Évaluations

Sommaire

La question de l'évaluation des algorithmes constitue le dernier point essentiel pour l'expérimentateur, dans la mise en place d'un système d'indexation automatique. On peut considérer celle-ci comme la conjonction de deux éléments principaux : un critère d'évaluation quantifiable, qui permet ainsi la comparaison numérique objective des méthodes, et un corpus de test, que l'on suppose représentatif du problème évalué, et qui fournit une base commune pour la comparaison des résultats numériques. Nous commencerons par détailler les corpora exploités dans cette étude pour la classification parole/musique et la détection de chant en section 10.1, puis nous présenterons le procotole et les critères d'évaluation utilisés dans la section 10.2.

Les sections suivantes détailleront les expériences mises en place pour évaluer notre système et valider les points théoriques présentés dans ce document. Ainsi, la première expérience, section 10.3 présentera nos travaux sur le corpus ESTER 1, à travers une étude sur les taxonomies multi-classes. L'expérience décrite en section 10.4 prolongera cette dernière par l'étude des algorithmes de post-traitements. Nous présenterons ensuite en section 10.1.2 les résultats de notre participation à la campagne d'évaluation ESTER 2, puis nous finirons par valider dans la section 10.6, l'application du système développé sur la tâche de détection de chant.

10.1 Corpora audio

Le contenu du corpus de test définit clairement le cadre expérimental, par la proportion relative des classes et leur disposition dans les fichiers, ce que l'expérimentateur va considérer comme une synthèse des difficultés afférentes au problème étudié. Ainsi par exemple, un corpus dans lequel une classe est sous-représentée favorisera implicitement les autres classes. Mais la proportion adéquate des classes dépend largement de l'application visée.

De plus, la constitution d'un bon corpus de test, dont la pertinence est reconnue par la communauté, permet à celle-ci de travailler sur une base d'évaluation commune et ainsi de comparer objectivement les résultats des diverses contributions. Une telle démarche s'accompagne généralement d'un corpus d'apprentissage commun afin de restreindre la variabilité aux algorithmes de classification. Pourtant, la plupart des corpora d'évaluation de la littérature ne sont pas rendus publics, principalement en raison de la protection des droits d'auteurs. Cependant, il existe heureusement plusieurs corpora dont le contenu est partagé par leurs auteurs, et que nous exploiterons dans cette étude.

Le meilleur exemple de corpus public reste cependant celui qui accompagne une campagne d'évaluation nationale ou internationale. En effet, devant le besoin d'un cadre d'évaluation comparative exprimé par la communauté en indexation audio, ces dernières années ont vu fleurir un bon nombre de ces campagnes d'évaluation. Leur but est non seulement de fournir aux participants un corpus dont le soin apporté à la constitution et à l'annotation est hors de portée des laboratoires de recherche, mais également d'imposer un protocole d'évaluation commun qui rend possible une comparaison entre les contributions, dont les modalités sont reconnues.

10.1.1 Campagne ESTER

La campagne d'évaluation ESTER [1] (Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques) est née en 2003 de la réunion d'intérêts communs à plusieurs laboratoires de recherche dans le domaine de la transcription automatique de la parole, et a été proposée par l'AFCP (Association Francophone de la Communication Parlée). La campagne définit un cadre commun pour les différents laboratoires en concurrence, dont les systèmes sont évalués par un acteur extérieur, représenté par le Centre d'Expertise Parisien de la DGA (Délégation Générale pour l'Armement).

La majorité des tâches définies concerne la transcription et l'indexation de la parole, et couvre toute la chaîne qui permet, à partir du signal audio, et en passant par la reconnaissance de locuteur et la transcription de la parole, d'obtenir une base textuelle indexée, axée sur la catégorisation automatique en entités nommées. La première de ces tâches, nommée SES (Segmentation en Événements Sonores) concerne en toute logique la localisation des segments de parole et de musique, qui permet d'appliquer les autres traitements sur les segments identifiés.

La première édition de la campagne ESTER s'est déroulée entre 2003 et 2005, depuis la publication du protocole d'évaluation et du corpus d'apprentissage [93] jusqu'à la publication des résultats comparatifs des participants [83]. Le corpus contient un certain nombre d'heures d'enregistrements d'informations radiophoniques annotées ainsi que des transcriptions textuelles de journaux. Nous n'exploitons que les enregistrements annotés dans le cadre de cette étude. Les documents sonores proviennent des radios suivantes : France Inter, France Info, RFI (Radio France International), RTM (Radio Télévision Marocaine), dont les proportions dans les corpora d'apprentissage et de test sont résumées dans le tableau 10.1.

Bien que les annotations fournies avec le corpus soient très minutieuses, l'effort a surtout été concentré sur la transcription de la parole et la délimitation des segments de classes acoustiques présente quelques erreurs. Nous avons donc reparcouru l'intégralité du corpus d'apprentissage, à l'aide de l'outil d'annotation *Transcriber* [2], et corrigé ces erreurs, ce qui nous a permis en outre d'affiner l'annotation pour distinguer les segments de chant et de parole sur fond bruité (que nous

1. On pourra se rendre sur le site dédié de l'AFCP : `http://www.afcp-parole.org/ester/index.html` pour trouver plus d'informations sur la campagne et le corpus ESTER.

2. Transcriber (`http://trans.sourceforge.net/`) est un outil libre de segmentation, d'annotation et de transcription dont nous avons détourné l'usage habituel pour l'annotation de segments audio.

Station	**Apprentissage**	**Test**
France Inter	35h	2h
France Info	10h	2h
RFI	25h	2h
RTM	20h	2h
France Culture	-	1h
France Musique	-	1h
Total	90h	10h

TABLE 10.1 – Contenu des ensembles d'apprentissage et de test de la campagne ESTER.

désignons par ParoleBr). Le tableau 10.2 synthétise les durées cumulées de chacune des classes pour les différents sous-ensembles du corpus ESTER. Les pourcentages sous les durées précisent la proportion de chaque classe dans le sous-ensemble. Cependant, bien que nous ayons annoté à titre personnel les sous-classes en question dans le corpus de test, aucune modification n'a été apportée à ce dernier lors de l'évaluation, afin de conserver la pertinence de la comparaison aux autres participants. Les 12 minutes manquantes au total par rapport aux 90 heures de données audio sont dues au fait que certains segments ne sont pas pris en compte (silence ou classe non définie).

On reprécise le sens des classes ici mises en jeu :

- **Chant** : désigne la présence de voix chantée, a priori en présence d'un fond musical instrumental.
- **Mix** : désigne la présence de voix sur fond musical.
- **Musique** : désigne la présence de musique sans voix chantée.
- **ParoleBr** : désigne la présence de voix parlée sur fond de bruit (par exemple enregistrements en extérieur).
- **Parole** : désigne la présence de voix parlée pure.

Ensemble	**Chant**	**Mix**	**Musique**	**ParoleBr**	**Parole**	**Total**
Apprentissage	0h38	8h02	1h50	4h48	64h33	79h53
	0.8%	*10.1%*	*2.3%*	*6.0%*	*80.8%*	
Test	0h02	1h14	0h15	0h31	7h51	9h54
	0.4%	*12.5%*	*2.6%*	*5.2%*	*79.3%*	
Total	0h41	9h16	2h05	5h19	72h25	89h48

TABLE 10.2 – Répartition des classes dans les sous-ensembles du corpus ESTER

Le constat le plus frappant est la sur-représentation de la classe de parole dans le corpus, qui est également due au fait que la transcription de parole est la tâche prédominante dans la campagne. Le corpus est en effet essentiellement constitué de bulletins d'informations radiophoniques. La forte proportion de parole sur musique par rapport à la musique provient des habillages musicaux qui accompagnent couramment la voix du présentateur, notamment durant la présentation des titres. On notera enfin la proportion non négligeable de parole bruitée dans le corpus, qui n'est pas prise en compte dans la campagne ESTER, mais qui nous permettra d'apporter une analyse plus fine des résultats.

Le corpus de la campagne ESTER nous sert de point de comparaison pour l'évaluation de nos contributions. Toutefois, il convient de rappeler que, celle-ci ayant été close avant le début de cette thèse, la portée de cette comparaison est d'un impact limité puisque nous avons nécessairement tiré le bénéfice des enseignements qu'apportent les résultats des autres participants, ainsi que des annotations disponibles de l'ensemble de test, qui étaient inconnues dans les conditions réelles de la campagne. Nous avons cependant eu la chance de pouvoir participer à la seconde édition de cette campagne, que nous décrivons ci-dessous.

10.1.2 Campagne ESTER 2

La campagne d'évaluation ESTER 2 a regroupé la plupart des acteurs de la première édition, en particulier les institutions organisatrices, auxquelles se sont greffés plusieurs acteurs industriels. Elle a débuté en janvier 2008 par la mise à disposition d'un ensemble d'apprentissage et d'un autre de développement, pour l'estimation des résultats. Après la diffusion de l'ensemble de test et la campagne de test courant novembre 2008, la campagne s'est terminée en avril 2009 sur un atelier de clôture et une publication des résultats des participants [82].

Le tableau 10.3 indique la répartition du corpus audio parmi les médias et les sous-ensembles qui le constituent. Un nouveau média a été introduit dans le corpus ESTER 2, la radio Africa 1, qui se caractérise par une prise de son plus bruitée que les autres radios, et qui vient donc compliquer la tâche de classification audio. TVME est le nouveau nom de la Radio Télévision Marocaine (RTM), qui était présente dans le corpus ESTER. L'essentiel du corpus provient de la radio RFI, avec environ 70 heures d'enregistrements.

