HAL
open science

Une mesure d'expertise pour le crowdsourcing

Hosna Ouni, Arnaud Martin, Laetitia Gros, Mouloud Kharoune, Zoltan Miklos

To cite this version:

Hosna Ouni, Arnaud Martin, Laetitia Gros, Mouloud Kharoune, Zoltan Miklos. Une mesure d'expertise pour le crowdsourcing. Extraction et Gestion des Connaissances (EGC), Jan 2017, Grenoble, France. hal-01432561

HAL Id: hal-01432561

https://hal.science/hal-01432561

Submitted on 11 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Une mesure d'expertise pour le *crowdsourcing*

Hosna Ouni*, Arnaud Martin*, Laetitia Gros**, Mouloud Kharoune*, Zoltan Miklos*

*UMR 6074 IRISA, DRUID team, Université de Rennes 1, Lannion, France
hosnaouni@gmail.com, {arnaud.martin, mouloud.kharoune,zoltan.miklos}@univ-rennes1.fr,
http://www-druid.irisa.fr
** Orange Labs, 2, av. Pierre Marzin, F-22307 Lannion Cedex, France
laetitia.gros@orange.com

Résumé. Le *crowdsourcing*, un enjeu économique majeur, est le fait d'externaliser une tâche interne d'une entreprise vers le grand-public, la foule. C'est ainsi une forme de sous-traitance digitale destinée à toute personne susceptible de pouvoir réaliser la tâche demandée généralement rapide et non automatisable. L'évaluation de la qualité du travail des participants est cependant un problème majeur en *crowdsourcing*. En effet, les contributions doivent être contrôlées pour assurer l'efficacité et la pertinence d'une campagne. Plusieurs méthodes ont été proposées pour évaluer le niveau d'expertise des participants. Ce travail a la particularité de proposer une méthode de calcul de degrés d'expertise en présence de données dont l'ordre de classement est connu. Les degrés d'expertise sont ensuite considérés sur des données sans ordre pré-établi. Cette méthode fondée sur la théorie des fonctions de croyance tient compte des incertitudes des réponses et est évaluée sur des données réelles d'une campagne réalisée en 2016.

1 Introduction

Le *crowdsourcing*, concept lancé par Howe (2006), stimule la participation collective à l'élaboration de certaines tâches qu'une entreprise ne souhaite pas réaliser en interne par faute de ressources ou de temps et qu'il est compliqué voire impossible de confier à un ordinateur. Il s'inscrit dans une logique de partage dérivée de l'essor du web 2.0. En effet, l'échange des idées et des savoir-faire se réalise par l'intermédiaire d'une plateforme internet.

Plusieurs plateformes telles que Amazon Mechanical Turk (AMT), Microworker et Foule Factory sont destinées aux petites tâches que la machine est incapable d'effectuer rapidement et de façon fiable. Ces tâches sont généralement simples et courtes à l'instar de l'analyse des émotions, la catégorisation des produits ou la comparaison de designs.

Néanmoins, l'environnement d'une plateforme de *crowdsourcing* est incertain car finalement peu maîtrisé. Par suite, l'évaluation de la qualité et la fiabilité des contributions et des travailleurs eux-même est indispensable pour le bon déroulement d'un tel processus. Ainsi, plusieurs travaux ont été proposés pour identifier les experts ou les travailleurs sérieux de la plateforme.

