IRENE, le système commun IRISA – ENST d'indexation d'émissions radiophoniques.

Guillaume Gravier, François Yvon, Mathieu Ben

IRISA/METISS & ENST/INFRES

ggravier@irisa.fr, mben@irisa.fr, yvon@infres.enst.fr

1. Introduction

Nous décrivons dans cet article les différentes composantes du système d'indexation d'émissions radiophoniques développé conjointement par l'IRISA et l'ENST Paris dans le cadre de la campagne ESTER.

La description du système d'indexation est divisée en trois modules. Le premier module permet de détecter les portions du document contenant de la parole ainsi que celles contenant de la musique (tâche SES). Le deuxième module vise à établir une transcription orthographique du document (tâche TRS). Enfin, le troisième module permet de déterminer les portions du document où un locuteur donné, connu au préalable, est présent (tâche SVL).

Le module de transcription est un développement commun IRISA-ENST (IRENE). Les modules de suivi de locuteur et de détection de parole sont développés par l'IRISA.

2. Détection parole/musique

Le module de détection de parole et de musique permet de déterminer les portions du document contenant de la parole et celles contenant de la musique. L'objectif principal de ce module est de permettre de détecter les zones contenant de la parole pour pouvoir y appliquer les traitements adéquats. Il a donc été optimisé pour cette tâche, sans chercher à optimiser la détection de musique.

L'approche adoptée utilise un modèle de Markov caché ergodique à quatre états représentant les classes sonores 'parole', 'parole et musique', 'musique' et 'silence'. Des densités multi-gaussiennes à 256 composantes sont associées aux états. Les paramètres du modèle sont estimés sur des données issues du corpus d'apprentissage ESTER phase 2.

Le signal est représenté par 14 coefficients cepstraux, l'énergie ainsi que les dérivées première et seconde, soit un vecteur de dimension 45. Une normalisation par centrage et réduction glissants des composantes du vecteur de paramètres, sur un horizon de 4s, est utilisée.

Les performances de cette approche en terme d'omissions (faux rejet, %fr) et d'insertions (fausse alarme, %fa) sont données dans le tableau 1.

	parole		musique	
corpus	%fr	%fa	%fr	%fa
dev	1,25	5,76	77,70	0,23
eva	1.93	14.01	78.43	1.03

Table 1: Résultats du suivi de classes sonores sur les corpus ESTER dev et eva phase 2.

La détection parole/musique s'effectue en 0.03 fois le temps réel (TR) sur un processeur Pentium IV, 3,2 GHz, RAM 2 Gb.

3. Transcription

Les trois transcriptions soumises lors de la campagne d'évaluation proviennent de deux systèmes différents, l'un basé sur le décodeur publique Sphinx, l'autre sur notre décodeur Sirocco en combinaison avec HTK. Les deux systèmes partagent la même segmentation en groupes de souffle et se basent sur le même lexique et sur le même modèle de langage. En revanche, les modèles acoustiques sont propres à chaque décodeur.

3.1. Segmentation en groupe de souffle

La première étape du module de transcription est la segmentation des zones de parole en (pseudo) groupes de souffle. Cette segmentation se fait en deux étapes.

Dans un premier temps, chaque segment de parole issu du module de détection de la parole est segmenté selon la bande passante. Cette segmentation se fait à l'aide d'un modèle de Markov caché ergodique à deux états représentant la parole de qualité studio et téléphone respectivement. La représentation du signal utilisée est la même que pour la détection de parole. Les paramètres de ce modèle sont estimés sur les données de développement de la phase 2.

La deuxième étape consiste en une segmentation basée sur l'énergie pour détecter les pauses correspondant aux inspirations. Cette détection de pauses se fait indépendamment sur chacun des segments issus de la segmentation en bande passante, à l'aide d'une modélisation bi-gaussienne de l'énergie des trames. Les pauses de plus de 0.3s, qui correspondent normalement à une inspiration ou à un silence, sont alors utilisées pour découper le signal en groupes de souffle. Quelques heuristiques permettent d'éviter la présence de segments trop long ou la détection à tord de silence.