Station	**Apprentissage**	**Développement**	**Test**
France Inter	26h40	2h40	3h40
RFI	68h00	1h20	1h10
Africa 1	4h50	2h15	1h30
TVME (ex RTM)	-	1h00	1h00
Total	99h30	7h15	7h20

TABLE 10.3 – Contenu des sous-ensembles de la campagne ESTER 2.

Cette seconde édition a vu l'essor des recherches sur le sujet de la reconnaissance d'entités nommées. Toutefois, un soin supplémentaire a été apporté à l'annotation de la tâche SES, et le contenu du corpus s'est diversifié pour mieux prendre en considération les problèmes de la détection de la musique et des enregistrements bruités. On constate ainsi dans le tableau 10.4 que les parts de musique et de parole sur musique (mix) sont rehaussées en terme de durée totale (4 heures de plus de mix et 2 heures de plus de musique). Néanmoins la parole demeurent forteme majoritaire dans le corpus.

Ensemble	**Mix**	**Musique**	**Parole**	**Total**
Apprentissage	12h42	3h32	82h36	98h51
	12.8%	*3.6%*	*83.6%*	
Développement	0h22	0h08	5h34	6h04
	6.2%	*2.2%*	*91.6%*	
Test	0h22	0h26	6h12	7h01
	5.3%	*6.2%*	*88.5%*	
Total	13h27	4h06	94h23	111h57

TABLE 10.4 – Répartition des classes dans les sous-ensembles du corpus ESTER2.

10.1.3 Corpus de Scheirer

Le corpus que Scheirer a constitué en 1996 pour l'évaluation de ses travaux sur la classification parole/musique [207] est diffusé par l'auteur, et a été repris par la suite dans plusieurs publications de la communauté [195][43][5][13], dont deux publications de Ellis et de ses coauteurs [247][24] qui ont complété l'annotation originale du corpus.

Celui-ci est constitué d'un ensemble de 160 extraits de 15 secondes collectés au hasard à la radio, la moitié étant des extraits de parole pure et l'autre moitié de musique pure. Il ne contient donc pas d'extraits de parole sur fond musical.

La répartition en fichiers de classes homogèness a un impact non négligeable sur l'évaluation des résultats puisque l'absence de transition entre classes (qui constituent des zones plus difficilement caractérisables) facilite beaucoup la tâche de classification. Nous verrons ainsi que les

post-traitements les plus simples (cumul des résultats sur des fenêtres de décision longues) améliorent facilement les résultats.

La taille réduite du corpus et l'absence de test sur la classe mix limitent l'importance des résultats sur ce dernier, mais la base nous permettra avant tout de comparer nos résultats à ceux des auteurs l'ayant exploité, sur le problème de classification parole/musique.

10.1.4 Corpus Jamendo pour la détection de chant

La recherche sur le problème de la détection de chant est plus récente que la classification parole/musique et le sujet est beaucoup moins traité par la littérature. Il existe donc peu de corpora publics couvrant ce sujet. On peut citer le corpus de Holzapfel et Stylianou, constitué pour l'identification de chanteur [105], et qui sera par la suite exploité pour la détection de chant [148], qui cumule 3h12 de musique, mais se limite au genre particulier du Rembetiko (musique traditionnelle grecque).

Nous avons donc constitué un corpus [3], introduit dans [192], qui pourra, nous l'espérons, servir de base commune à la communauté pour l'évaluation de la détection de chant. Afin de pouvoir diffuser les données audio, nous avons réuni un ensemble de titres musicaux téléchargés depuis le site Jamendo [1], un site communautaire de partage de musique sous licence *Creative Commons* (c'est-à-dire libre de droits). Le corpus, d'une durée totale de 6 heures de musique, est constitué de 93 titres répartis entre les sous-ensembles d'apprentissage (61 titres), d'évaluation (16 titres) et de test (16 titres), et est constitué d'exemples de musique pop ou rock, qui constitue le genre majoritaire sur les radios généralistes. Les chansons ont été annotées avec une précision de l'ordre d'un dixième de seconde sur les frontières de segments. L'annotation du chant demeure néanmoins complexe car il existe en réalité énormément d'interruptions de la voix durant une même phrase, et certaines consonnes prolongées sont parfois très ambigües.

Le tableau 10.5 résume la répartion des classes dans les sous-ensembles du corpus. À nouveau la classe musique représente les segments sans voix chantée. Comme sur le corpus de Scheirer, on constate que les deux classes sont à peu près équilibrées sur tous les sous-ensembles, ce qui semble être un constat assez général sur la musique populaire.

Ensemble	**Chant**	**Musique**	**Total**
Apprentissage	2h05 *52.6%*	1h53 *47.4%*	3h58
Développement	0h31 *51.4%*	0h29 *48.6%*	1h00
Test	0h32 *49.6%*	0h33 *50.4%*	1h06
Total	3h09	2h55	6h05

TABLE 10.5 – Répartition des classes dans les sous-ensembles du corpus Jamendo. À nouveau Chant désigne la présence de voix chantée, sur fond musical éventuel, tandis que Musique désigne la présence de musique instrumentale, sans voix chantée.

10.2 Protocole d'évaluation

Le découpage du signal audio en une séquence de trames étant la méthode presque unanimement employée dans le domaine de la classification audio, c'est généralement le taux moyen d'erreur de classification par trames qui est employé pour évaluer les performances des algorithmes, parfois associé à la matrice de confusion, qui permet de distinguer le taux d'erreur sur les classes considérées.

3. Le corpus Jamendo est disponible à l'adresse suivante :
`http://www.telecom-paristech.fr/~ramona/icassp08/`.

Toutefois, la discrétisation de l'annotation induite par le découpage en trames suppose lors de l'évaluation, l'homogénéité en termes de classe sur chaque trame. L'impact est généralement minime mais peut devenir non-négligeable lorsque les trames de décision atteignent une taille de l'ordre de plusieurs secondes, puisque certains segments peuvent alors être ignorés lors de l'évaluation. De plus, l'usage d'un tel critère pour une évaluation comparative implique d'imposer le même pas d'avancement à tous les systèmes.

Le protocole des campagnes d'évaluation ESTER 1 et 2 exploite une alternative qui permet en plus d'évaluer le cas des classes se chevauchant. En effet, bien que nous ayons justifié la pertinence pour la classification d'un problème à trois classes, dont l'une est la superposition des deux autres (voir la section 2.2), le protocole d'évaluation ESTER, que nous suivons dans ce document, se base sur un problème à deux classes (parole et musique) pouvant se chevaucher. Ce chevauchement constitue ce que nous avons défini comme la classe mix.

Ainsi, comme le montre la figure 10.1, on réunit dans un premier temps les trames successives en segments, puis l'on sépare ces derniers en autant de séquences de segments que de classes, en tenant compte des classes superposées.

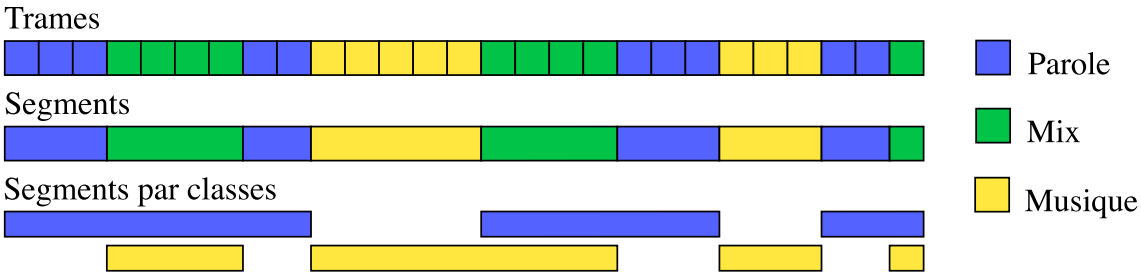


FIGURE 10.1 – Conversion des résultats à 3 classes sur les trames vers les problèmes sur segments pour chaque classe.

Par la suite le protocole suivant est appliqué sur chaque classe (voir la figure 10.2) :

- On inclut d'abord entre chaque paire de segments consécutifs de la classe un segment de « non-classe » qui représente l'absence de cette classe.
- On applique ensuite une tolérance aux frontières qui permet d'ignorer lors de l'évaluation les décalages par rapport à la frontière réelle d'une durée inférieure à un seuil τ_{tol} qui est fixé à 0.25 s dans les campagnes ESTER. Concrètement on exclut de l'évaluation les τ_{tol} secondes qui précèdent et suivent chaque frontière réelle.
- On reporte sur les segments estimés les zones ignorées par la tolérance, et l'on introduit également la « non-classe » sur les zones non ignorées qui ne sont pas estimées dans la classe.
- On extrait de la comparaison des segments pré-traités réels et estimés, les segments de bonne classification, de fausse alerte (faux positif) et de détection manquée (faux négatif).

