

Étape : système TRS du LIUM

**Paul Deléglise, Yannick Estève, Fethi Bougares,
Mickael Rouvier**

LIUM, Université du Maine



18 juin 2012

Principes généraux

- Système multi-passes :
 - ① segmentation : LIUM_SpkDiarization
 - ② CMLLR sur 1 GMM (pré-adaptation)
 - ③ passe 1: → graphe pour CMLLR (adaptation pour les modèles SAT) (Sphinx 3.7)
 - ④ passe 2 : signal transformé → graphe (Sphinx3.7, SAT MPE)
 - ⑤ passe 3 : réévaluation acoustique du graphe avec triphones inter-mots → graphe (Sphinx4)
 - ⑥ passe 4 : réévaluation du graphe en 4G
 - ⑦ passe 5 : réévaluation en 5G
 - ⑧ passe 6 : réseaux de confusion → séquence des mots les plus probables sans seuillage
- Construire plusieurs systèmes en faisant varier certaines passes, par exemple :
 - P5: ML à repli ou ML continu
 - P3 : MPE ou MLP
- fusion de graphes pour construire les réseaux de confusion + Rover

Corpus acoustiques

- ESTER1 : apprentissage + développement
- ESTER2 : apprentissage sauf AFRICA1
- EPAC : 60 heures
- sous-total : 230 heures
- ETAPE apprentissage : 17 heures 30
- news (2007-2008) : 227 heures
- podcast débats 2011, transcriptions rapides (pas d'indication temporelle, ni de locuteur, sentences longues > 70s) 99 heures
- total 511 heures studio + 60 heures téléphone

Modèles acoustiques

- 33 phonèmes + silence + 4 fillers (inspiration, pi, bruit, musique)
- passe1: taille du modèle : 2500 états partagés X 22 gaussiennes X 39 paramètres PLP
- passe2 :
 - ① sphinx 3.7 : 10000 états partagés X 48 gaussiennes X 39 paramètres PLP MPE/SAT
 - ② RASR : 15000 mélanges, 2783349 gaussiennes avec une variance partagée paramétrisation LDA.
- passe3 : même taille que P2 avec soit PLP MPE/SAT, soit 79 paramètres en 2 flux 39 PLP + 40 MLP

Paramétrisation MLP

- utilisation de quicknet de ICSI
- apprentissage en sablier
 - entrée : 9 trames de 39 MFCC \rightarrow 351
 - couches cachées 4000 et 40 (\neq phonèmes pour des raisons historiques)
 - en sortie : 102 = 3 états X(33 phonèmes + 1 pour les fillers)
- décodage en bouteille
 - réseau de tailles 351 X 4000 X 40
 - ACP sur les 40 sorties
 - construction d'un double flux 39 PLP X 40 MLP.

Corpus linguistiques : données

- corpus audio + PFC : 8,3 M de mots
- googleNews : 207M de mots
- french gigaWord : 1035M de mots
- le Monde (ESTER1+ESTER2) : 373M de mots
- OCR journaux télé : 9M de mots
- télétexte journaux TF1 : 2M de mots
- Savoie (site web du Dauphiné libéré) : 0,97 M de mots
- Google Ngram

Construction du vocabulaire

- construction d'un modèle unigramme par combinaison linéaire optimisée sur le dev d'étape
- sélection des 159k mots les plus probables
- phonétisation BDLEX + LIAPHON filtrée et pondérée avec l'alignement sur le corpus d'apprentissage des modèles acoustiques
- \Rightarrow filtrage des Google Ngram : 69M de 5-grams comportant au plus 1 OOV

Modèles de Langage

- modèle à repli Kneyser Ney modifié sans coupure
- un modèle pour chaque corpus + interpolation optimisée sur dev ETAPE pour la perplexité
- poids :

audio	GNE	GW	Monde	OCR	télétext	Savoie	GNgram
0,59	0,14	0,047	0,13	0,06	0,02	0,0063	0,03

- nombre de N-gram.