Le tableau 2 donne la longueur moyenne des segments pour la segmentation automatique et pour une segmentation basée sur la transcription manuelle[1] sur un sous-ensemble de 4h du corpus de développement phase 2 et sur le corpus d'évaluation phase 2. Les performances d'un système de transcription automatique sont aussi données.

Sur chaque segment obtenu, le sexe du locuteur est déterminé selon un critère de maximum de vraisemblance avec des modèles de mélange de gaussiennes à 256 composantes. Le taux d'identification correcte du sexe sur le corpus de développement de la phase 1 est de 96%.

La segmentation en groupes de souffle s'effectue en 0.02 x TR, la majeure partie du temps étant dévolue à la segmentation

[1] Segmentation basée sur les balises transcriber '¡Sync¿', équivalente à la segmentation STM de référence.

	manuelle		automatique	
corpus	durée (s)	%erreur	durée (s)	%erreur
dev phase 2	3.1	31,4	3.8	33,1
eva phase 2	3.3	35,4	3.7	37,9

Table 2: Durée moyenne des segments et taux d'erreur avec les segmentations automatique et manuelle sur un sous-ensemble de 4h du corpus dev et sur eva phase 2 (système Sirocco 3g).

	nb. de n-grammes			perplexité	
système	n=2	n=3	n=4	dev	eva
sphinx (primaire)	4,25M	4,25M	–	85.1	101.6
sirocco 3g	8,5M	7,3M	–	83.7	99.6
sirocco 4g	8,5M	7,7M	5M	75,9	91.6

Table 3: Taille et perplexité sur les corpus de développement et d'évalation des modèles de langage utilisés.

en bande passante (0.015 x TR). La détermination du sexe du locuteur s'effectue en 0.014 x TR.

3.2. Ressources

3.2.1. Lexique et modèles de langage

Les deux systèmes utilisent un vocabulaire de 65 000 mots[2]. La liste des mots est constituée en interpolant de manière optimale, au sens de la perplexité sur le corpus de développement, des modèles unigrammes construits sur des sous-ensembles du corpus "Le Monde" et du corpus d'apprentissage phase 2. Les transcriptions phonétiques des mots du lexique sont soit extraites des bases de données lexicales de l'ENST, en particulier ILPho, soit établies manuellement. En moyenne, il y a 1,8 variantes de prononciation par forme graphique, correspondant à 118517 transcriptions phonétiques. Les taux de mots hors vocabulaire sont de 0.9% sur le corpus de développement et de 1,2% sur le corpus de test.

Le modèle de langage est construit en interpolant de manière optimale, au sens de la perplexité du corpus de développement, des modèles de langage quadrigramme appris sur des sous-ensemble du corpus "Le Monde" (années 96-03) et un modèle trigramme appris sur les transcriptions du corpus d'entraînement de la phase 2. Tous les modèles sont estimés avec le SRI-SLM Toolkit et le lissage de Knesser et Ney, puis éventuellement simplifiés (en trigramme), et élagués avec différents facteurs de réduction.

3.2.2. Modèles acoustiques

Les deux systèmes utilisent des modèles de Markov cachés à trois états avec une topologie gauche-droite et des densités gaussiennes associées aux états. La paramétrisation du signal consiste en 12 coefficients cepstraux et l'énergie normalisée, plus les dérivées première et seconde. Dans le système Sphinx, les coefficients cepstraux sont normalisés par soustraction de la moyenne. Pour le système Sirocco, l'ensemble des 39 coefficients sont normalisés par centrage et réduction.

Les modèles acoustiques pour le système Sphinx sont des modèles de phones contextuels (triphones) dépendant du sexe et possèdent 6 111 états distincts, soit environ 200 000 gaussiennes. Le système Sirocco utilise un jeu de monophones avec 64 gaussiennes par état (soit un total de 7 170 gaussiennes) et un jeu de triphones avec 6 244 états distincts et 32 gaussiennes par états, soit également environ 200 000 gaussiennes.