On calcule ainsi les grandeurs d_{OK}, d_{FA} et d_{DM} respectivement définies comme les durées cumulées des segments corrects, de fausse alerte et de détection manquée. On définit également les durées cumulées de segments estimés $d_{\mathrm{EST}} = d_{\mathrm{OK}} + d_{\mathrm{FA}}$ et de segments réels $d_{\mathrm{REEL}} = d_{\mathrm{OK}} + d_{\mathrm{DM}}$,

Les valeurs de rappel R et de précision P sont alors calculées de la manière suivante :

$$R = \frac{d_{\mathrm{OK}}}{d_{\mathrm{OK}} + d_{\mathrm{DM}}}, \qquad P = \frac{d_{\mathrm{OK}}}{d_{\mathrm{OK}} + d_{\mathrm{FA}}},$$

ce qui revient à définir le rappel et la précision comme le temps cumulé de détection correction sur le temps où, respectivement, la classe est réellement présente et où la classe est détectée.

La F-mesure, qui représente le critère global d'évaluation, est définie comme la moyenne harmonique des deux précédentes mesures :

$$F = \frac{2RP}{R+P}.$$

Elle constitue ainsi un compromis entre les deux, dont l'effet est beaucoup plus pénalisant que la moyenne arithmétique si l'une des valeurs est particulièrement faible (puisque $F = 0$ si $R = 0$ ou

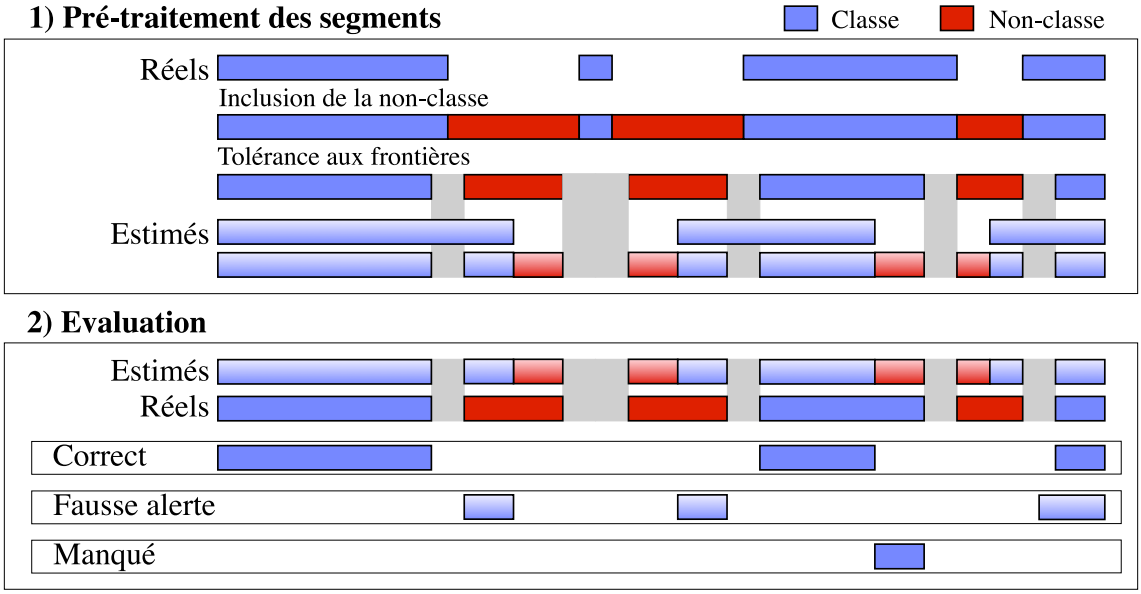


FIGURE 10.2 – Post-traitement des segments pour la prise en compte de la tolérance aux frontières et extraction des segments corrects, de fausses alertes, et manqués, pour le calcul des critères d'évaluation.

$P = 0$).

Les résultats de la campagne ESTER mentionnent également, comme mesures d'erreur, les classiques taux de fausse alerte FA $= \frac{d_{\text{FA}}}{d_{\text{EST}}}$ (ou faux positif) et de faux rejet MD $= \frac{d_{\text{DM}}}{d_{\text{REEL}}}$ (ou faux négatif).

10.3 Expérience 1 : comparaison des taxonomies

Nous comparons, dans cette première expérience, différentes taxonomies de classification multi-classes pour le problème de la classification parole/musique, synthétisées dans la figure 10.3. Nous étudions par ailleurs l'influence de la prise en compte des trames de chant sur les performances.

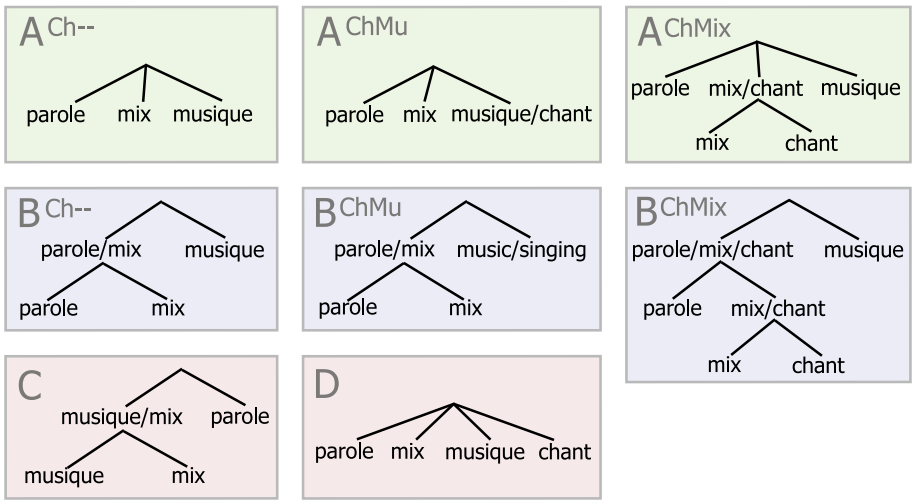


FIGURE 10.3 – Représentation hiérarchique des huit taxonomies multi-classes comparées dans l'expérience 1.

Nous proposons ici deux taxonomies principales dont l'une (A) est basée sur un cadre non-

hiérarchique de combinaison *one-vs-one*, et l'autre (B) sur une approche hiérarchique d'arbre binaire de classification, qui consiste d'abord à séparer la musique des exemples contenant de la parole, puis à distinguer parmi ces derniers les exemples de parole pure et de mix.

Les suffixes *Ch–*, *ChMu* et *ChMix* désignent des variantes qui différent par leur prise en compte des exemples de chant. Dans les taxonomies *Ch–*, les exemples de chant ne sont pas pris en compte lors de l'apprentissage ; la classe musique ne contient alors que de la musique instrumentale sans voix chantée. Cette première approche est basée sur l'hypothèse que les exemples chant sur fond musical constituent une source de confusion possible avec la classe mix (en supposant que le chant est proche de la voix parlée). Dans la variante *ChMu*, on conserve, sans modification, les exemples contenant de la voix chantée dans l'ensemble des exemples de musique. Le chant n'est bien sûr pas considéré comme une classe supplémentaire lors de la phase d'évaluation. Enfin dans l'approche *ChMix*, partant de la proximité supposée du chant avec le mix, on associe dans un premier temps ces deux classes pour les séparer par la suite (les exemples de chant seront ensuite associés à la classe musique lors de l'évaluation).

La taxonomie C est une variante de l'approche hiérarchique qui consiste d'abord à séparer la parole pure de tout signal contenant de la musique. Celle-ci est toutefois beaucoup moins intuitive que la B parce que la musique est généralement en retrait par rapport à la parole dans les exemples de mix.

Enfin, la taxonomie D se base sur un paradigme *one-vs-one* incluant également la classe de chant (dont les exemples détectés sont par la suite associés à la classe de musique).

Les nœuds binaires des arbres représentés impliquent l'application simple d'un SVM discriminatif sur les classes des fils. Nous avions de plus mentionné dans la section 5.1.5.3 la possibilité de définir un cadre hybride en introduisant des noeuds non-binaires dans un arbre de classification, traités par une approche *one-vs-one*, que nous retrouvons dans la taxonomie A *ChMix*. Les labels « classe1/classe2 » désigne une classe formée pour l'apprentissage par l'union des exemples des deux classes.

Cette étude est basée sur le corpus ESTER 1 (section 10.1.1), sur lequel nous avons annoté les occurrences de voix chantée. Nous utilisons les descripteurs présentés dans la section 6.5, dont nous sélectionnons par l'algorithme IRMFSP (présenté dans la section 7.3.2) les d plus pertinents. Les SVM exploitent un noyau RBF gaussien, dont nous discutons l'affinage ci-dessous, et l'apprentissage est effectué sur un maximum de 20000 exemples par classe. L'application des taxonomies multi-classes se base sur l'algorithme d'estimation pondérée des probabilités a posteriori, proposé en section 5.1.5.4, lesquelles sont lissées par un filtrage médian (voir section 8.2).

10.3.1 Affinage du noyau par la mesure d'Alignement

Nous commençons par comparer les procédures de recherche par maillage (section 4.2.1) et d'optimisation par maximisation du critère d'Alignement, introduit en section 4.4.1 pour l'affinage du paramètre σ du noyau. Dans cette partie, le nombre de descripteurs sélectionnés est fixé arbitrairement à $d = 20$. La recherche par maillage est effectuée sur un ensemble de 12 valeurs logarithmiquement réparties entre 0.2 et 15. Les SVM impliquées dans les taxonomies sont donc apprises pour chacune de ces valeurs ; la valeur de σ maximisant la F-mesure globale sur l'ensemble de validation est choisie.