1-gram	2-gram	3-gram	4-gram	5-gram
159k	37M	202M	465M	700M

- en 5G : construction d'un modèle continu en sous-échantillonnant les corpus selon les poids du LM classique après filtrage.

Construction du primaire

- fusion des graphes entre MPE-CSLM, MLP-CSLM, MLP5G, MPE4G, RASR-4G \rightarrow F5
- rover entre F5, MLP4G, MLP5G, MPE-CSLM, MLP4G-segbase, RASR-4G

- hésitations dans la référence et enlevées de l'hypothèse

Nb	cor	sub	sup	ins	WER
82237	77.40	10.68	8.65	2.28	21.62

- aucune hésitation (nglm)

Nb	cor	sub	sup	ins	WER
79847	79.71	11.01	8.91	2.35	22.26

Résultats du primaire

- évaluation sans hésitations (ni dans la référence, ni dans les hypothèses)

condition	Nb	cor	sub	sup	ins	erreur locu- teurs	WER/ Speaker at- tribued WER
sans sur- posée	79847	79.7	11.0	8.9	2.3	-	22.3
toutes zones (NIST estimation)	96638	72.4	12.5	14.8	2.1	12.2	29.4/ 41.6
superposées seules	16791	37,4	19,3	42,8	0,9	-	63,1

Résultats divers et variés

- évaluation sans hésitations (ni dans la référence ni dans les hypothèses), zones non-superposées

système	Nb	cor	sub	sup	ins	WER
primaire	79847	79.71	11.01	8.91	2.35	22.26
F5	79847	78.99	11.59	9.06	2.29	22.94
MLP-5G	79847	79.02	12.70	7.92	2.98	23.60
MLP-4G	79847	78.90	12.78	7.96	3.01	23.75
MLP-CSLM	79847	78.80	13.02	7.81	3.18	24.02
MPE-4G	79847	78.46	12.76	8.41	2.88	24.05
MPE-CSLM	79847	78.68	13.28	7.67	3.48	24.43

Variations sur le LM

- variations sur les corpus utilisés pour le LM : suppression des corpus OCR ou OCR et télétexte ou new 2007&2008 et podcast.

système	Nb	cor	sub	sup	ins	WER
base MLP4G	79847	78.90	12.78	7.96	3.01	23.75
sans OCR	79847	78.76	12.85	8.02	3.06	23.93
sans OCR ni TÉLÉ	79847	78.76	12.86	8.01	3.06	23.93
sans news ni podcast	79847	78.58	12.88	8.18	2.99	24.05

Les différentes passes

système	Nb	cor	sub	sup	ins	WER
passe 1	79847	70.00	19.57	10.06	4.35	33.98
passe 2	79847	74.36	16.52	8.76	4.08	29.36
passe 3	79847	77.24	14.05	8.34	3.16	25.55
passe 4	79847	78.66	13.02	7.96	3.03	24.01
R de confusion p4	79847	78.90	12.78	7.96	3.01	23.75
passe 3 de MPE	79847	76.97	14.21	8.45	3.21	25.87

Projet ASH : description

- collaboration de systèmes en parallèle pour une latence minimale
- décodage direct en une seule passe sans adaptation
- 2 systèmes auxiliaires fournissent en cours de décodage leurs hypothèses sous forme de sac de n-grams (BONG) à un système primaire qui modifie les siennes en conséquence.

ASH : résultats

système	Nb	cor	sub	sup	ins	WER
RASR	79847	69.51	16.01	14.10	2.97	33.08
speeral	79847	65.23	22.98	11.40	4.90	39.28
sphinx-BONG	79847	69.58	16.47	13.57	2.52	32.57
rover	79847	71.00	12.40	16.20	1.77	30.38
sphinx-base	79847	0.13	18.20	11.28	3.98	33.46
rover-base	79847	73.17	16.06	10.39	4.41	30.86