Ce papier propose une solution permettant de calculer la pertinence des réponses des participants à partir des réponses à une campagne lancée par Orange Labs. Durant cette campagne,

on se propose de faire évaluer par les travailleurs de la plateforme de *crowdsourcing* la qualité sonore restituée par différentes solutions de codage audio. La procédure consiste à faire écouter aux participants 12 extraits musicaux de différentes qualités et à leur demander d'en évaluer la qualité audio sur une échelle à 5 catégories (Excellente, Bonne, Moyenne, Médiocre, Mauvaise). A chaque catégorie est associée une note allant de 1 (pour Mauvais) à 5 (pour Excellent) voir ITU (1996). Parmi les extraits sonores présentés dans un ordre aléatoire, 5 sont de qualité connue (ajout d'un bruit modulé par le signal, avec différents rapports signal/bruit (MNRUs : *Modulated Noise Reference Unit* voir ITU (1996))). Ces signaux sont utilisés dans les tests comme références et ancrages sur l'échelle de qualité. Dans cette étude, les signaux MNRU vont permettre de définir des degrés d'expertise. L'idée est de structurer les réponses des travailleurs par des graphes, représentant un ordre de préférences (issu des notes) entre les signaux MNRU, puis de comparer ces graphes avec celui de référence issu des notes théoriques attendus. Ensuite, cette estimation sera prise en compte pour sélectionner les participants experts pour les 7 autres extraits sonores afin d'atteindre l'objectif souhaité de classement de ces signaux (*i.e.* des codeurs dont ils sont issus).

La comparaison des graphes consiste à quantifier la similarité entre deux graphes. Il s'agit d'un problème très courant surtout dans les réseaux sociaux car les graphes révèlent des propriétés topologiques qu'on cherche à comprendre et à comparer. Malheureusement, on manque de références permettant de construire une échelle de comparaison de leurs caractéristiques géométriques. En effet, il n'y a pas de métrique ou de méthode de comparaison de graphes universelle. De plus, du point de vue algorithmique, les méthodes classiques pour aborder ce genre de problèmes sont complexes.

De plus, les réponses étant fournies pas des humains dans un environnement non contrôlé, contrairement aux tests d'écoute classiques en laboratoire, il est nécessaire de modéliser les imperfections sur les réponses. La théorie des fonctions de croyance permet de répondre à cette problématique et d'offrir un cadre théorique pour réaliser la combinaison des informations issues de différentes sources.

Nous proposons donc dans ce travail une approche originale permettant une estimation d'une mesure d'expertise à partir d'une comparaison de graphes dans le cadre de la théorie des fonctions de croyance. Ainsi, la section suivante présente les concepts de base de la théorie des fonctions de croyance. La section 3, après un bref rappel des approches existantes, propose une approche originale pour la représentation des réponses sous forme de graphes et le calcul de degrés d'expertise. Finalement, l'évaluation de la méthode sur des données réelles fait l'objet de la section 4.

2 La théorie des fonctions de croyance

La théorie des fonctions de croyance issue des travaux de Dempster (1967) et de Shafer et al. (1976) permet une représentation à la fois des incertitudes et des imprécisions mais aussi de l'ignorance d'une source (ici la réponse d'un contributeur). Considérant un ensemble $\Omega = \{\omega_1, \omega_2, .., \omega_n\}$ qui représente les réponses possibles à une question, une fonction de masse est définie sur 2^Ω (ensemble de toutes les disjonctions de Ω) et à valeur dans $[0, 1]$ avec les contraintes :

$$\begin{cases} \sum_{A \subseteq \Omega} m(A) = 1 \\ m(\emptyset) = 0 \end{cases} \tag{1}$$

La fonction de masse $m(A)$ représente la part de croyance allouée à la proposition A et qui ne peut pas être affectée à un sous-ensemble strict de A. Elle peut être vue comme une famille d'ensembles pondérées ou une distribution de probabilité généralisée. Un ensemble A est un élément focal si $m(A) \neq 0$. Par exemple, si nous considérons la fonction de masse $m(\{\omega_1, \omega_2\}) = 0.8$, $m(\Omega) = 0.2$, cette quantité représente une imprécision sur ω_1 ou ω_2 et une incertitude car la valeur affectée à cette proposition est 0.8.