[2]La notion de mot doit ici être comprise comme une forme graphique.

		sirocco 3g		sirocco 4g	
corpus	sphinx	passe 1	passe 2	passe 1	passe 2
dev	–	36,7	32,0	–	–
eva	35,7	42,5	37,9	42,6	38,6

Table 4: Performances obtenus par les différents systèmes sur les corpus de développement et d'évaluation de la phase 2 (version 1.6 des transcriptions du corpus d'évaluation) .

Les paramètres des modèles sont estimés uniquement sur la base des données d'apprentissage disponibles pour la phase 2. La transcription des données d'apprentissage est phonétisée à l'aide d'un lexique, établi comme le lexique de décodage. Une procédure de réalignement permet de déterminer les variantes de prononciations. Pour le système Sphinx, le réalignement est réalisé avec les modèles monophones obtenus après une première passe d'apprentissage. Pour le système Sirocco, les modèles issus de la phase 1 sont utilisés pour choisir avant l'apprentissage les variantes de prononciations.

3.3. Décodeurs

3.3.1. Description

Le décodage avec Sphinx se fait en une seule passe utilisant les modèles contextuelles. Le décodage avec Sirocco et HTK se fait en deux passes : une première passe avec Sirocco permet de générer un graphe de mots, avec des monophones pour la variante trigramme soumise et des modèles contextuels à l'intérieur des mots pour la variante quadrigramme; la deuxième passe permet de réévaluer le graphe de mot avec des modèles contextuels entre mots.

Les résultats obtenus par les trois systèmes soumis sont résumés dans le tableau 4.

Le temps de calcul pour le décodeur Sphinx est de 15 x RT sur un processeur Pentium 4, 3 GHz, 2G RAM. Pour le système Sirocco avec des monophones en première passe, le temps de calcul total est de 9 x RT pour le système soumis avec une première passe en 5 x RT. Ce temps peut-être réduit à 7 x RT sans perte de performance en utilisant des techniques d'élagage lors de la deuxième passe, ce qui n'a pas été fait pour la campagne. Le temps de calcul pour le système quadrigramme n'a pas été mesuré proprement mais semble s'établir autour de 15 x RT avec une première passe en 12 x RT.

3.3.2. Quelques expériences complémentaires

Pour des raisons de temps, les paramètres du système sirocco 3g n'ont pas été optimisés de manière systématique pour la campagne. Après une meilleure optimisation, les taux d'erreurs sont de 30,1% et 34,8%, respectivement sur un sous-ensemble de 4h du corpus de développement et sur le corpus d'évaluation.

Les mauvaises performances du système quadrigramme par rapport au système trigramme peuvent paraître surprenante. La différence de performance s'explique en fait par l'utilisation de modèles contextuels lors de la première passe dans le système sirocco 4g. Si l'utilisation de tels modèles permet une amélioration des performances à l'issu de la première passe[3], cet avantage s'efface après la deuxième passe. En re-

[3]Une telle amélioration n'est pas visible dans les résultats du tableau 4 à cause du mauvais réglage des paramètres. En revanche, elle est apparue lors de la phase d'optimisation des paramètres effec-

vanche, en utilisant le modèle de langage quadrigramme avec les monophones lors de la première passe, on obtient une légère amélioration : 34,3% au lieu de 34,8% sur le corpus de test.

3.4. Perspectives

De nombreuses perspectives restent à envisager, comme l'ajout d'un module d'adaptation au locuteur, l'utilisation de modèles plus complexes (plus de gaussiennes, séparation homme/femme dans le système Sirocco, etc) ou encore l'amélioration de la segmentation.

Dans l'immédiat, nous nous intéressons à chercher une meilleure exploitation du modèle quadrigramme qui ne permet à l'heure actuelle qu'un gain marginal. L'utilisation de modèle de langage thématique est aussi une extension envisagée à nos travaux actuels.

A terme, nous souhaitons intégrer l'utilisation de modèles contextuels en première passe dans Sirocco.

4. Suivi de locuteurs

Le module de suivi de locuteurs permet de déterminer les portions d'un document où un locuteur donné est présent. Dans l'approche proposée, le suivi des locuteurs se fait en deux étapes. Une première étape permet de segmenter le document en détectant les changements de locuteurs. La seconde étape consiste à détecter la présence ou l'absence de chaque locuteur référencé sur chacun des segments issus de la première étape.

