Vers un système embarqué de classification d'événements sonores : étude de l'impact de la quantification des descripteurs

Marie-Anne LACROIX, Nancy BERTIN, Romuald ROCHER, Pascal SCALART

Univ Rennes, CNRS, IRISA
Marie-Anne.Lacroix@loria.fr, Nancy.Bertin@irisa.fr
Romuald.Rocher@irisa.fr, Pascal.Scalart@irisa.fr

Résumé – Le déploiement d'applications de classification d'événements sonores sur système embarqué est susceptible de dégrader les performances d'algorithmes n'ayant fait leurs preuves que sur une implémentation logicielle sans contraintes matérielles. Nous étudions expérimentalement l'impact de la quantification des descripteurs audio en entrée de deux classifieurs de référence. Nous concluons à l'acceptabilité de la quantification sur 8 bits, sans bénéfice d'un apprentissage adapté.

Abstract – Deployment of sound event detection applications on embedded systems might degrade performance of algorithms which proved efficient on software implementations without hardware constraints. We experimentally study the impact of quantizing audio features at the input of two reference classifiers. We conclude that 8-bit quantization is acceptable, with no benefits from a matched training.

1 Introduction

La classification d'événements sonores est un domaine de recherche en pleine expansion et aux applications nombreuses, par exemple la surveillance acoustique pour la sécurité des biens et des personnes. Le développement des objets connectés et la miniaturisation des microphones permet aujourd'hui d'envisager le déploiement de telles applications sur système embarqué. L'implémentation matérielle menace cependant les performances des algorithmes de l'état de l'art, performances souvent évaluées sur des implémentations logicielles (par exemple lors du challenge *DCASE*, qui fait référence en la matière [5]).

Dans ces travaux, nous nous penchons sur cette question, et en particulier celle de l'impact de la quantification des descripteurs audio en entrée des classifieurs, nécessaire pour calculer en virgule fixe. Nous présentons les deux systèmes de classification étudiés dans la section 2 et la quantification appliquée aux descripteurs audio dans la section 3. Notre contribution, expérimentale, est présentée en section 4 ; nous y comparons les performances des systèmes sans ou avec quantification des descripteurs, au décodage et à l'apprentissage. Nous concluons à l'acceptabilité de la quantification sur 8 bits, sans bénéfice d'un apprentissage adapté.

2 Systèmes de classification étudiés

Dans ces travaux, nous choisissons d'utiliser un classifieur discriminatif et un génératif issus de l'état de l'art : *(i)* le système de référence proposé au challenge DCASE 2016 [5], basé sur des modèles de mélanges gaussiens *(GMM)* binaires et *(ii)* un classifieur basé sur un réseau de neurones convolutif récurrent *(CRNN)* inspiré du système victorieux en 2017 [1] et des systèmes de référence des éditions 2019 et 2020 [6].

2.1 Vue globale du système

La classification d'événements sonores est mise en œuvre dans le cadre classique de l'apprentissage supervisé. Lors de la phase d'apprentissage, le signal d'entrée est segmenté en trames temporelles et transformé en un vecteur de descripteurs acoustiques, puis un modèle (dépendant du classifieur choisi) est appris en confrontant les descripteurs aux annotations du signal. Lors de la phase de décodage, le signal inconnu appliqué en entrée du modèle résulte en un vecteur de probabilité de présence d'un événement de chacune des classes au sein de la trame d'analyse. Après un éventuel post-traitement, les probabilités sont binarisées en fonction d'un seuil θ_{detect} (ici réglé empiriquement pour maximiser la F-mesure en virgule flottante, *cf. infra*) pour décider des étiquettes finales.

2.2 Classifieur GMM

Le premier classifieur étudié se fonde sur un modèle GMM dit *soustractif*, qui consiste à apprendre deux modèles : le premier pour détecter la présence d'événements sonores et le second pour détecter l'absence de chaque classe. Ce classifieur utilise un vecteur d'entrée constitué de 19 coefficients cepstraux en échelle mel ou *MFCC* (le premier est supprimé), ainsi que les coefficients des dérivées première et seconde (premier coefficient inclus), soit 20 ΔMFCC (i.e. vitesse) et 20 Δ^2MFCC (i.e. accélération), ce qui produit un ensemble de 59 descripteurs. La modélisation GMM représente indépendam-

ment chacune des C classes par un mélange de n_c lois normales multi-dimensionnelles.