La manipulation des données imparfaites issues de plusieurs sources distinctes nécessite de fusionner les informations. On parle donc de la combinaison des fonctions de masse permettant l'aboutissement à un état de connaissance générique et pertinent. L'opérateur de combinaison conjonctive proposé par Smets (1990) est donné pour deux fonctions de masse issues de deux sources par :

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B_1 \cap B_2 = A} m_1(B_1) m_2(B_2) \tag{2}$$

La masse affectée sur l'ensemble vide à l'issue de cette combinaison peut être interprétée comme l'inconsistance de la fusion.

Afin de mesurer l'écart à une fonction de masse attendue, par exemple pour prendre une décision ou définir une mesure, plusieurs distances ont été proposées. La distance de Jousselme et al. (2001) est la plus communément utilisée pour ses propriétés de répartitions des pondérations en fonction de l'imprécision des éléments focaux. Elle est donnée par :

$$d_J(m_1, m_2) = \frac{1}{2}(m_1 - m_2)^T \underline{\underline{D}}(m_1 - m_2) \tag{3}$$

avec :

$$\underline{\underline{D}}(X, Y) = \begin{cases} 1 \text{ si } X = Y = \emptyset \\ \dfrac{|X \cap Y|}{|X \cup Y|} \ \forall X, Y \in 2^{\Omega} \end{cases} \tag{4}$$

3 Caractérisation d'experts dans le crowdsourcing

3.1 Positionnement du travail

L'identification des experts sur les plateformes de crowdsourcing a fait l'objet de plusieurs travaux récents. Il faut distinguer deux types d'approches, celles tenant compte de questions dont on connaît la réponse (nommées données d'or) et celles où aucunes connaissances *a priori* n'est disponible. En effet, dans le contexte de ce dernier type d'approches, Ben Rjab et al. (2015, 2016) se sont intéressés à calculer un degré d'exactitude et de précision en supposant que la majorité a raison et en définissant ce degré à partir de la *distance de* Jousselme et al. (2001) entre sa réponse et la moyenne des réponses des autres participants. D'autre part, Dawid et Skene (1979) et Ipeirotis et al. (2010) ont utilisé l'algorithme Expectation-Maximisation (E.M) permettant dans une première phase d'estimer la bonne réponse pour chaque tâche, en utilisant des étiquettes affectées par les participants, puis, d'évaluer la qualité des travailleurs en comparant les réponses soumises à la bonne réponse inférée. Également, Smyth et al. (1995) et Raykar et al. (2010) ont utilisé cette approche pour les classements binaires et les étiquetages catégoriques. Raykar et Yu (2012) ont généralisé l'idée sur les classements ordinaires (associer

des notes de 1 à 5 en fonction de la qualité d'un objet ou d'un service). Ces méthodes proposent de calculer la "sensibilité" (les vrais positifs) et la "spécificité" (les vrais négatifs) pour chaque label. Ainsi, le participant est un spammer si son score est proche de 0 ; un expert parfait a un score égal à 1.

Cependant, ces algorithmes proposent de déterminer la qualité des réponses des participants quand la vérité est inconnue alors que dans notre cas, les notes correctes théoriques attribuées aux signaux MNRUs sont connues. En effet, nous cherchons plutôt à identifier les experts en nous fondant sur des données correctes de référence et définir un degré d'expertise proportionnel à la similarité entre les réponses d'un participant et les réponses connues à l'avance. Ainsi, notre travail est fondé sur des données d'or qui servent à estimer directement la qualité des participants tel que proposé par Le et al. (2010). Les données d'or sont des questions dont les réponses correctes sont connues à l'avance et qui sont injectées d'une façon arbitraire. Les données d'or ont l'avantage de mesurer explicitement la précision des travailleurs et d'être utilisées pour prendre des décisions concernant le travailleur : pouvons-nous utiliser leur travail ? Faut-il les laisser finir la tâche ? Est-ce qu'ils méritent un bonus ? En second lieu, c'est processus transparent dans le sens où on s'assure que les travailleurs comprennent les détails nuancés ou difficiles des exigences de la tâche.