La figure 10.4 montre, pour chaque taxonomie, les F-mesures calculées après affinage du noyau sur l'ensemble de test pour les deux méthodes (la recherche par maillage est en teintes foncées et l'optimisation sur l'alignement en couleurs claires), avec et sans post-traitement par lissage médian (respectivement en vert et en bleu). Le dernier sous-histogramme indique la moyenne sur toutes les taxonomies.

Il est clair qu'en l'absence de post-traitement, la recherche par maillage se montre plus efficace que l'alignement pour l'affinage du noyau. Toutefois, le constat s'inverse lorsque l'on applique le filtrage médian. Ceci s'explique par le fait que l'alignement apporte un meilleur affinage pour quelques-uns des discriminateurs impliqués dans la taxonomie. Le filtrage médian corrige alors efficacement les erreurs accidentelles (sur une ou deux trames adjacentes) et compense ainsi le léger désavantage (de l'ordre de quelques dizièmes de pourcent sur la F-mesure) des SVM affinés par l'alignement.

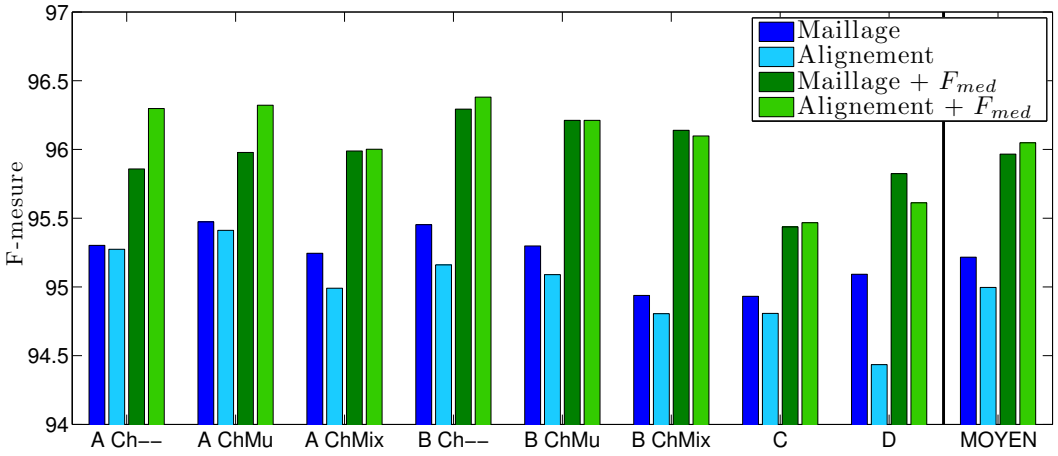


FIGURE 10.4 – Comparaison des résultats (en F-mesure) sur les différentes taxonomies après affinage des noyaux par recherche par maillage ou par optimisation de l'alignement, avec et sans filtrage médian.

On confirme donc que l'affinage du paramètre par la maximisation de l'alignement permet d'obtenir des performances comparables à la recherche par maillage, à un coup fortement réduit, puisqu'une seule opération d'apprentissage de SVM est nécessaire. Ce résultat est pour nous essentiel, parce qu'il permet la mise en place d'un système de classification audio dont l'apprentissage est entièrement automatisé et ne nécessite pas de corpus de validation.

10.3.2 Résultats sur le corpus ESTER 1

La figure 10.5 montre l'évolution de la F-mesure avec le nombre d de descripteurs sélectionnés, pour chacune des huit taxonomies présentées précédemment.

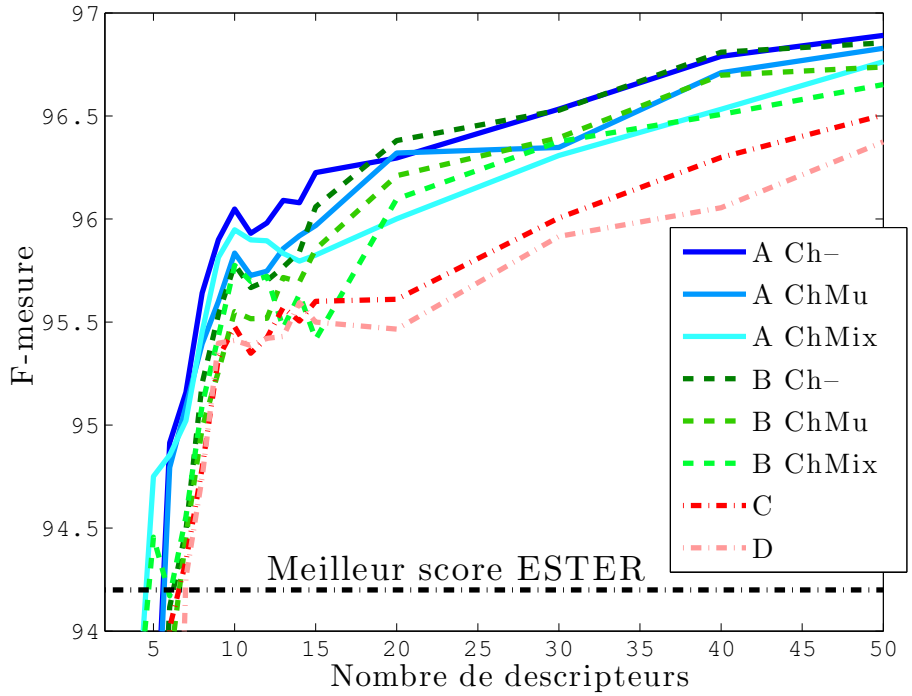


FIGURE 10.5 – Evolution de la F-mesure pour les 8 taxonomies multi-classes en fonction de la dimension d du vecteur de descripteurs, et comparaison au meilleur résultat de la campagne ESTER 1.

	globale			**parole**			**musique**		
Participant	F	%fa	%fr	F	%fa	%fr	F	%fa	%fr
$d = 50$	96.9	2.0	4.5	99.4	13.0	0.5	78.8	1.5	29.6
$d = 10$	95.9	3.3	5.4	99.1	19.4	1.0	73.8	2.5	33.2
$d = 2$	93.3	11.9	4.1	98.9	16.2	1.5	64.8	11.6	20.3
1er ESTER	94.2	2.1	9.5	98.8	30.1	1.5	52.7	1.2	61.7
2e ESTER	93.1	1.3	12.1	98.9	9.7	1.9	33.7	1.0	78.5
3e ESTER	92.7	11.7	5.7	99.2	36.6	0.7	54.8	10.9	38.7
4e ESTER	90.7	1.3	16.2	97.4	8.0	4.9	17.8	1.1	89.6

TABLE 10.6 – Performances des participants à la campagne ESTER 1 pour la tâche SES de segmentation.

On remarque en premier lieu que les configurations A et B (en traits pleins, respectivement bleus et verts) sont sensiblement plus efficaces que les configurations C et D (en pointillés rouges et roses), de manière presque uniforme sur l'ensemble des dimensions d. En effet, la taxonomie C est pénalisée par l'union de deux classes trop éloignées acoustiquement (mix et musique), tandis que dans la D, la classe de chant, trop peu fournie en exemples, est trop faiblement caractérisée et réduit ainsi les performances globales. L'écart avec les approches A et B reste limité à moins de 1% mais, sur un score de 96%, l'avantage de ces dernières représente une réduction relative de 25% sur l'erreur, ce qui confirme l'importance du choix de la meilleure taxonomie.

En revanche, on ne remarque pas d'écart notable entre les taxonomies A et B, à part à basses dimensions ($d < 20$) où l'approche *one-vs-one* (A) se montre plus efficace que l'approche hiérarchique. De plus, l'influence de la classe de chant est très similaire sur les deux cas. A haute dimension ($d > 15$), l'absence des exemples de chant est la plus profitable, tandis que leur inclusion dans une classe impliquant les exemples de mix se révèle moins efficace que l'union plus naturelle chant/musique. Toutefois, à dimension très basse ($d < 7$), la première union (chant/mix) devient plus efficace, probablement parce que la diversité apportée par les exemples de chant compense en partie les défauts de caractérisation dus au faible nombre de descripteurs.

La F-mesure augmente avec la dimension du vecteur de descripteurs, approchant asymptotiquement les 97%. Toutes les taxonomies testées dépassent, pour $d > 7$, le meilleur score obtenu durant la campagne ESTER 1 (indiqué par la ligne noire en pointillés, et égal à 94.2%), ce qui montre l'efficacité du système proposé pour cette tâche. On note même que certaines taxonomies (A *ChMix* et B *ChMix*) demeurent d'ailleurs efficaces à très basse dimension ($d = 5$), avec une F-mesure autour de 94.5%. Ainsi, pour une complexité raisonnable ($d = 10$), la meilleur et la pire taxonomie apportent respectivement un gain absolu de 2% et 1.3% sur le meilleur résultat d'ESTER. Tous les systèmes proposés dans le cadre de la campagne étaient basés sur les descripteurs MFCC et leurs dérivées premières et secondes, pour une dimension entre 33 et 40. L'usage de techniques de sélection de descripteurs apporte donc ici un réel avantage en termes de performances et de complexité.

