Campagne ESTER : une première version d'un système complet de transcription automatique de la parole grand vocabulaire

Équipe SAMoVA

(**S**tructuration **A**nalyse et **Mo**délisation de la **V**idéo et de l'**A**udio)

Martine de Calmès, Jérôme Farinas,

Isabelle Ferrané, Julien Pinquier

Motivations

✧ Participation de l'IRIT à la campagne ESTER

- Test à blanc : pas de système TRS

- Evaluation janvier 2005
  - Tâche SES :
    système primaire & système contrastif
  - Tâche TRS et TTR :
    première version de notre système de transcription

✧ Ouverture vers d'autres tâches

- Système complet : indexation de documents audio vidéo,
  recherche d'entités nommées, ...

Participation à la tâche SES

✧ Système SES dit « contrastif »

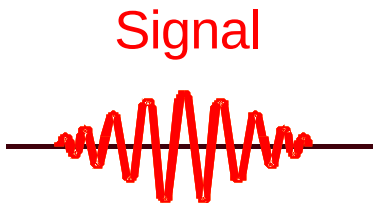
- Deux sous-systèmes de classification
  - **Sous-système 1** : détection de la parole
  - **Sous-système 2** : détection de la musique

- Paramètres
  - Modulation de l'énergie à 4Hz
  - Modulation de l'entropie
  - Nombre de segments par seconde
  - Durée des segments

- Décision
  - Maximisation des scores (vraisemblances)

Participation à la tâche SES

Système contrastif



Détection de parole

Modulation de l'entropie

Modulation de l'énergie à 4 Hz

Fusion (scores)

Classification Parole / NonParole

Détection de musique

Segmentation

Nombre de segments

Durée des segments

Fusion (scores)

Classification Musique / NonMusique

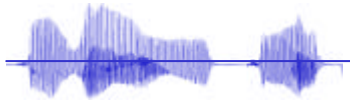
Participation à la tâche SES

✧ Système SES dit « primaire »

- Fusion de deux systèmes
  - **Système A** : système dit « contrastif »
  - **Système B** : système classique ➔ **GMM et approche spectrale**

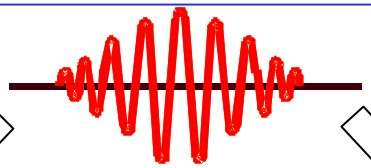
- Paramètres
  - Coefficients cepstraux (12 MFCC, Énergie et dérivées)
  - Modulations : énergie à 4 Hz et entropie
  - Coefficients spectraux (28 et Énergie)
  - Segments : nombre et durée

- Décision
  - Maximisation des scores (vraisemblances)



Participation à la tâche SES

Système primaire



Signal

Détection de parole

Détection de musique

Segmentation

Système contrastif

Système classique

Paramètres

Coefficients cepstraux

Modulation de l'entropie

Modulation de l'énergie à 4 Hz

Coefficients spectraux

Nombre de segments

Durée des segments

Modélisation

GMM

Seuil

Seuil

GMM

Seuil

Seuil

Scores

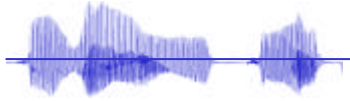
Fusion (max)

Fusion (max)

Décision

Classification parole/non-parole

Classification musique/non-musique



Participation à la tâche SES

✧ Bilan des résultats obtenus pour la tâche SES

- **Système contrastif**
  - (+) Détection de la parole (F-measure = ~0.98)
  - (-) Détection de la musique (F-measure = ~0.49)

  ➔ **seuils appris sur des données autres que celles d'ESTER**

  ➔ **problème : parole sur faible fond musical** (détectée comme zone de non-musique du fait du nombre élevé de segments)

- **Système primaire** (apport des GMM)
  - (+) Détection de la parole (F-measure = ~0.99)
  - (-) Faible apport des GMM (F-measure =~0.529)

  ➔ **scores de confiances associés aux paramètres durée et nombre de segments sont trop importants**

Participation aux tâches TRS et TTR

✧ Système de transcription – tâche TRS

- **Ressources logicielles** libres
  HTK - SLM CMU Toolkit - Julius

- **Prétraitements** (HTK)
  - MFCC : 13 coefficients avec énergie
  - Normalisation : soustraction cepstrale

- **Segmentation** (système classique IRIT)
  - détection de la parole par GMM
  - 200 ms lissage
  - durée segment max 8 mn (découpage forcé sinon)
    (+ détection de courtes pauses intégrée au décodage de Julius.)

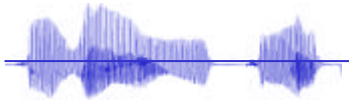
Participation aux tâches TRS et TTR

✧ Modèles acoustiques (HTK)

- Apprentissage : corpus de la phase 1 (~31 heures)
- Etiquetage phonétique de l'IRISA/ENST

- Paramètres :
  - fenêtre 16 ms
  - recouvrement 8 ms
  - fenêtrage de Hamming
  - 12 MFCC + E +delta + accel
  - normalisation de l'énergie
  - pré emphase : 0.97
  - fréquence min = 300 Hz ; max = 8000 Hz

- **Monophones** : 35 phonèmes + 2 phonèmes "silence"
  - 32 gaussiennes
  - HMM 3 états



Participation aux tâches TRS et TTR

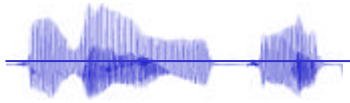
✧ Lexique Phonétique

- **Ressources lexicales :**
  BDLEX + ressource spécifique (noms propres, …)

- **Taille :**
  61 223 formes orthogr. → 119 297 variantes phon.