Pour évaluer l'impact de l'utilisation des données d'or, Ipeirotis (2010) a examiné la performance de l'algorithme de Dawid et Skene (1979) modifié pour tenir compte de l'existence de données d'or. En variant le pourcentage des données d'or (0%, 25%, 50% et 75%), il a essayé de mesurer à chaque fois l'erreur de classification (dans quelle mesure l'algorithme estime la classe correcte des exemples) et l'erreur d'estimation de la qualité (dans quelle mesure l'algorithme estime la qualité des travailleurs). Il a trouvé que ce genre de données ne fait pas de différence par rapport au modèle non supervisé. Par contre, il a considéré que l'utilisation des données d'or est nécessaire dans quelques cas tel que le travail sur des ensembles de données très déséquilibrés (évaluer toutes les classes). Selon Ipeirotis (2010), les raisons les plus importantes sont le gain de la confiance des personnes non techniques (en proposant une approche de contrôle de qualité) et le calibrage des résultats lorsque la sensibilité des utilisateurs influent sur leurs réponses.

Également, Philips (2011) traite les données d'or comme un outil pour associer des scores de confiance aux contributeurs. Les participants doivent donc dépasser des seuils de confiance minimum pour continuer à travailler sur une tâche. Si à tout moment un contributeur tombe en dessous du seuil de confiance, on exclue son travail.

Dans ce travail, nous allons pondérer les réponses des contributeurs à partir des relations sur les réponses des données d'or. Cette pondération, réalisée à partir d'un degré d'expertise, pourra aller jusqu'à ne plus considérer les contributeurs trop éloignés des réponses attendues sur les données d'or.

3.2 Calcul d'un degré d'expertise

Comme déjà mentionné, les réponses des contributeurs de la plateforme sont représentées par des graphes. Pour cela, considérant un participant p qui a associé les notes présentées dans le tableau 1 aux MNRUs (les données d'or), le graphe orienté et pondéré correspondant est construit (*cf.* figure 1). Le graphe est orienté pour la lisibilité des préférences ($a \rightarrow b$ signifie a est préféré à b) et la facilité du travail sur les successeurs et les prédécesseurs. Il est conçu comme suit :

TAB. 1 – *Exemple de notes*

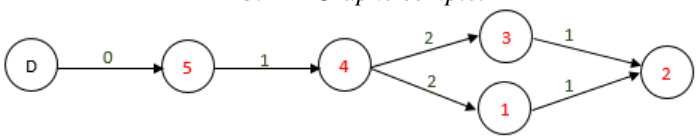
MNRU	1	2	3	4	5
Note associée	2	1	2	4	5

— L'insertion du point de départ virtuel D, correspondant toujours à la note 5, qui sert à l'extraction de la note associée au morceau $MNRU_i$ de la façon suivante :

$$note.ass(MNRU_i) = 5 - d_G(D, i) \tag{5}$$

— A chaque itération k, l'ensemble des nœuds ayant la k^{eme} note la plus élevée est recherché et cet ensemble est ajouté au graphe dans une même profondeur en respectant les normes suivantes :
- Coût de l'arc : différence entre les notes des MNRUs de deux profondeurs consécutives.
- Nom du nœud : numéro du MNRU

FIG. 1 – *Graphe complet*

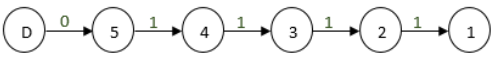


Pour calculer les degrés d'expertise, nous considérons le graphe de référence qui correspond aux notes théoriques attendues pour les MNRU et données par le tableau 2 qui correspond au graphe 2. Les graphes correspondants aux réponses des participants sont comparés

TAB. 2 – *Notes correctes*

MNRU	1	2	3	4	5
Note associée	1	2	3	4	5

FIG. 2 – *Graphe de référence*



à ce graphe de référence et une fonction de masse est ainsi calculée pour chaque réponse des contributeurs.