Le tableau 10.6 détaille les résultats des trois meilleurs participants à la campagne ESTER 1, et les confronte à ceux de la taxonomie la plus efficace (A *Ch–*) sur différentes dimensions d. On remarque que même à très faible dimension ($d = 2$), la taxonomie choisie surpasse le deuxième participant, ce qui confirme à nouveau la pertinence de la sélection de descripteurs adjointe à l'emploi des SVM. L'amélioration la plus notable concerne la détection de la musique, sur laquelle le système proposé apporte un gain absolu de 12 à 26%, dû principalement à une forte réduction du taux de fausse alerte (colonne %fa). Ceci s'explique par le fait que sur la plupart des autres systèmes, l'accent a été mis sur la bonne reconnaissance des regions de parole (en raison de l'importance des autres tâches sur la parole dans la campagne).

10.4 Expérience 2 : post-traitements

Nous poursuivons l'étude précédente sur le corpus ESTER 1 pour montrer les effets des algorithmes de post-traitement dynamiques présentés dans la partie III.

Le système exploité ici est identique au système correspondant aux résultats de la seconde ligne ($d = 10$) du tableau 10.6, basé sur la taxonomie A *Ch–*, soit une approche *one-vs-one* sur les trois

classes de parole, de mix et de musique, les exemples de chant étant exclus de cette dernière classe.

Nous comparons ici les gains en performances apportés respectivement par le filtrage médian (qui est appliqué dans l'expérience précédente), le post-traitement par HMM proposé dans la section 8.3.2.2, ainsi que les 5 algorithmes hybrides basés sur un principe de détection de rupture, présentés dans le chapitre 9.

Nous avons vu dans la section 9.6, que la détermination des frontières de segments avec ces dernières approches, se base en définitive sur l'application d'un seuil empirique τ (voir section 9.6) qui fixe l'habituel compromis entre frontières fausses et manquées (faux positifs et faux négatifs). La valeur optimale du seuil est ainsi déterminée en recherchant le maximum de la F-mesure globale calculée après application du post-traitement sur l'ensemble de validation. La figure 10.6 montre l'évolution de la F-mesure en fonction du seuil τ (échelonné entre -1 et 3), pour les cinq métriques proposés, à savoir le critère BIC (*Bayesian Information Criterion*) en vert, les mesures LLR (*Log Likelihood Ratio*) et KCD (*Kernel Change Detection*), toutes deux basées sur les SVM à une classe, respectivement en rose et rouge, et enfin les mesures DIV (Divergence de Kullback-Leibler) et BAT (Distance probabiliste de Bhattacharyya) en bleu clair et foncé, toutes deux calculées dans l'espace RKHS, espace image de la transformation implicite appliquée par le noyau. Toutes les mesures de détection de rupture sont calculées sur deux fenêtres glissantes de 9 trames longues (qui correspondent chacune à 5 secondes de signal).

Les résultats mesurés sans post-traitement et avec le filtrage médian sont respectivement indiqués par les lignes pointillées noire et grise.

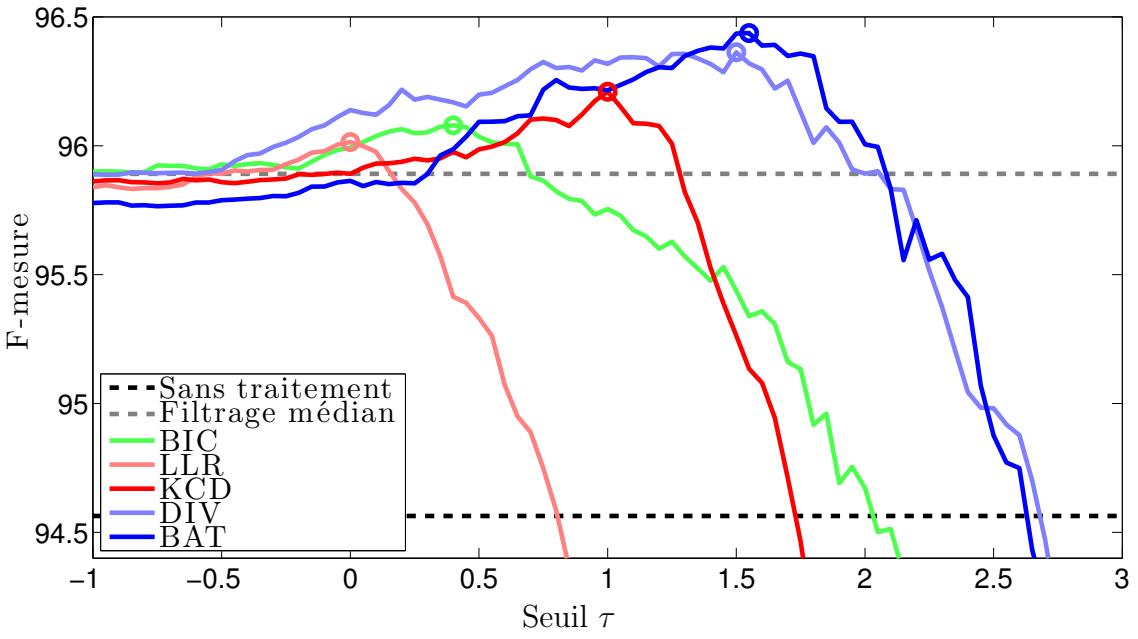


FIGURE 10.6 – Évolution de la F-mesure globale sur l'ensemble de validation par rapport au seuil τ, pour chacune des mesures de détection de rupture.

On remarque en premier lieu que les cinq courbes suivent toutes un profil similaire. Lorsque le seuil est très faible ($\tau < 0$), les métriques ont à peu près les mêmes performances, par ailleurs proches de celles du filtrage médian. En effet, le nombre de maxima dans les courbes de détection de rupture étant limité (du fait de la procédure de recherche, présentée dans la section 9.6) en deçà d'un certain seuil, tous les pics sont retenus et l'algorithme n'évolue plus. On est alors dans une situation de « sur-segmentation », où le traitement est appliqué sur une série de segments très courts, dont l'échelle avoisine celle de la fenêtre de filtrage médian, ce qui explique la proximité observée entre les différentes approches.

Lorsque le seuil τ augmente, les performances augmentent de manière quasi monotone pour atteindre un maximum (indiqué sur la figure par un cercle, pour chaque métrique), au delà duquel celles-ci chutent brutalement. En effet, plus on augmente le seuil, plus le nombre de maxima est

restreint et plus les segments sont larges. Ainsi lorsque l'on dépasse le seuil optimal, on fusionne alors les classes de segments de plus en plus importants, ce qui explique que les effets s'amplifient rapidement. Bien sûr si l'on pousse le seuil à l'extrême on ne détecte plus aucune frontière, ce qui revient à assigner la même classe à l'ensemble du signal audio, à priori la classe de parole, puisque celle-ci est majoritaire. On n'aura donc pas un résultat nul, mais très pénalisé.

La figure montre sans équivoque la supériorité des métriques basées sur les distances probabilistes dans l'espace RKHS, par rapport aux autres métriques. Nous rappelons que les fenêtres glissantes ont une largeur de 9 trames, ce qui signifie que les SVM à une classe impliqués dans les métriques KCD et LLR, effectuent l'apprentissage de la classe sur ces seules 9 trames. Ainsi le critère LLR, qui n'implique en réalité qu'un seul SVM (appris sur la fenêtre passée et évalué sur la fenêtre future), montre sa faiblesse par rapport aux autres métriques, du fait d'une caractérisation si réduite. Le critère KCD, qui implique bien deux SVM à une classe, fournit de meilleurs résultats, mais on peut supposer que les 9 exemples sont insuffisants pour caractériser assez précisément l'axe des hyperplans délimitant les classes, qui définissent le critère lui-même. Ainsi, les métriques RKHS, qui n'impliquent que la modélisation gaussienne des exemples dans l'espace RKHS se comportent beaucoup mieux par rapport au faible nombre d'exemples.

Il est important de préciser que le choix des longueurs de fenêtre n'est pas arbitraire et résulte d'une détermination empirique sur l'ensemble de validation, que nous ne détaillons pas ici, qui traduit le compromis entre précision du modèle et précision temporelle des fenêtres modélisées. En effet, l'augmentation de la largeur des fenêtres implique que celles-ci ont plus de chance d'inclure des changements de classes, qui viennent compenser l'effet bénéfique sur la modélisation.

	globale			**parole**			**musique**		
Système	**F**	**%fa**	**%fr**	**F**	**%fa**	**%fr**	**F**	**%fa**	**%fr**
Sans traitement	94.56	5.68	6.15	98.73	15.70	1.81	67.84	5.16	33.59
Filtre médian	95.89	3.34	5.38	99.06	19.44	0.98	73.80	2.51	33.17
HMM à 2 gauss.	95.45	5.03	4.96	99.01	10.90	1.46	72.98	4.72	27.09
HMM à 5 gauss.	96.04	4.07	4.55	99.08	**8.31**	1.44	76.67	3.85	**24.21**
HMM à 10 gauss.	96.15	3.18	5.00	99.12	9.88	1.29	76.26	2.83	28.43
Hybride BIC	96.08	3.40	4.97	99.16	23.49	0.59	74.48	2.37	32.66
Hybride LLR	96.01	3.55	4.99	99.11	22.83	0.73	74.54	2.56	31.93
Hybride KCD	96.21	3.15	4.92	99.16	22.61	0.64	75.51	2.15	31.93
Hybride DIV	96.36	3.34	**4.48**	99.10	29.92	**0.43**	**77.19**	1.97	30.09
Hybride BAT	**96.44**	**2.93**	4.65	**99.22**	22.68	0.50	76.81	**1.91**	30.86

TABLE 10.7 – Comparaison des résultats obtenus sur le corpus ESTER avec les différents paradigmes de post-traitement présentés (filtrage médian, HMM et segmentation aveugle).