4.1. Détection des changements de locuteurs

La détection des changements de locuteurs dans un document se fait par une approche basée sur le critère BIC (Bayésian Information Criterion). Les modèles de segments utilisés sont des modèles gaussiens avec des matrices de covariance pleines. La segmentation se fait en trois passes sur les portions étiquetées "parole" issues de la détection parole/musique. Le signal est représenté par les énergies en sortie d'un banc de 24 filtres en échelle MEL. Une version en trois passes de la segmentation BIC est utilisée (cf. documentation audioseg). La fenêtre en première passe est de 3s avec un incrément de 0,6s. En deuxième passe, la taille de la fenêtre est de 4s avec un incrément de 0,1s.

Le tableau 5 donne la longueur moyenne des segments et la pureté globale obtenues avec le système de segmentation en locuteur, pour chacun des fichiers du corpus de développement. Le compromis longueur moyenne/pureté des segments a été réglé en jouant sur le facteur de pénalisation λ intervenant dans le calcul des ΔBIC (ici $\lambda = 0,85$).

Cette segmentation en locuteur s'effectue en 0,094 x TR (0,003 x TR pour la paramétrisation et 0,091 x TR pour la détection des changements de locuteurs).

fichiers	longueur moyenne des segments	pureté globale
20030418_0700_0800_inter	8.6 s	97%
20030418_0800_0900_inter	10.5 s	98%
20030418_1200_1300_info	11.3 s	97%
20030418_1700_1800_info	11.3 s	98%
20030508_1400_1500_rfi	11.3 s	95%
20030509_1400_1500_rfi	11 s	96%
20030717_0700_0715_rtm	12.3 s	98%
20030717_1300_1320_rtm	10.9 s	99%
20030717_2000_2020_rtm	9 s	99%
20030717_2300_2315_rtm	5.5 s	99%
20030719_0700_0715_rtm	7.1 s	99%
20030719_1300_1320_rtm	22.9 s	99%
20030719_2000_2015_rtm	14 s	97%
20030719_2300_2310_rtm	13.1 s	≈100%

Table 5: Longueur moyenne des segments et pureté globale obtenues pour chaque fichier du corpus ESTER dev phase 2.

4.2. Détection des locuteurs par approche GMM-UBM

Lorsque la détection des changements de locuteurs a été effectuée, une approche classique basée sur un système GMM-UBM est utilisée pour détecter individuellement chaque locuteur référencé, sur chacun des segments étiquetés 'parole'.

4.2.1. Estimation des modèles

La paramétrisation fournit 16 coefficients cepstraux LFCC, les 16 coefficients deltas correspondants et le coefficient de delta log-énergie. Ces coefficients sont centrés et réduits sur une fenêtre glissante de 3 secondes et aucun procédé de sélection de trames n'est appliqué.

Le système GMM-UBM utilise un modèle du monde universel (UBM) obtenu par concaténation de 2 modèles du monde dépendant du sexe. Ces modèles 'femme' et 'homme' sont des modèles de mélange de gaussiennes (MMG) à 256 composantes et matrices de covariance diagonales, entraînés respectivement sur 10 heures et 3 heures[4] de parole issues du corpus d'apprentissage phase 2.

Les modèles des locuteurs référencés ont été dérivés du modèle UBM par adaptation MAP (maximum a posteriori) des moyennes avec une itération d'EM. L'ensemble des données disponibles pour chaque locuteur dans le corpus 'train' a été utilisé pour adapter le modèle correspondant. Le facteur de confiance intervenant dans cette adaptation MAP ($\tau = 2,5$) a été optimisé sur le corpus de développement.