L'idée est ici d'extraire un ensemble de critères hétérogènes permettant de contourner les différences entre deux nœuds, ce qui est proche de la notion de "signature des nœuds" introduite par Jouili (2011), bien que les critères considérés ne soient pas les mêmes.

Une particularité de la modélisation proposée est que tous les graphes ont les mêmes nœuds (même nombre et même attribut). Ainsi, l'idée est de comparer tous les couples $(N_{(1,i)}, N_{(2,i)})$

où $N_{(1,i)}$ est le nœud d'attribut i dans le graphe de référence et $N_{(2,i)}$ est le nœud de même attribut appartenant au graphe à comparer.

Pour ce faire, nous avons défini pour chaque nœud du graphe quatre critères représentant les différentes erreurs possibles que nous avons identifiées. Ces critères sont représentés puis fusionnés à l'aide des fonctions de masse où le cadre de discernement considéré est :

$$\Omega = \{E, NE\} \tag{6}$$

où E représente l'assertion *Expert* et NE *Non expert*. Nous cherchons ainsi à mesurer la croyance dans le fait qu'un contributeur soit un expert (*i.e.* détermine l'ordre correct sur les MNRUs) en fonction des notes qu'il a attribuées sur les MNRUs.

1. Degré d'exactitude de la note associée : Ce critère est caractérisé par la différence de position d'un nœud considéré entre la référence et la réponse du contributeur. La dissimilarité est calculée à l'aide de la distance Euclidienne dans (7).

$$d_1(N_{1,i}, N_{2,i}) = |d_{G_1}(D, N_{1,i}) - d_{G_2}(D, N_{2,i})| \tag{7}$$

où $d_G(D, N_i)$ est la profondeur du nœud N_i par rapport au nœud D.
La fonction de masse correspondante à ce critère est donnée par :

$$\begin{cases} m_1(N_{1,i}, N_{2,i})(E) = 1 - \dfrac{d_1(N_{1,i}, N_{2,i})}{d_{max}} \\ m_1(N_{1,i}, N_{2,i})(NE) = \dfrac{d_1(N_{1,i}, N_{2,i})}{d_{max}} \end{cases} \tag{8}$$

où d_{max} est la distance maximale entre deux nœuds. Compte tenu du fait que les graphes considérés ne représente que 5 notes, $d_{max} = 4$.

2. Degré de confusion entre les MNRUs : Ce critère mesure la proportion des nœuds de même note/distance au point de départ D que le nœud concerné. La dissimilarité de Jaccard sera ainsi employé pour la comparaison des contenus des ensembles dans (9)

$$d_2(N_{1,i}, N_{2,i}) = \frac{|I_{N_{1,i}} \bigcap I_{N_{2,i}}|}{|I_{N_{1,i}} \bigcup I_{N_{2,i}}|} \tag{9}$$

où $I_{N_i} = \{N_j \in V; d_G(D, N_j) = d_G(D, N_i)\}$, avec V l'ensemble des nœuds du graphe.
La fonction de masse associée est donnée par :

$$\begin{cases} m_2(N_{1,i}, N_{2,i})(E) = d_2(N_{1,i}, N_{2,i}) \\ m_2(N_{1,i}, N_{2,i})(NE) = 1 - d_2(N_{1,i}, N_{2,i}) \end{cases} \tag{10}$$

Cette masse représente une valeur minimale de 0.2.