Le tableau 10.7 compare les résultats sans post-traitement et après application d'un filtrage médian, d'un lissage HMM, ou des métriques de segmentation aveugle proposées. Les critères de F-mesure, fausse alerte (%fa) et faux rejet (%fr) sont indiqués pour les deux classes parole et musique, ainsi que pour l'ensemble des segments (classe « globale »).

On constate en premier lieu que les différences en terme de F-mesure globale sont assez réduites (le gain maximal absolu est de 1.88% avec l'approche hybride Bhattacharrya, soit tout de même une réduction relative de l'erreur d'environ 35%), ce qui montre sans surprise que les effets du post-traitement ne peuvent se substituer au soin à apporter à la mise en place et à l'affinage du système de classification. Les erreurs accidentelles (de l'ordre d'une ou quelques trames consécutives) sont facilement corrigées en considérant les résultats proches dans le temps, mais les erreurs structurelles (une classe mal apprise ...) sont trop étalées dans le temps pour satisfaire le formalisme des méthodes de post-traitement. Cependant, bien que la F-mesure constitue le critère global d'évaluation, les différences sont plus marquées et plus facilement interprétables sur les autres critères présentés.

On constate que les deux méthodes de post-traitements proposées dans cette thèse (HMM et hybride par segmentation) améliorent les performances par rapport au filtrage médian. L'augmentation de la complexité des modèles de probabilités des HMM (c'est-à-dire le nombre de gaussiennes

des modèles GMM) favorise, sans surprise, les résultats avec cette méthode ; toutefois, au delà de 10 gaussiennes, on ne note pas d'amélioration notable.

Bien que les écarts entre les trois approches sont ténus, ils restent cependant significatifs au regard du volume de la base de test, et se traduisent surtout dans les critères de fausse alerte et de faux rejet. Ainsi on remarque que les HMM réduisent fortement les taux de fausse alerte sur la parole et de faux rejet sur la musique, ce qui signifie que la part de trames de musique prises pour de la parole est fortement réduite par ce post-traitement. Les approches hybrides par segmentation aveugle sont caractérisées par la correction opposée, à savoir à la réduction du nombre de trames de parole prises pour de la musique, et donc des taux de fausse alerte en musique et de faux rejet en parole. La parole étant largement majoritaire dans le corpus, l'effet des approches par segmentation est donc plus efficace que celui du post-traitement par HMM, qui prend pourtant en compte la proportion des classes dans l'apprentissage de la matrice de transition $\boldsymbol{A}$ (voir section 8.3.1).

Confirmant les observations sur la figure 10.6, les métriques probabilistes dans l'espace RKHS (DIV et BAT) sont les plus efficaces et apportent un gain absolu de 0.5% environ sur le filtrage médian, soit une réduction relative de l'erreur de 12%.

10.5 Résultats à ESTER 2 et sur le corpus de Scheirer

Campagne ESTER 2

Nous avons également participé à la campagne d'évaluation ESTER 2, pour laquelle nous avons appliqué le système présenté dans la première expérience (sur ESTER 1) en exploitant la taxonomie A *Ch–*, c'est-à-dire un paradigme *one-vs-one* sur les trois classes de parole, musique et mix. Nous avons également exclu les exemples de chant des données d'apprentissage de la classe musique. Le nombre de descripteurs sélectionnés est ici fixé à $d = 50$, et la sélection exploite l'algorithme IRMFSP, nos recherches sur les algorithmes de sélection de descripteurs n'ayant pas encore abouti à l'époque de la campagne de test d'ESTER 2. Les résultats sur la tâche SES sont résumés dans le tableau 10.8, et proviennent en partie de la publication de clôture de la campagne [82].

Malheureusement dans les résultats publics de la campagne, seuls les taux d'erreur, de détection manquée (md), de fausse alerte (fa) et de F-mesure pour chaque classe ont été publiés. Les mesures de rappel et de précision ne sont pas disponibles, et la F-mesure globale, qui servait de critère de référence pour la campagne ESTER 1, et qui évaluait le compromis entre les détections de parole et de musique, ne fait plus partie du protocole d'évaluation, principalement parce que certains participants n'ont publié des résultats que sur la détection de parole (en raison de sa proéminence dans le corpus, et sur les autres tâches). Ainsi, l'IRIT est notre seul concurrent sur la tâche de détection de la musique. On remarquera enfin dans le tableau que les taux d'erreur ont également été renseignés sur les données propres à chaque station de radio du corpus.

Système	**Erreur(%)**					**md(%)**	**fa(%)**	**F**
	Africa	**Inter**	**RFI**	**TVME**	**Globale**			
Classe de parole								
IRISA	1,65	1,42	0,58	2,44	1,49	**0.37**	16.42	99,20
IRIT	2,05	0,85	0,65	2,47	1,31	0.72	9.28	99,29
LIMSI	2,55	**0,52**	**0,26**	**1,71**	**1,08**	0.80	**4.91**	**99,42**
RTL	**1,40**	1,10	0,61	2,07	1,23	0.50	11.01	99,34
Classe de musique								
IRIT	**6,63**	5,17	5,93	4,63	5,51	43.13	**0.77**	69,80
TPT/RTL	12,40	**2,95**	**3,92**	**4,10**	**5,25**	**12.56**	4.33	**78,85**

TABLE 10.8 – Comparaison des résultats des participants à la tâche SES de la campagne ESTER 2. Notre système est désigné par TPT/RTL (pour TELECOM ParisTech / RTL).

Les résultats que nous obtenons sur le corpus ESTER 2 (participant TPT/RTL pour TELECOM ParisTech / RTL) sont très proches de ceux mesurés sur ESTER 1, ce qui souligne la proximité entre les deux corpora. Le meilleur résultat sur la détection de la parole est obtenu

par le système du LIMSI, tant en termes de F-mesure que de taux d'erreur et de fausse alerte, ainsi que sur la majorité des stations de radio du corpus. On notera toutefois que les F-mesures des quatre participants sont assez proches (les écarts mutuels étant de l'ordre de 0.1%), et notre système obtient le résultat le plus proche du meilleur participant. Le taux de détections manquées est d'ailleurs inférieur, même si notre taux de fausse alerte est significativement plus élevé. Il est difficile cependant de comparer les deux systèmes, en raison de notre participation conjointe à la tâche de détection de musique. En effet, l'optimisation du système résulte d'un compromis entre trois classes, en comptant la classe mix, tandis que l'optimisation des résultats de détection de parole n'implique que deux classes (parole et non-parole) et simplifie ainsi le problème.

Les écarts de performances sur la tâche de détection de musique sont d'ailleurs beaucoup plus marqués que dans le cas précédent, et notre système marque sur le second participant une avance nette en termes de F-mesure (+9% en absolu, et 30% de réduction relative d'erreur) et de taux de détection manquée (65% de réduction relative). En contrepartie, l'IRIT présente un taux de fausse alerte largement réduit par rapport au notre (82% de réduction relative), ce qui semble indiquer que notre système modélise une classe de musique plus étendue que celle modélisée par l'IRIT. L'inversion de ce constat sur la classe de parole confirme cette intuition.

En définitive, il est difficile d'analyser plus en profondeur les résultats sur la campagne ESTER 2, en raison du manque de participants à la tâche de détection de musique.

Corpus de Scheirer

S'il existe peu de corpora publics pour l'évaluation de la classification parole/musique (à l'exception des campagnes d'évaluation), le corpus de Scheirer, que nous avons introduit dans la section 10.1.3, est l'un des plus cités et des plus repris dans la littérature. Ainsi, il constituera une troisième opportunité d'évaluer notre système, et par la même occasion, de le comparer aux publications internationales, contrairement à la portée des campagnes ESTER, qui reste nationale.

Toutefois il est difficile de comparer directement les résultats des auteurs en raison de divergences dans les protocoles d'évaluation suivis. La démarche fixée à l'origine par Scheirer et Slaney [207] consiste à diviser le corpus audio en un ensemble de test contenant 10% des 160 fichiers (soit 16 fichiers) et un ensemble d'apprentissage contenant les 90% restant. Les fichiers ne sont pas découpés entre les deux bases de manière à ne pas bénéficier des similarités au sein d'un même extrait. Le taux d'erreur est évalué en reproduisant un grand nombre de fois ce processus, le découpage étant aléatoire à chaque itération. Les auteurs publient principalement les résultats obtenus sur les trames d'une seconde (qui correspondent à notre trame long-terme), en mentionnant en plus les résultats calculés en étendant la fenêtre de décision à 2.5 s, de manière à reproduire le protocole qu'applique Saunders dans son article [205]. Williams et Ellis [247] poussent cette dernière idée à l'extrême en publiant les résultats obtenus en étendant la fenêtre de décision à l'ensemble de l'extrait de 15 s, mais ne publient malheureusement pas de résultats sur les trames d'une seconde, contrairement à Casagrande et al. [43], qui se limitent justement à cette échelle.