Nous utilisons les paramètres suivants. Chaque mélange est composé de 4 gaussiennes, apprises par le passage successif de l'ensemble des données d'apprentissage pendant 80 époques. Les paramètres à apprendre (moyennes et covariances) sont initialisés aléatoirement ; l'initialisation est répétée plusieurs fois, et celle ayant la plus faible inertie est conservée [1].

Ce modèle, naguère compétitif, présente encore l'intérêt d'être utilisable en pratique lorsque l'on dispose de peu de données étiquetées.

2.3 Classifieur CRNN

Le second classifieur est un CRNN. En entrée du réseau, on applique un mel-spectrogramme de $n_{\text{seq}} = 256$ trames temporelles par lot et 40 bandes mel (issues de 128 filtres), dont les coefficients sont centrés et réduits pour chaque trame [2, 3]. Le réseau est constitué de (i) trois couches convolutives suivies de (ii) deux couches récurrentes et (iii) d'un réseau pleinement connecté en sortie.

Les principaux paramètres des couches convolutives sont : une taille de filtre de (3×3) ; un nombre de filtres de 128 ; une fenêtre glissante de *pooling* de 1 ; une taille de fenêtre du *pooling* de (2×1) et un taux de *dropout* de 0.4. Les trois couches convolutives (convolution, puis activation non-linéaire de type *ReLU*, puis dropout) sont suivies de deux couches récurrentes de type *GRU* bidirectionnel pour prendre en compte la temporalité des descripteurs. Un *dropout* de taux 0.5 est appliqué aux neurones actuels et aux neurones récurrents. Enfin, une fonction d'activation non linéaire (de type tangente hyperbolique) est appliquée en sortie de ces couches.

Le dernier étage du CRNN est constitué de deux couches denses dans lesquelles tous les neurones sont connectés (sans *dropout*). La première couche diminue la taille des matrices de $n_{\text{seq}} \times 32$ en entrée à $n_{\text{seq}} \times 16$ en sortie. La seconde diminue encore cette taille pour parvenir à une matrice indiquant les probabilités de chaque classe pour chaque trame, soit ici une taille de $n_{seq} \times 6$. Ces deux couches denses sont enveloppées dans une couche dite *TimeDistributed* qui permet d'appliquer la couche à chacune des n_{seq} trames indépendamment. La sortie du réseau est obtenue à l'aide d'une couche d'activation (de type sigmoïde) de manière à obtenir les probabilités de sortie.

Chaque lot est constitué de 32 matrices de taille $40 \times n_{\text{seq}}$. On réalise alors un apprentissage par validation croisée à 4 plis sur 100 époques. En raison de la variabilité des résultats en sortie des réseaux neuronaux, on effectue cet apprentissage 5 fois et on moyennera les performances obtenues lors des 5 tests pour déterminer les résultats du système.

1. Lors d'expériences préliminaires (non présentées ici), nous avons exploré la possibilité d'approcher la matrice de covariance par une matrice diagonale. Il s'avère que cette approximation classique, visant à réduire la complexité calculatoire, dégrade davantage les performances dans le cas de descripteurs quantifiés. Nous l'avons donc exclue de la suite de cette communication.

3 Quantification des descripteurs

Afin d'étudier l'impact de la quantification des descripteurs du signal audio, ces derniers subissent une quantification vectorielle fractionnée [4]. Celle-ci consiste à subdiviser le vecteur des descripteurs en plusieurs paires de sous-vecteurs, chacune étant quantifiée indépendamment à l'aide de l'algorithme LBG (*Linde Buzo Gray*). Une telle approche conduit à un coût calculatoire bien plus faible par rapport à une quantification vectorielle appliquée sur la totalité du vecteur des descripteurs car le nombre n total de bits affectés au codage est alors partagé entre tous les sous-vecteurs.

Pour le classifieur GMM, seuls les 20 MFCC sont quantifiés à l'aide de l'algorithme LBG. En effet, afin de réduire au maximum le budget d'éléments binaires, nous choisissons de ne pas quantifier directement les 20 ΔMFCC et 20 Δ^2MFCC, qui peuvent être approchés à partir des MFCC quantifiés au sein du nœud collecteur. Le vecteur de 59 descripteurs est ainsi reconstitué, en supprimant le coefficient constant. On économise ainsi le coût d'encodage binaire de 39 descripteurs, soit près des deux tiers du vecteur initial. Dans le cas du CRNN, une telle économie n'est pas possible. Nous quantifions donc l'ensemble des 40 coefficients du spectre mel, soit 20 paires de descripteurs encodées par l'algorithme LBG comme indiqué précédemment.

