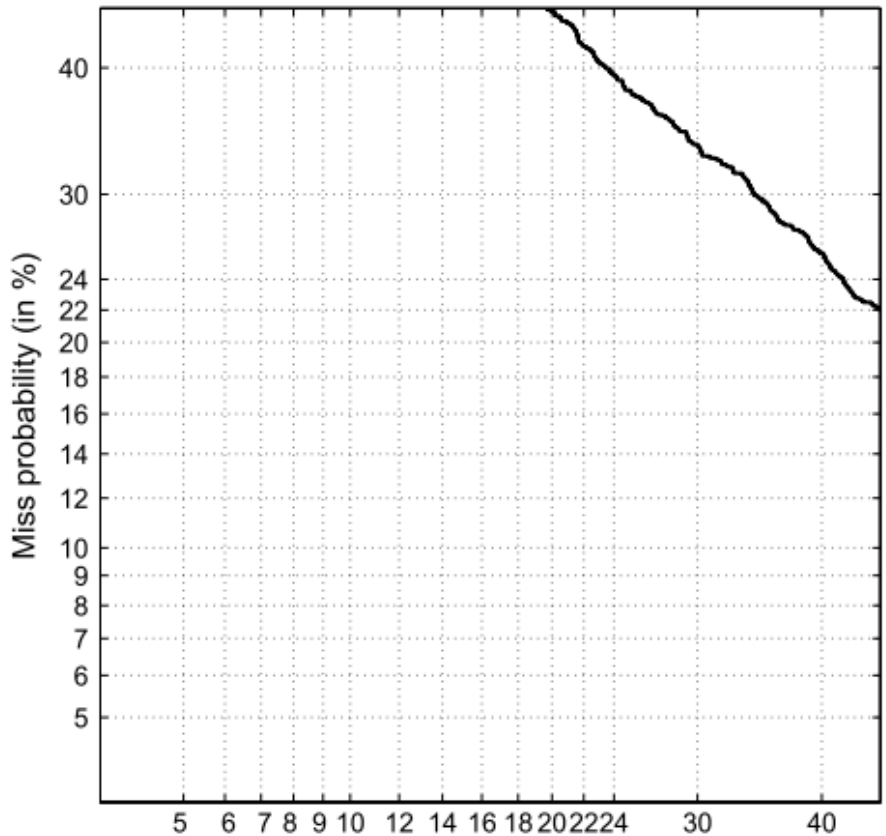


Caractéristiques prosodiques pour la reconnaisance de la langue : Techniques & Résultats

Résultats

Résultats obtenus

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



Caractéristiques prosodiques pour la reconnaisance du locuteur : Techniques & Résultats

Techniques

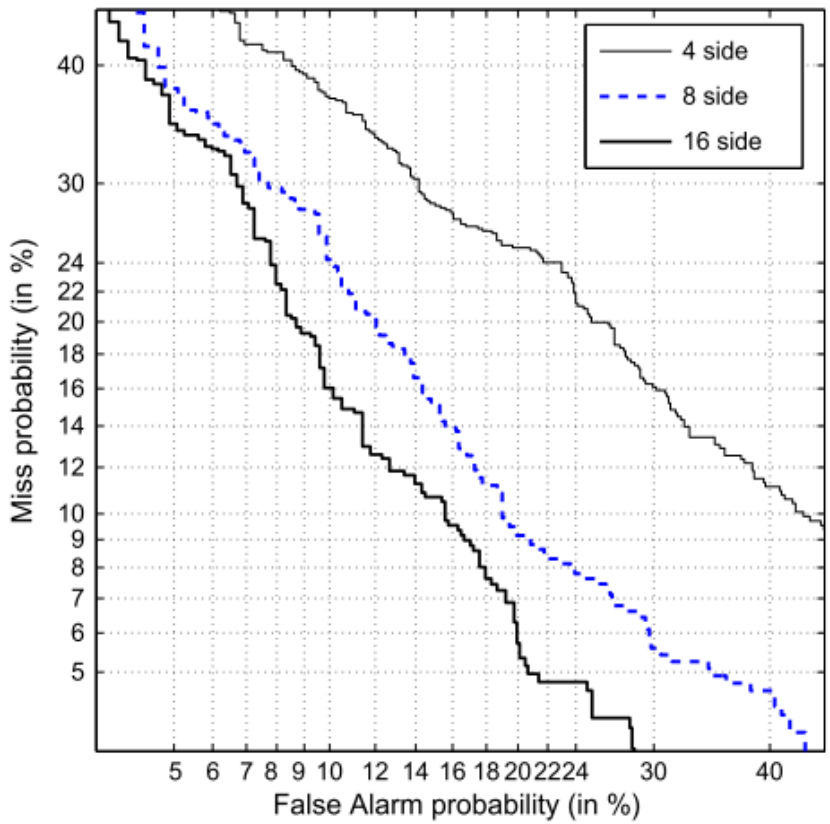
Technique

- Pour illustrer la reconnaissance du locuteur, les vecteurs de sept dimensions de caractéristiques prosodiques dérivés de parole correspondant à deux locuteurs de sexe masculin dans la base de données NIST 2003 sont utilisés
- Ces vecteurs sont appliqués sur un réseau neuronal auto associatif (AANN)

Caractéristiques prosodiques pour la reconnaisance du locuteur : Techniques & Résultats

Résultats

- Notre système de vérification du locuteur basé prosodie donne lieu à un EER de 12,4%, 15% et 23% pour des conversations à 16, à 8 et à 4.



Caractéristiques prosodiques pour la reconnaisance du locuteur : Techniques & Résultats

Résultats

Résultats

- Les meilleures performances sont obtenues dans le cas de conversation à 16 et à 8 ; ce qui montre qu'un apprentissage accru est nécessaire pour bien saisir les caractéristiques prosodiques.
- Les performances de notre système basé sur la prosodie est proche des résultats rapportés pour le NIST 2001 à l'aide des caractéristiques dérivées sans utiliser ASR.

Conclusions

Résumé des techniques

- Détéctions des VOPs grâce la transformée de Hilbert
- Utilisation des réseaux de neurones pour la reconnaissance de la langue et du locuteur

Apports

- Non necessité d'un ASR
- Approche syllabique

Conclusions

Résumé des techniques

- Détéctions des VOPs grâce la transformée de Hilbert
- Utilisation des réseaux de neurones pour la reconnaissance de la langue et du locuteur

Apports

- Non necessité d'un ASR
- Approche syllabique

Critiques

Performances

- Les performances des caractéristiques prosodiques pour la reconnaissance du locuteur vérifiées par NIST SRE 2003, semble être significatives en particulier pour les cas où les données vocales étaient disponible pour former les modèles.

<2->Limites

- Le succès de reconnaissance de la langue a été limité par le nombre de données vocales disponibles pour l'apprentissage des réseaux de neurones.

For Further Reading I

Carolin Schmid, Cédric Gendrot , Martine Adda-Decker.
Une comparaison de la déclinaison de F0 entre le français et l'allemand journalistiques.
, 2012.