3-4. Degré de mauvais ordre précédent (sur l'ensemble de prédécesseurs) et **de mauvais ordre suivant** (sur l'ensemble de successeurs). Le participant peut considérer un morceau meilleur qu'un autre, contrairement à ce qui est attendu. Ainsi ces critères mesurent ces erreurs d'inversion par rapport au précédent ou suivant. Afin de définir ces degrés, nous introduisons la définition des ensembles suivants respectivement pour l'ensemble des prédécesseurs ($P^C_{N_i}$ correctes et $P^{NC}_{N_i}$ non correctes) et l'ensemble des successeurs ($S^C_{N_i}$ correctes et $S^{NC}_{N_i}$ non

correctes) :

$$\begin{cases} P^C_{N_i} = \{N_j \in V / N_j \in Pred_{G_1}(N_i)\} \\ P^{NC}_{N_i} = \{N_j \in V / N_j \in Succ_{G_1}(N_i)\} \end{cases} \text{ et } \begin{cases} S^C_{N_i} = \{N_j \in V / N_j \in Succ_{G_1}(N_i)\} \\ S^{NC}_{N_i} = \{N_j \in V / N_j \in Pred_{G_1}(N_i)\}\} \end{cases}$$

où $Succ_G(N)$ et $Pred_G(N)$ sont respectivement l'ensemble des successeurs et l'ensemble des prédécesseurs du nœud N dans le graphe G.
A partir de ces définitions, les distances d_3 et d_4 sont données par les équations (11) et (12).

$$\begin{cases} d_{3,1}(N_{1,i}, N_{2,i}) = \dfrac{|P^C_{N_{2,i}} \bigcap P_{N_{1,i}}|}{|P_{N_{1,i}} \bigcup P_{N_{2,i}}|} = m_3(N_{1,i}, N_{2,i})(E) \\ d_{3,2}(N_{1,i}, N_{2,i}) = \dfrac{|P^{NC}_{N_{2,i}}|}{|P_{N_{2,i}}|} = m_3(N_{1,i}, N_{2,i})(NE) \end{cases} \tag{11}$$

$$\begin{cases} d_{4,1}(N_{1,i}, N_{2,i}) = \dfrac{|S^C_{N_{2,i}} \bigcap S_{N_{1,i}}|}{|S_{N_{1,i}} \bigcup S_{N_{2,i}}|} = m_4(N_{1,i}, N_{2,i})(E) \\ d_{4,2}(N_{1,i}, N_{2,i}) = \dfrac{|S^{NC}_{N_{2,i}}|}{|S_{N_{2,i}}|} = m_4(N_{1,i}, N_{2,i})(NE) \end{cases} \tag{12}$$

Le reste de la masse sera associé à l'ignorance. La masse associée à l'ignorance peut être également dérivée des nœuds extrêmes qui n'ont pas des prédécesseurs (tous les nœuds sauf le nœud (5)) ou bien des successeurs (tous les nœuds sauf le nœud (1)).

Les équations (7), (8), (9), (10), (11) et (12) permettent de calculer les fonctions de masse par critère pour **chaque couple de nœuds** $(N_{1,i}, N_{2,i})$, respectivement du graphe de référence et du graphe correspondant à la réponse d'un participant, et d'attribut i. L'étape suivante définit une fonction de masse sur le **graphe tout entier** en faisant la moyenne des fonctions de masse sur tous les nœuds, calculées pour chaque critère :

$$\begin{cases} m_k(G_1, G_2)(E) = \dfrac{\displaystyle\sum_{i=1}^{O(G)} m_k(N_{1,i}, N_{2,i})(E)}{O(G)} \\ m_k(G_1, G_2)(NE) = \dfrac{\displaystyle\sum_{i=1}^{O(G)} m_k(N_{1,i}, N_{2,i})(NE)}{O(G)} \end{cases} \tag{13}$$

où $O(G)$ est l'ordre du graphe (*i.e.* le nombre de sommets, ici 6).
Afin d'obtenir une fonction de masse pour la réponse considérée, nous combinons les fonctions de masse des quatre critères. Finalement, le degré d'expertise est donné en calculant la distance de Jousselme et al. (2001) entre la fonction de masse ainsi obtenue et la fonction de masse catégorique sur l'élément expert tel que Essaid et al. (2014).