4 Étude expérimentale

4.1 Données et protocole

Les données d'apprentissage et de test sont issues de la base TUT2017 utilisée dans de nombreux travaux, et pour lesquels des résultats de référence sont disponibles [5]. Cette base est constituée d'enregistrements réels stéréo réalisés en environnement urbain et échantillonnés à 44,1 kHz. 6 classes sont représentées, dans des proportions inégales : *brakes squeaking* (crissement de freins), *car* (voiture) et *large vehicle* (véhicule lourd), *children* (enfants), *people speaking* (discussion) et *people walking* (bruits de pas). Les événements sont de durée variable, typiquement de l'ordre de quelques secondes, et peuvent être simultanés. La base, d'une durée de 92 minutes, est fournie avec des annotations complètes (étiquetage fort).

Les performances en sortie du classifieur sont évaluées suivant deux critères classiques en détection d'événements [6] : le taux d'erreur (ER) et la F-mesure ($\mathcal{F}$). Ces critères sont calculés sur des segments d'une seconde, puis moyennés soit sur tous les échantillons de test (« micro-moyennage »), soit classe par classe avant de moyenner les scores par classe (« macro-moyennage ») afin de tenir compte des déséquilibres.

Dans un premier temps, on réalise l'entraînement des deux classifieurs à partir de descripteurs codés en virgule flottante sur 64 bits chacun et seule la phase de test est exécutée avec des descripteurs quantifiés (section 4.2). On observe les performances pour des paires de descripteurs de test codées en virgule fixe successivement sur 1, 2, 4, 6 et 8 bits, ainsi que celles

obtenues pour des descripteurs codés en virgule flottante sur 64 bits. Ces résultats sont comparés au cas où l'entraînement est réalisé sur données quantifiées (section 4.3).

4.2 Apprentissage virgule flottante sur 64 bits

Classifieur GMM soustractif. La table 1 présente les performances obtenues par le GMM décrit section 2.2. On observe une augmentation globale de la F-mesure quand croît la longueur de quantification, ce qui semble logique (les valeurs quantifiées approchant mieux les valeurs d'origine en virgule flottante sur 64 bits). Toutefois, pour deux classes (*large vehicle* et *people speaking*), il existe une longueur de classification optimale (respectivement 2 bits et 6 bits) au-delà de laquelle la performance se dégrade. Dans tous les cas, les valeurs du taux d'erreur et de la F-mesure pour une quantification fractionnée par paires de descripteurs sur 8 bits sont généralement très proches ou légèrement supérieures aux performances d'origine.

Classifieur CRNN. Nous étudions maintenant l'impact de la quantification sur le système CRNN de la section 2.3. Les performances sont données dans la table 2 et dans les mêmes conditions que précédemment. On constate que l'évolution des résultats en fonction du nombre de bits de quantification est moins univoque. En effet, la F-mesure maximale et le taux d'erreur minimal de chaque ligne ne sont pas systématiquement obtenus pour une même expérience : pour la classe *large vehicle*, la F-mesure maximale est atteinte pour les descripteurs d'origine ; pour *car*, avec un codage sur 1 bit ; pour *people walking*, avec un codage sur 8 bits, etc. Les classes sont diversement impactées par la longueur de la quantification, les variations de F-mesure allant de valeurs négligeables (*car*, *brakes squeaking*) à un doublement (*children*, *people speaking*). Il convient cependant de noter que sur 8 bits, la F-mesure diffère toujours des performances d'origine de 1 ou 2% environ.

4.3 Apprentissage en virgule fixe

Classifieur GMM. Dans la table 3 et au contraire des résultats attendus, on observe une dégradation générale des performances en comparaison de celles obtenues pour un classifieur GMM entraîné sur les données non quantifiées (sauf à 1 bit, où les performances sont toujours très médiocres). En outre, si les variations de la F-mesure et du taux d'erreur suivent des tendances similaires au cas précédent (section 4.2), leurs valeurs ne tendent plus toujours vers celles d'origines. Pour un codage des paires de descripteurs sur 8 bits, toutes les classes montrent une différence entre leur taux d'erreur et celui d'origine inférieure à 0.10, mais seules quatre ont atteint une F-mesure proche de celle d'origine (moins de 2.5% de différence ici). Les dégradations peuvent être sévères, par exemple dans la classe minoritaire *children*, qui perd 13.3% de F-mesure et n'est plus détectée. Nous pouvons conclure que l'apprentissage spécifique à partir de données quantifiées n'est pas souhaitable comparativement au classifieur d'origine (64 bits).