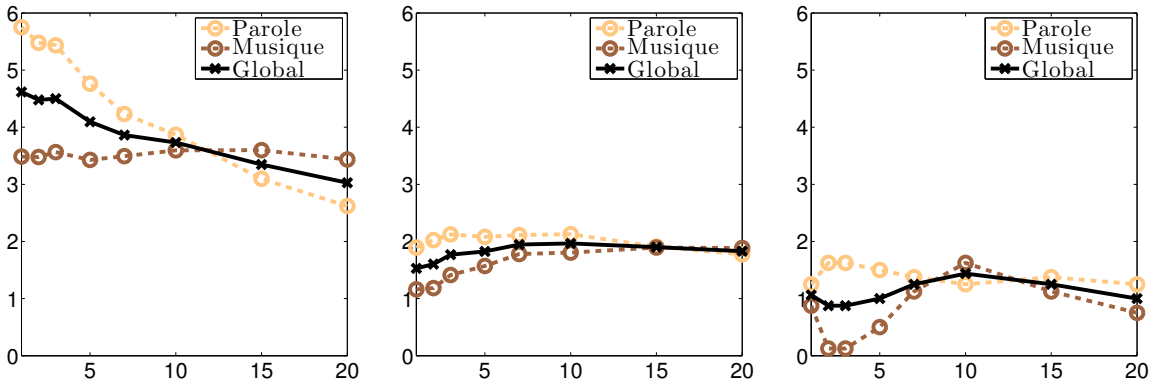
Fenêtre de décision			**Trame longue (1 s)**			**2.5 s**	**Tout 15s**
Système	**Dim**	**Ref**	**Parole**	**Musique**	**Global**	**Global**	**Global**
Saunders	-	[205]	-	-	-	2.0	-
Scheirer & Slaney	3	[207]	6.7±1.9	4.9±3.7	5.8±2.1	1.4	-
	8		6.2±2.2	7.3±6.1	6.7±3.3	-	-
Casagrande et al.	-	[43]	-	-	6.7	-	-
Williams & Ellis	3	[247]	-	-	-	1.3	0.0
	4		-	-	-	1.7	0.0
Ramona	1		5.7±3.9	3.5±2.5	4.6±2.3	1.5±2.2	1.0±2.3
	20		2.6±3.7	3.4±3.1	3.0±2.2	1.8±2.5	1.0±2.3

TABLE 10.9 – Comparaison des résultats de la littérature sur le corpus de Scheirer (à l'exception de Saunders qui est donné ici à titre indicatif), pour différentes longueurs de fenêtres de décision. L'écart type des résultats sur les itérations est précédé du symbole ±.

Le tableau 10.9 synthétise l'ensemble des résultats des auteurs mentionnés. Nous indiquons le résultat de Saunders, mesuré sur des fenêtres de décision de 2.5 s, bien que celui-ci ne soit pas

calculé sur le même corpus. Scheirer et Slaney, ainsi que Williams et Ellis, publient des résultats pour différentes combinaisons de descripteurs, en faisant varier le nombre de descripteurs sélectionnés. Nous indiquons ici les résultats les plus significatifs des deux auteurs. En revanche, dans la contribution de Casagrande et al. la sélection des descripteurs pertinents fait partie du processus de classification et ne constitue pas une variable du protocole expérimental. Le système que nous appliquons sur ce corpus est constitué d'un unique classifieur SVM, dont le noyau est sélectionné par minimisation du critère d'alignement, appris sur un sous-ensemble des descripteurs proposés, sélectionnés par l'algorithme SAS (minimisation de l'alignement sur noyau pondéré) que nous avons présenté dans la section 7.5.1. La taille restreinte du corpus est ici propice à l'application de cette méthode qui, bien qu'elle soit plus légère que ses concurrentes, reste très coûteuse sur un grand nombre d'exemples. Nous mentionnons dans le tableau les résultats obtenus avec le meilleur descripteur et avec les 20 meilleurs descripteurs. Nos résultats sont estimés sur 100 itérations.

Les résultats sur les trames d'une seconde sont les plus significatifs puisqu'ils valident sans ambiguïté la pertinence du système de classification employé ainsi que l'algorithme de sélection de descripteurs. En effet, avec un unique descripteur, notre système réduit l'erreur à 4.6%, ce qui représente un gain relatif de 20% sur l'erreur minimale de 5.8% de Scheirer de Slaney ; le système de Casagrande et al. ne présente qu'une erreur minimale de 6.7 %. L'erreur est d'ailleurs encore réduite si l'on augmente le nombre de descripteurs et descend jusqu'à 3% pour $d = 20$. La figure 10.7(a) montre l'évolution de l'erreur globale ainsi que pour chaque classe, par rapport au nombre de descripteurs sélectionnés. On observe que la décroissance de l'erreur globale est due à la décroissance de l'erreur sur la parole, tandis que la détection de la musique semble être essentiellement due au premier descripteur sélectionné, et évolue peu lorsque l'on augmente la dimension.



(a) 1 seconde (b) 2.5 secondes (c) 15 secondes

FIGURE 10.7 – Évolution de l'erreur globale et des erreurs de classes par rapport au nombre de descripteurs sélectionnés, pour notre système. Les trois figures correspondent aux différentes longueurs de fenêtres de décision impliquées dans le protocole expérimental.

Les figures 10.7(b) et 10.7(c) montrent également l'évolution des erreurs pour les fenêtres de décisions respectives de 2.5 s et de 15 s (c'est-à-dire la totalité des extraits). Si l'élargissement de la fenêtre de décision réduit sans surprise l'erreur, on constate par contre avec étonnement que l'augmentation de la dimension devient un handicap pour la classification. Ainsi on obtient de meilleurs résultats avec un seul descripteur (1.5%) sur les fenêtres de 2.5 s, qu'avec 20 descripteurs (1.8%), le premier résultat étant d'ailleurs légèrement inférieur mais très proche de ceux de Scheirer et Slaney, et de Williams et Ellis. L'homogénéisation de la décision sur l'ensemble de l'extrait réduit encore l'erreur mais l'évolution par rapport à la variation du nombre de descripteurs est encore moins intuitive. Williams et Ellis affichent une erreur nulle sur cette échelle, mais nous restons très sceptiques sur ce résultat, car il semble que celui-ci soit calculé sur une unique itération. Or, nous rencontrons de nombreuses itérations dans notre expérience où l'erreur est également nulle sur la fenêtre de 15 s.

10.6 Expérience 4 : Détection de voix chantée

Nous concluons cette partie expérimentale en appliquant le système décrit sur le problème auxiliaire de la détection de chant, afin de montrer que l'architecture en question peut s'adapter à différents problèmes de classification. Cependant, parce que les variations de classes sont beaucoup plus fréquentes sur ce problème (chaque pause dans le phrasé du chanteur induit une transition de musique instrumentale), nous nous restreignons ici à l'échelle temporelle courte (soit des trames de 32 ms, avec un pas d'avancement de 16 ms). Ceci limite donc considérablement le nombre de descripteurs mis en jeu puisque les résultats d'intégration temporelle ne sont pas exploitables ici. Nous employons tout de même les descripteurs long-termes en répétant leur valeur sur l'ensemble des trames courtes couvertes par chaque trame longue. On obtient ainsi 116 composantes pour caractériser les classes de chant et de musique instrumentale, parmi lesquelles on sélectionne les d plus pertinentes par l'algorithme IRMFSP. La classification se fait par une unique machine SVM apprise sur une base contenant 20000 exemples de chaque classe.

L'emploi d'une fenêtre de décision très courte implique nécessairement une plus forte variabilité de la sortie des SVM, puisque beaucoup moins d'information est exploitée pour le calcul de chaque descripteur. Ainsi, on peut constater sur la figure 10.8(a), que la probabilité a posteriori de la classe de chant (en bleu) est assez bruitée et oscille autour du seuil de décision de 0.5 (indiqué par une ligne grise). Ceci se traduit par une séquence estimée de classes (les points rouges) très instable, qui contient de très fréquentes erreurs accidentelles (l'annotation réelle est représentée en noir, immédiatement au-dessus et en dessous des estimations en rouge). Nous avons donc profité ici des techniques de post-traitement que nous avons introduites précédemment. La figure 10.8(b) montre le résultat du filtrage médian sur les probabilité a posteriori, et l'estimation déduite, sur laquelle on peut constater que quelques transitions accidentelles subsistent. Nous verrons qu'on obtient de meilleurs résultats encore en appliquant le post-traitement par HMM que nous avons proposé, et dont le résultat sur l'exemple précédent est illustré sur la figure 10.8(c). On remarque que même si les frontières des segments ne correspondent pas exactement à l'annotation réelle, la séquence de classes est beaucoup plus stable et les transitions correspondent mieux (à un décalage près) à la vérité terrain.