- **Vocabulaire :**
  transcription corpus apprentissage (~ 35 000 formes)
  textes du Monde 1987-2002 et 2003 (~ 26 500 formes)
  mots les plus fréquents (12 fois au min)

- **Taux de couverture :**
  1358 mots hors vocabulaire sur 96 003 mots
  → 1,39% corpus de développement



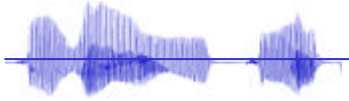
Participation aux tâches TRS et TTR

✧ Modèles de Langage (CMU SLM)

- Connaissances linguistiques
  Julius 2 passes → un modèle **bigramme** et un modèle **trigramme**
- **Vocabulaire ouvert**
- Méthodes de **prélèvement** (discount)
- Technique de **repli** (back-off)

- Sources fournies dans le cadre de la campagne :
  (1) **Ester** : transcriptions corpus d'apprentissage
  (2) **Le Monde 1987-2002** / (3) **Le Monde 2003**
  (4) MMCC (non utilisé)

- Normalisation corpus écrit → valeurs numériques, unités de mesures, site web, ...



Participation aux tâches TRS et TTR

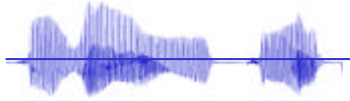
✧ Modèles de Langage

- Différentes versions des modèles bigramme / trigramme :
  - Source : (1) ; (1) + (3) ; (1) + (2) + (3)
  - Méthode de prélèvement (Good-Turing / Witten-Bell)
  - Valeur de cutoff (0, 1, 2, (B) et 0,1,2,3 (T))

- Pas d'interpolation mais un mélange des bigrammes (resp. trigrammes) issus de chaque source.

- Perplexité évaluée sur le corpus de développement
  WER non évalué → critères d 'évaluation incomplets

- Modèles utilisés : (1)+(2)+(3) – Good Turing – cutoff 0
- Perplexité : Modèles bigramme = 190,27 ; trigramme = 113,32

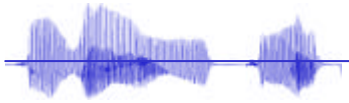


Participation aux tâches TRS et TTR

✧ Architecture logicielle, matérielle et temps de traitement

- **Logiciel** : Julius rev.3.4.2 (standard) - compilé avec gcc -O2
  - Base setup : **standard**
  - Tunings : N-gram, WordsInt, ShortWordTree, StrictIWCD2, WordPairApprox, **ShortPauseSegment**

- **Matériel** :
  (1) Sur SGI Altix 3300 (12 processeurs Itanium2 900 MHz 24 Go)
  10 processeurs utilisés (environ 1.2 Go par processus)
  ➔ ~ 11h soit 11xRT
  (2) Sur un PC dernière génération (PIV 3GHz, 2Go RAM)
  ➔ ~ 4xRT



Participation aux tâches TRS et TTR

✧ Système de transcription – tâche TTR

◆ Caractéristiques du système

➔ mêmes composantes acoustiques, lexicales et linguistiques que le système évalué pour la tâche TRS

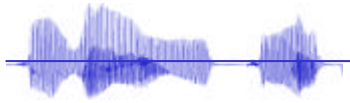
◆ Temps de traitement et architecture matérielle

- Logiciel : Julius rev.3.4.2 (standard)

- Base setup : **fast**
- Tunings : N-gram, WordsInt, WordPairApprox, ShortPauseSegment

- Matériel :

Sur un PC dernière génération : ~0.7xRT



Participation aux tâches TRS et TTR

✧ Bilan des résultats obtenus pour les tâches TRS et TTR

- **Faibles performances du système de transcription**

  TRS : ~64,5% de taux d ’erreur sur les mots

  TTR : ~70.6% de taux d ’erreur sur les mots

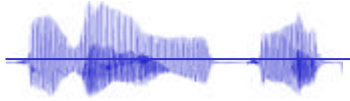
  ➔pas de système existant initialement

  ➔modèles acoustiques :

  - appris sur un tiers du corpus d ’apprentissage
  - monophones

  ➔modèles de langage :

  - pas d’évaluation préalable du taux d’erreur sur les mots
  - pas de tests contrastifs sur les différents modèles générés
  - pas d ’optimisation de la taille des modèles



Objectifs à court et moyen termes

✧ Segmentation en événements sonores

- Bon résultats sur la parole
- Améliorer les performances de la détection de musique

  ➔ réestimation des modèles

  ➔ traiter la parole propre, la parole téléphonique et la parole sur fond musical

✧ Transcription

- **Modèles acoustiques**

  ➔prendre en compte la totalité du corpus d ’apprentissage de la phase 2 de la campagne ESTER (adaptation en cours)

  ➔augmenter le nombre de gaussiennes (court terme)

  ➔modéliser les phonèmes en contexte (moyen terme)

Perspectives à court terme

✧ Transcription (suite)

- **Modélisation lexicale**
  - ➔améliorer la couverture lexicale
  - ➔ prendre en compte de l ’actualité (noms propres, …)
  - ➔utilisation de MHATLex
- **Modélisation du langage**
  - ➔interpoler les modèles
  - ➔tenir compte de la période de parution de la source
  - ➔spécialiser les modèles en fonction des radios
  - ➔ ….

Conclusion

✧ Campagne ESTER

- des ressources et une infrastructure
  - pour l'évaluation des systèmes
  - pour la stimulation des activités des recherches

  ➔ IRIT : développement d'un premier système complet de transcription

- de nombreuses perspectives d'amélioration des différents modèles
- une ouverture vers d 'autres applications et d'autres tâches

➔ **forte motivation pour participer à la prochaine campagne**