Classifieur CRNN. Enfin, on étudie l'influence d'un apprentissage spécifique sur le CRNN en analysant les performances relevées dans la table 4. Comme pour le GMM, on observe que l'apprentissage sur les données quantifiées n'est favorable que dans le cas d'un codage des paires de descripteurs sur 1 bit. Au-delà, les résultats sont mitigés. Certaines classes voient leur F-mesure augmenter – mais également leur taux d'erreur croître – (*large vehicle* et *people walking*), tandis que la F-mesure décroît pour d'autres classes (*children* et *people speaking*). La différence de performances reste faible pour les autres classes. Par ailleurs, on note que, pour un codage sur 8 bits, la F-mesure et le taux d'erreur approchent davantage les performances d'origine que pour un apprentissage sur les données non quantifiées.

5 Conclusion

Cette étude a porté sur la problématique de la classification d'événements sonores à l'aide d'un classifieur GMM soustractif et d'un CRNN. Nous avons mesuré les conséquences de l'utilisation de données d'entrées (descripteurs) quantifiées relativement au cas usuel où les données sont représentées en pleine précision (virgule flottante 64 bits). Les résultats expérimentaux montrent qu'une longueur de quantification suffisante permet d'obtenir des performances similaires à celles obtenues pour des descripteurs codés en virgule flottante (64 bits). Par contre, il ne semble pas nécessaire d'adapter la phase d'apprentissage avec des données quantifiées. Nous pouvons également remarquer que le GMM soustractif semble plus robuste à la quantification des descripteurs que ne l'est le CRNN. Au même titre que la compression, la latence, ou encore la consommation globale d'énergie, cette problématique de quantification des données d'entrée constitue l'un des élements clefs d'analyse de tout dispositif opérationnel de classification sonore embarquée.

Références

[1] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *ICASSP*, 2017, pp. 771–775.

[2] E. Cakir, "Deep neural networks for sound event detection," Ph.D. dissertation, Tampere University, 2019.

[3] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4," in *Detect. Classif. Acoust. Scenes Events*, 2019.

[4] "Speech processing, Transmission and Quality aspects (STQ) ; Distributed speech recognition ; Front-end feature extraction algorithm ; Compression algorithms," ETSI, ES 201 108, 2003.

[5] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 Challenge," *IEEE/ACM TASLP*, 2019.

[6] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Detect. Classif. Acoust. Scenes Events*, 2019, pp. 253–257.

TABLE 1 – Performances du *GMM* avec apprentissage sur descripteurs non quantifiés. $\theta_{\text{detect}} = 35$.

	Sans quantif.		Codage 1 bit		Codage 2 bits		Codage 4 bits		Codage 6 bits		Codage 8 bits	
	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$
Total micro-moy.	1.13	45.2%	1.29	28.1%	1.52	39.7%	1.39	41.7%	1.20	44.9%	**1.12**	**45.7%**
Total macro-moy.	1.92	35.2%	2.63	21.9%	2.64	30.6%	2.57	33.0%	2.08	34.7%	**1.86**	**35.6%**
brakes squeaking	1.36	18.9%	**1.02**	2.2%	1.08	4.0%	1.64	13.3%	1.44	14.8%	1.35	**19.0%**
car	**0.74**	**60.4%**	0.89	37.2%	0.83	54.7%	0.83	55.4%	0.77	58.6%	**0.74**	60.3%
children	3.49	13.3%	8.00	5.3%	6.60	9.7%	6.33	8.4%	4.22	11.2%	**3.13**	**13.5%**
large vehicle	2.63	36.9%	1.94	35.1%	**1.76**	**42.9%**	2.01	41.6%	2.53	37.6%	2.68	36.2%
people speaking	2.03	30.1%	2.14	20.3%	3.18	30.9%	2.70	34.2%	2.04	**36.1%**	**1.96**	32.6%
people walking	**1.26**	51.9%	1.78	31.5%	2.40	41.1%	1.91	45.2%	1.49	50.0%	1.30	**52.1%**

TABLE 2 – Performances du *CRNN* avec apprentissage sur descripteurs non quantifiés. $\theta_{\text{detect}} = 0.5$.