4.2.2. Détection des locuteurs

Lors de la phase de suivi, l'information d'appartenance des locuteurs à une ou plusieurs radios spécifiques est utilisée pour ne chercher que les locuteurs concernés. Le score brut utilisé pour la détection est la moyenne des log de rapports de vraisemblances par trame sur l'ensemble du segment considéré. Les log du rapport de vraisemblances par trame sont estimés pour chaque locuteur par la technique N-best avec N=10. Le score brut est ensuite normalisé par T-norm en utilisant 250 modèles de normalisation dont 117 femmes et 133 hommes ayant entre 30 secondes et 10 minutes de parole dans le corpus d'apprentissage et ne figurant pas dans la liste des locuteurs à suivre. Le score normalisé est ensuite comparé à un seuil dépendant du sexe, optimisé sur le corpus de développement pour maximiser la F-mesure.

4.2.3. Résultats

Les résultats obtenus sur les corpus de développement et d'apprentissage de la phase 2 sont résumés dans le tableau 6.

Le temps total de calcul nécessaire pour effectuer le suivi

tuée postérieurement à la campagne.

[4] en raison d'un manque de temps, la quantité de parole utilisée pour estimer les paramètres du modèle du monde 'homme' a été limitée à 3 heures.

corpus	F-mesure	%fr	%fa
dev	0,812	13.8	0,08
eva	0,817	24,6	0,005

Table 6: Résultats obtenus par le système de suivi de locuteurs sur les corpus ESTER dev et eva phase 2.

des locuteurs référencés (dans leurs radios respectives uniquement) correspond à environ 0,4 x TR dont 0,035 x TR pour la détection de parole, 0,095 x RT pour la segmentation en locuteurs et 0,26 x TR pour la détection des locuteurs.

4.3. Quelques éléments d'analyse des résultats

Pour analyser le comportement du système en différents points de fonctionnement, des courbes DET représentant les variations des taux de faux rejets et de fausses acceptations en fonction du seuil de décision sont utilisées.

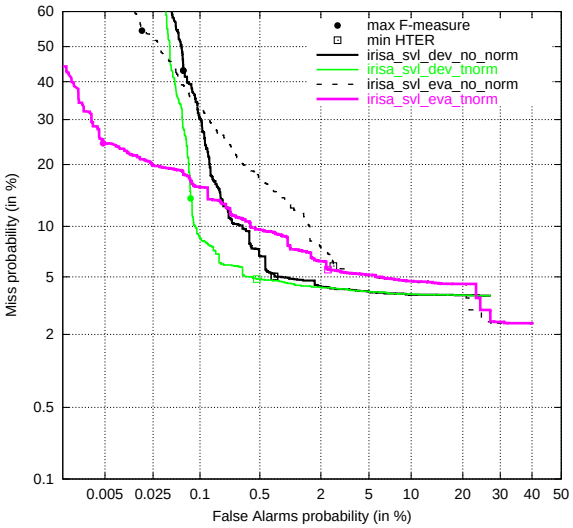

Figure 1: Courbes DET du système SVL sans normalisation de scores (no-norm) et avec T-norm (tnorm) - corpus dev et eva

La figures 1 représente les courbes DET obtenues sur les corpus de développement et d'évaluation, sans normalisation de score et avec normalisation T-norm. Ces courbes montrent que la T-norm apporte une amélioration significative pour les taux de fausse acceptation bas, et notamment aux points correspondant au maximum de la F-mesure. Ce résultat est assez surprenant car la T-norm a pour but de compenser des différences entre les conditions des données de test et celles des données d'apprentissage. Dans le cas des données ESTER, les conditions d'enregistrement sont en général peu variables entre les tests et l'apprentissage. Par contre, la longueur des segments de test varie beaucoup, ce qui peut amener de la variabilité au niveau des scores obtenus sur les différents segments. La T-norm permet peut être de compenser en partie cette variabilité en apportant ainsi de la robustesse au seuil de décision.

Sur la figure 2, les courbes DET correspondant aux locuteurs femmes (female) et hommes (male) obtenues sur le corpus 'dev' sont représentées. On constate d'après ces courbes que les performance sont meilleures pour les locuteurs femmes que pour les locuteurs hommes.

