

Détection de parole superposée pour la tâche ETAPE SES-2

Claude Barras, Cong-Thanh Do

LIMSI-CNRS, Orsay, France

Journées ETAPE
Rennes, 19 juin 2012

Motivations

Annotation de disfluences et de parole superposée

Détection de la parole superposée dans ETAPE

Conclusions

Intérêt croissant pour la parole superposée

- rare en parole préparée et actualité (ignorée en STT)
- fréquente en parole spontanée (téléphone, réunions, talk-shows)

Pourquoi détecter la parole superposée ?

- pour la mettre de côté : erreurs de transcription
- pour la transcrire ? pas toujours nécessaire
- pour comprendre la nature de l'interaction

Comment la détecter ?

- séparation de source ou localisation en situation multi-canal
- décodage conjoint des voix superposées
- analyse spectrale et harmonique à court terme

Annotation de disfluences et de parole superposée

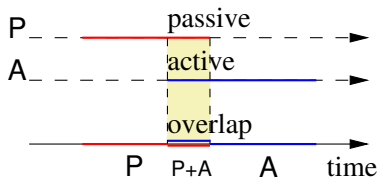
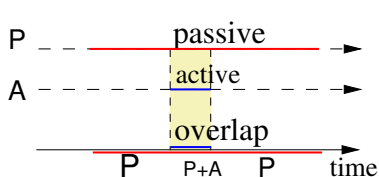
Adda-Decker et al. *Annotation and analysis of overlapping speech in political interviews*. LREC 2008.

Corpus et outils

- corpus de débats politiques l'Heure de vérité
- outil Transcriber adapté pour l'annotation de superpositions
- marquage des mots sans localisation temporelle associée

2 situations principales

- avec ou sans changement de locuteur primaire



Etiquetage en 4 classes

- suite à un processus itératif d'annotation
- augmentation de l'accord inter-annotateurs

signaux d'écoute (backchannels)

A: c'est simplement /le fait/ /B: hmm/ de...

commentaires ou précisions

A: une dernière question /sur/ /B: très courte/ sur votre...

prises de parole interruptives

A: et dans /ce cas.../

B: /Je veux/ revenir à...

prises de parole anticipées

A: et c'est cela qui mène à l'action /humanitaire ?/

B: /eh bien/ je pense

Résultats

Superpositions courtes mais fréquentes

- 3 à 4 par minute, 5% des mots prononcés

Cas non-intrusif (backchannels, anticipation)

- superposition courte (1 à 2 mots), peu d'impact

Cas intrusif (interruptif, commentaire)

- superposition plus longue (3 à 4 mots)
- forte augmentation des disfluences, surtout des répétitions

Influence du rôle du locuteur (journaliste ou invité)
sur la production de disfluences

Défi

- la parole est une alternance de sons voisées, non voisées et de silences courts avec des variations importante du niveau d'énergie
- en interaction normale, les vrais superpositions sont très courtes : la plupart du temps, la voix d'un des locuteurs domine l'autre

Approche choisie

- pré-traitement de complexité limitée
- analyse à court terme du timbre
- prise en compte de la structure harmonique

Modèles de voix isolée et de voix superposée

- paramètres cepstraux classiques utilisés en reconnaissance de parole (12 PLP+E+ Δ + $\Delta\Delta$, 30ms win, 10ms step)
- trois modèles $\{\lambda_i\}_{i=0\dots 2}$ à 256 Gaussiennes appris sur le train d'ETAPE suivant l'alignement forcé (λ_0 : silence, λ_1 : parole isolée et λ_2 : parole superposée)

Décision

- décodage de Viterbi avec durée minimale par état
- décodage de Viterbi avec pénalité de transition
- rapport de log-vraisemblance lissé sur une fenêtre de Hamming comparé à un seuil

$$l_t = \sum_{j=0}^{d-1} H(j) \cdot \log \frac{f(x_{t+j}|\lambda_2)}{f(x_{t+j}|\lambda_0) + f(x_{t+j}|\lambda_1)}$$

Analyse multi-pitch

- identifier le mélange de sons voisés en parole superposée
- peigne à suppression d'harmoniques (PSH) :
Liénard et al. *Using sets of combs to control pitch estimation errors*, Acoustics'08.
- 0, 1 ou 2 valeurs de pitch pour chaque trame

Paramètre dérivé

- soit $h_t \in \{0, 1, 2\}$ le nombre d'hypothèses de pitch à la trame x_t
- lissage sur une fenêtre de Hamming

$$p_t = \sum_{j=0}^{d-1} H(j) \cdot h_{t+j}$$

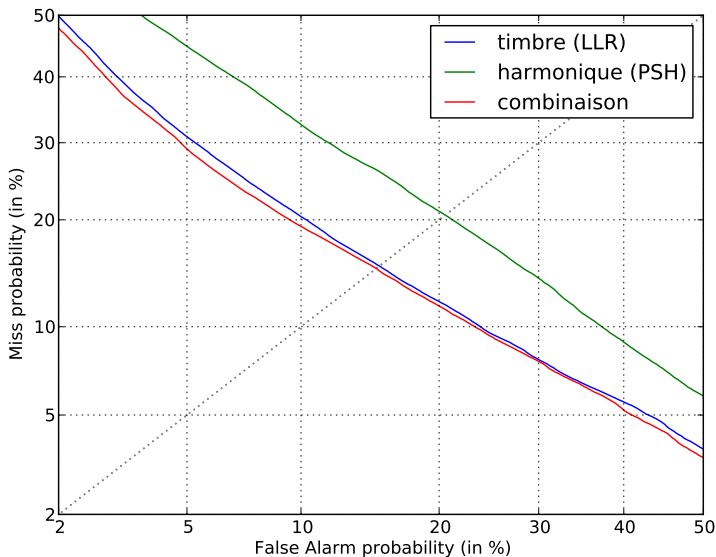
Evaluation

- données de développement ELDA avec alignement forcé
4,9% du temps de parole en superposition
- métrique proposée dans le plan d'évaluation : rappel, précision et F1-mesure en durée de parole superposée détectée
- mesures alternatives : EER, taux d'erreur. . .

Résultats

Système		EER (%)	F-mesure
Timbre	Viterbi + durée 3 sec.	-	0.486
	Viterbi + pénalité 70	-	0.520
	LLR lissé l_t	15.1	0.545
Harmonique	PSH lissé p_t	20.5	0.453
Combinaison	$0.3 l_t + 0.7 p_t$	14.9	0.558

Evaluation SES-2



- pas encore de score calculé sur le corpus d'évaluation
- premier système simple de détection de parole superposée
- résultats encourageants (F-mesure supérieure à 50%, mais insuffisant si on compte le taux d'erreur en trame)
- à intégrer à SRL