	Sans quantif.		Codage 1 bit		Codage 2 bits		Codage 4 bits		Codage 6 bits		Codage 8 bits	
	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$
Total micro-moy.	1.07	**39.1%**	1.16	34.2%	1.12	38.4%	**1.06**	38.8%	1.07	38.6%	**1.06**	38.2%
Total macro-moy.	1.67	**25.2%**	1.57	19.3%	**1.05**	23.9%	1.54	24.7%	1.60	24.8%	1.64	24.6%
brakes squeaking	1.0	0.0%	1.0	0.0%	1.0	0.0%	1.0	0.0%	1.0	0.0%	1.0	0.0%
car	**0.84**	51.5%	1.03	**52.8%**	0.91	51.5%	0.85	51.6%	0.86	50.5%	0.87	49.9%
children	1.97	0.7%	**1.01**	1.0%	**1.01**	**2.1%**	1.48	**2.1%**	1.86	1.6%	2.28	**2.1%**
large vehicle	2.87	**30.1%**	3.04	20.4%	2.51	21.8%	**2.46**	26.9%	2.56	27.5%	2.50	27.9%
people speaking	1.92	23.0%	**1.56**	13.0%	1.86	**25.9%**	1.86	23.7%	1.89	24.0%	1.83	21.6%
people walking	1.43	45.8%	1.76	26.7%	1.76	41.9%	1.58	43.8%	1.44	45.5%	**1.33**	**46.2%**

TABLE 3 – Performances du *GMM* avec apprentissage sur descripteurs quantifiés. $\theta_{\text{detect}} = 35$.

	Sans quantif.		Codage 1 bit		Codage 2 bits		Codage 4 bits		Codage 6 bits		Codage 8 bits	
	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$
Total micro-moy.	1.13	**45.2%**	1.91	30.6%	1.53	33.4%	1.20	42.4%	1.15	42.6%	**1.12**	43.5%
Total macro-moy.	1.92	**35.2%**	5.56	25.7%	2.99	27.2%	2.19	31.4%	2.13	32.5%	**1.91**	31.6%
brakes squeaking	1.36	18.9%	8.53	15.5%	2.23	**19.0%**	1.47	9.8%	1.39	18.7%	**1.26**	16.5%
car	**0.74**	**60.4%**	0.85	58.0%	0.89	42.9%	0.79	54.7%	0.75	56.6%	**0.74**	59.0%
children	**3.49**	**13.3%**	17.78	4.5%	6.87	1.3%	4.91	0.9%	4.80	0.9%	**3.49**	0.0%
large vehicle	2.63	36.9%	**1.96**	16.0%	3.73	32.6%	2.27	**40.9%**	2.52	37.4%	2.60	36.9%
people speaking	**2.03**	30.1%	2.21	23.0%	2.33	25.9%	2.29	**31.5%**	2.09	29.5%	2.09	26.0%
people walking	1.26	**51.9%**	2.02	36.9%	1.88	41.7%	1.41	50.6%	**1.23**	51.7%	1.26	51.4%

TABLE 4 – Performances du *CRNN* avec apprentissage sur descripteurs quantifiés. $\theta_{\text{detect}} = 0.5$.

	Sans quantif.		Codage 1 bit		Codage 2 bits		Codage 4 bits		Codage 6 bits		Codage 8 bits	
	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$	ER	$\mathcal{F}$
Total micro-moy.	1.07	39.1%	**0.87**	**41.9%**	1.19	34.0%	1.17	35.4%	1.05	39.9%	1.11	38.1%
Total macro-moy.	1.67	**25.2%**	**1.14**	20.8%	1.99	22.3%	1.82	23.2%	1.59	24.9%	1.70	25.0%
brakes squeaking	1.0	0.0%	1.0	0.0%	1.0	0.0%	1.0	0.0%	1.0	0.0%	1.0	0.0%
car	0.84	51.5%	0.86	**60.5%**	0.85	50.8%	**0.83**	48.4%	0.85	54.3%	**0.83**	50.8%
children	1.97	0.7%	**1.11**	0.0%	3.82	0.0%	2.36	0.6%	1.63	0.0%	2.07	**1.9%**
large vehicle	2.87	30.1%	**1.31**	21.7%	2.33	**31.7%**	2.81	29.8%	2.39	27.2%	2.73	31.6%
people speaking	1.92	**23.0%**	**1.35**	9.8%	2.32	16.7%	2.23	19.5%	2.24	19.5%	2.16	19.8%
people walking	1.43	45.8%	**1.21**	33.1%	1.60	34.7%	1.68	41.0%	1.43	**48.4%**	1.45	45.7%