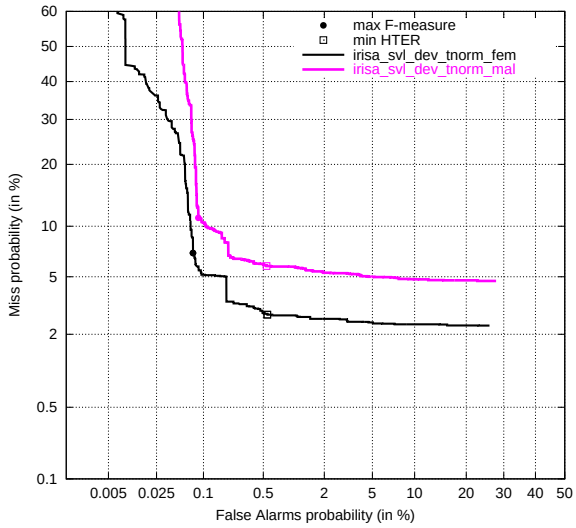

Figure 2: Courbes DET du système SVL avec T-norm pour les locuteurs femmes (fem) et homme (mal) - corpus dev

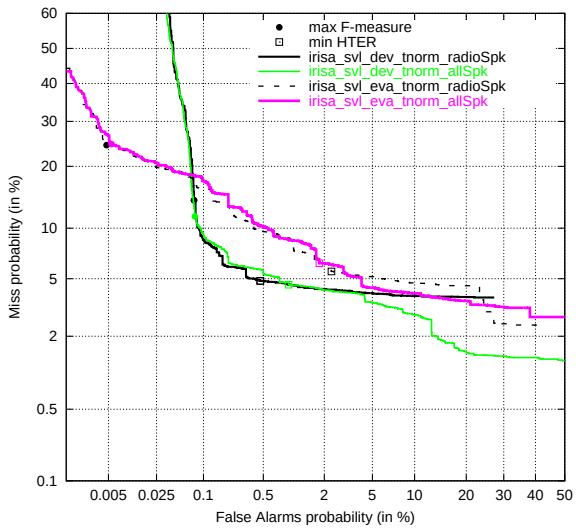

Figure 3: Courbes DET du système SVL - corpus dev et eva

La figure 3, compare les courbes DET obtenues lorsque l'on cherche à détecter l'ensemble des locuteurs référencés pour chaque fichier de test, aux courbes DET obtenues lorsqu'on se limite aux locuteurs spécifiques de chaque radio. Ces courbes montrent que l'utilisation de l'information "radio" n'apporte pas d'amélioration significative des performances.

4.4. Perspectives

En contraste du système primaire de suivi de locuteur, un système basé sur une détection des locuteurs dans un espace des modèles a été expérimenté. Ce système utilise des scores calculés à partir de distances euclidiennes entre les modèles MMG, au lieu des classiques log de rapport de vraisemblances. Lors de la phase de suivi, un modèle MMG est adapté de l'UBM pour chacun des segments de test. Les distances entre modèles sont ensuite rapidement calculables ce qui permet d'accélérer considérablement le suivi de l'ensemble des locuteurs, en particuli-

er lorsque ceux-ci sont nombreux. D'autre part, l'application de la T-norm est également fortement accélérée.

Ce système n'a pour le moment pas permis d'obtenir des performances aussi bonnes que le système de base (F-mesure de 0,75 sur le corpus d'évaluation phase 2, au lieu de 0,81). Par contre, la détection des locuteurs est effectuée en 0,1 x TR avec ce système, au lieu de 0,26 x TR avec le système de base, soit une diminution de plus de 60% des temps de calcul.

Le système opérant dans l'espace des modèles semble mal se comporter pour les segments de tests très courts. Nous travaillons actuellement à l'amélioration de ce système qui pourrait permettre des gain de temps importants. Nous prévoyons également de mettre au point des techniques de compensation agissant dans l'espace des modèles pour diminuer l'influence de facteurs perturbants tels que les changements de conditions (studio/téléphone) ou les bruits de fond (parole+musique ou parole+bruit).

Un travail est également en cours sur la tâche de segmentation et de regroupement en locuteurs (SRL). Une des perspectives de ce travail est de faire interagir les systèmes SVL et SRL pour étudier par exemple les bénéfices que peut apporter le regroupement en locuteurs pour la tâche de suivi.