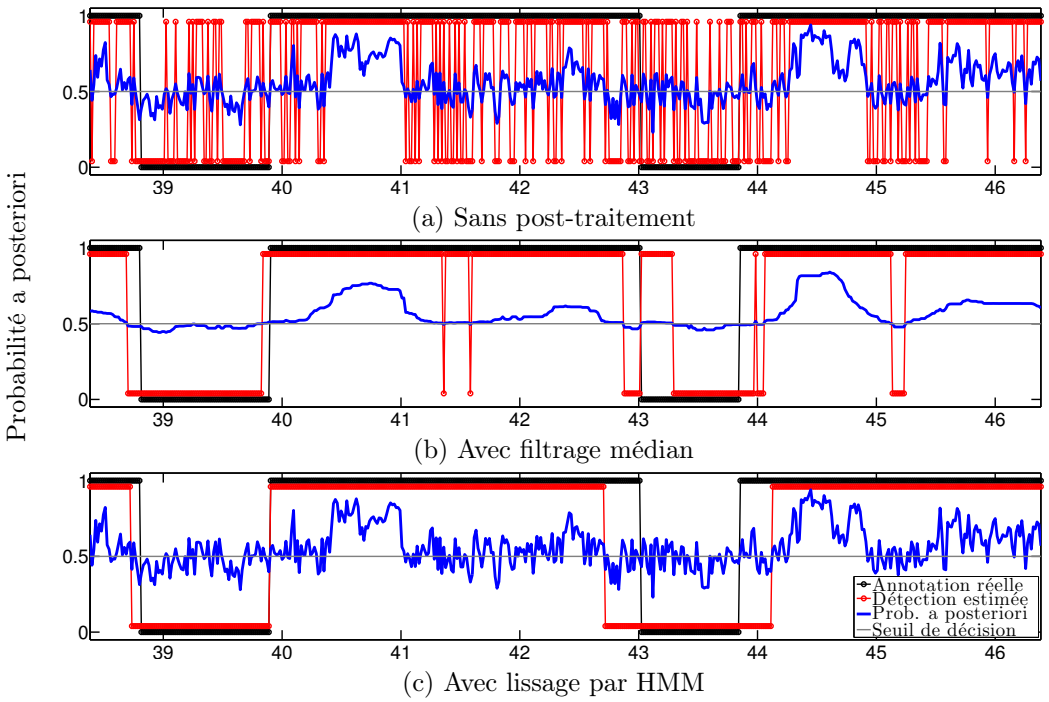


FIGURE 10.8 – Illustration de l'effet des post-traitements sur la probabilité a posteriori de la classe de chant. On constate que le filtrage médian (b) et le lissage par HMM (c) réduisent considérablement le nombre de transitions erronées par rapport à l'estimation de base (a).

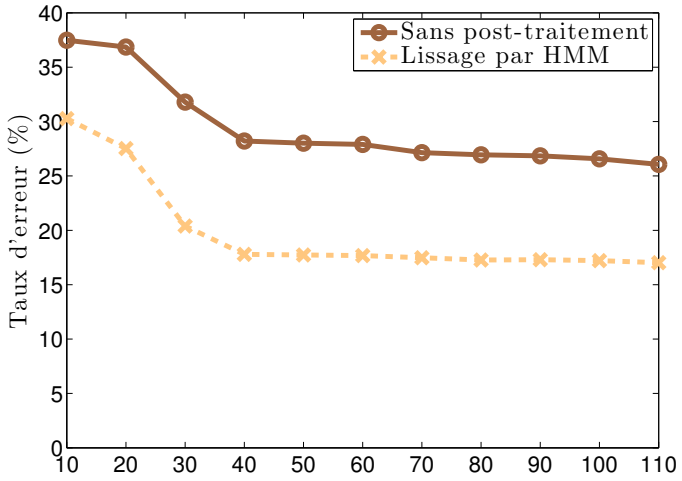


FIGURE 10.9 – Comparaison de l'évolution de la F-mesure sur le corpus Jamendo, avec ou sans post-traitement par HMM, en fonction du nombre de descripteurs d.

La figure 10.9 montre l'évolution de l'erreur de classification (calculée sur les trames) en fonction du nombre de descripteurs sélectionnés d, évoluant entre 10 et 110, en confrontant les résultats sans post-traitement (en ligne pleine marron) et après lissage par HMM (en pointillés jaunes). Le système est appris et testé sur le corpus Jamendo que nous avons introduit dans la section 10.1.4.

On constate que l'erreur décroît avec l'augmentation du nombre de composantes, mais reste assez stable au délà de $d = 40$, ce qui laisse supposer que l'essentiel de l'information discriminante est contenue dans les premiers descripteurs sélectionnés. Ainsi le taux de bonne classification de 82.2% pour $d = 40$ augmente légèrement jusqu'à 83.0% à $d = 110$, mais le premier cas constitue un meilleur compromis entre coût et performances ; nous fixons donc le nombre de descripteurs sélectionnés à $d = 40$ dans le reste de cette expérience.

Classe	**Chant**		**Musique**		**Global**	
Critère	**fr%**	**F%**	**fr%**	**F%**	**fr%**	**F%**
Sans post-traitement	74.8	72.6	68.5	70.4	71.8	71.6
Avec filtrage médian	84.6	81.6	76.4	80.5	80.7	81.1
Segmentation	88.0	84.8	74.0	80.3	81.3	82.7
Lissage par HMM	80.9	84.4	84.0	82.0	82.2	83.2
Vembu & Baumann [237]	87.7	70.5	35.8	47.4	62.7	62.4
Regnier & Peeters [196]	-	-	-	-	-	76.8

TABLE 10.10 – Comparaison du taux de bonne classification (fr%) et de la F-mesure (F) sur l'ensemble de test du corpus Jamendo, avec les différents algorithmes de post-traitement proposés, et confrontés à ceux de deux publications.

Le tableau 10.10 décrit plus en détail les résultats obtenus en termes de taux de bonne classification et de F-mesure, ainsi que sur chaque classe. Les performances sans pré-traitement et avec filtrage médian ou lissage par HMM, sont comparées. Nous avons également testé un post-traitement suivant l'approche hybride par segmentation aveugle, inspiré de celui proposé par Tsai et Wang [203], où les segments sont délimités par une procédure de détection d'attaques introduite par Duxbury et al. [67]. Nous avons également implémenté, à titre de comparaison, une autre approche par SVM tirée de la littérature [237], basée sur un vecteur de descripteurs de 38 composantes groupant les MFCC, les PLP et les LFPC. Nous avons conservé pour les paramètres du noyau RBF gaussien, les valeurs que l'auteur suggère. Les décisions de ce système sont basées sur des fenêtres de 190 ms, sur lesquelles seules les moyennes des valeurs des descripteurs sont considérées. Nous indiquons enfin le résultat en F-mesure globale de Regnier et Peeters [196], dont le système repose sur le seuillage empirique d'un unique descripteur basé sur une mesure combinée de vibrato et de trémolo, que nous avions évoqué dans la section 2.4.2 de l'état de l'art.

Notre système atteint 71.8% de bonne classification sans post-traitement, ce qui est largement supérieur aux résultats de l'approche de Vembu et Baumann, pourtant calculés sur des fenêtres de décision beaucoup plus conséquentes ; ceci montre la pertinence des descripteurs employés, ainsi que celle du processus de sélection, par rapport à un ensemble plus classique de descripteurs tirés de l'analyse de la parole. Le lissage par HMM offre ici les meilleurs résultats, par rapport aux deux autres post-traitements évalués, avec un taux de bonne classification de 82.2% et une F-mesure de 83.2%. On remarque toutefois que le traitement par segmentation est beaucoup plus efficace sur la classe de chant (88% contre seulement 80.9% pour les HMM), sans doute parce que celle-ci contient des attaques plus prononcées que le fond musical, qui favorisent donc le processus de segmentation. Les HMM, parce qu'ils appliquent un traitement distinguant les deux classes (chacune étant modélisé par un modèle GMM et des probabilités de transition propres), contrairement aux deux autres approches, appliquent en définitive le meilleur compromis entre les deux classes. On remarque d'ailleurs que l'on retrouve ce biais vers la classe de parole dans l'approche de Vembu de Baumann. Nous constatons que notre système se révèle plus efficace que celui proposé par Regnier et Peeters (qui post-traite également les décisions), même si celui-ci se distingue par l'étonnante efficacité de son unique descripteur.

Pour finir, le tableau 10.11 détaille les résultats sur quelques-uns des 16 fichiers audio du corpus de test, avec lissage par HMM. Bien que le système montre un léger avantage pour l'identification de la musique instrumentale, par rapport au chant, on remarque que c'est généralement la mauvaise identification de la musique qui fait chuter les résultats sur les pires fichiers (comme par exemple les numéros 4, 5, 7 et surtout 8). Après écoute des fichiers en question on constate que l'erreur provient d'un instrument particulier dont le timbre est proche de la voix. Ceci montre les limites de notre caractérisation de cette classe, dont l'extrême diversité peut impliquer des manifestations acoustiques très proches de la voix chantée.

Classe	**Chant**		**Musique**		**Global**	
Fichier audio	**fr%**	**F%**	**fr%**	**F%**	**fr%**	**F%**
1. À Poings Fermés.wav	84.5	92.2	98.7	95.7	93.7	94.6
2. 16 ans.wav	69.3	84.8	99.6	94.7	91.5	92.7
3. Une charogne.wav	87.1	91.7	79.4	72.3	85.3	86.8
4. Castaway.wav	94.4	87.3	55.1	73.4	79.0	82.8
5. Believe.wav	94.4	88.5	52.3	65.2	80.0	82.3
6. Si Dieu.wav	66.1	80.7	97.4	72.6	76.4	77.3
7. Elles disent.wav	76.1	78.7	63.6	59.9	71.8	72.1
8. L'Irlandaise.wav	93.8	64.2	33.4	47.5	57.7	57.1
Tous	80.9	84.3	84.0	81.8	82.2	83.2

TABLE 10.11 – Résultats détaillés de la détection de chant sur quelques-uns des fichiers de l'ensemble de test du corpus Jamendo.