4 Évaluation de la méthode en situation réelle

Historiquement, Orange Labs réalise des tests d'évaluation subjective de codeurs audio en laboratoire. Ces tests consistent à recruter des auditeurs dits naïfs (n'étant pas impliqués directement dans les travaux liés à l'évaluation de la qualité ou du codage audio), à leur présenter de

courtes séquences de parole ou de musique traitées selon différentes configurations de codage et à leur demander d'en évaluer la qualité audio sur des échelles adaptées. Les tests se déroulent dans des salles traitées acoustiquement et plus globalement dans un environnement parfaitement contrôlé. Ces méthodes en laboratoire sont efficaces mais restent coûteuses et peuvent avoir une portée limitée quant à la représentativité des résultats (par rapport à une utilisation de services in situ) ou des stimuli (nombre limité par exemple).

Dans l'objectif d'ajouter l'approche crowdsourcing aux méthodes de test, deux campagnes déployées sur une plateforme de *crowdsourcing* ont été réalisées en vue de comparer les résultats avec ceux obtenus en laboratoire. Chaque campagne consistait en une réplique d'un même test initialement réalisé en laboratoire pour la normalisation du codeur G729EV. Dans ce test laboratoire, 7 conditions de test i.e. solutions de codage étaient considérées, auxquelles s'ajoutaient les 5 conditions de référence MNRU. Au total, douze conditions étaient évaluées à travers 16 extraits musicaux. 32 personnes ont été recrutées et réparties en 4 groupes. Chaque groupe écoutait et évaluait 4 *hits* , un *hit* (*Human Intelligence Task*) étant ici un ensemble de 12 séquences audio correspondant aux douze conditions de tests présentées à travers 12 extraits musicaux différents. Ainsi, chaque *hit* contenait les douze conditions de test (7 conditions de codage et les 5 MNRU) présentées dans un ordre aléatoire avec un set d'extraits musicaux différent pour chaque *hit*. Après chaque séquence audio, les auditeurs étaient invités à noter la qualité sur une échelle de 1 (= Mauvaise) à 5 (= Excellente).

Pour les campagnes de crowdsourcing, les participants ont également été répartis en **4 panels** de façon à ce que chacun ne puisse appartenir qu'à un et un seul panel, comme en laboratoire. Suivant le plan expérimental du test laboratoire, à chaque panel étaient associés 4 *hits* de 12 séquences audio à évaluer sur la même échelle de qualité. Chaque *hit* faisait l'objet d'un micro-job sur la plateforme de crowdsourcing. Ainsi chaque participant pouvait faire entre 1 et 4 *hit* (les 4 *hits* de son panel, différents de ceux des autres panels). La participation d'un participant a été prise en compte s'il ou elle avait terminé au moins un *hit*, sachant qu'il ou elle pouvait cesser les écoutes avant la fin du *hit* contrairement au laboratoire. Les instructions étaient présentées en anglais par écrit aux participants avant le test. Une session d'apprentissage avec 8 séquences audio devait également être réalisée avant le test, comme en laboratoire. Deux campagnes ont ainsi été menées en considérant deux zones géographiques différentes. Dans la première campagne, tout travailleur anglophone pouvait participer, quel que soit son pays. Les travailleurs ayant participé à cette campagne étaient majoritairement en Asie. La deuxième campagne se limitait aux USA. Les mêmes conditions sont appliquées dans le sens où les participants appartenant à un même $Panel_i$ écoutent les mêmes morceaux pour les deux campagnes.

Les degrés d'expertise ont été calculés d'une part sur les données de laboratoire et d'autre part sur les notes issues des plateformes de *crowdsourcing*. Pour les données du laboratoire, les résultats sont présentés par intervalles de degré d'expertise de longueur 0.1 sur la figure 3. On a obtenu 31 personnes (sur 32) avec un degré d'expertise supérieur à 0.4 (seuil choisi en comparaison des données issues de la plateforme) montrant ainsi la fiabilité des réponses des personnes en laboratoire. L'intervalle $[0.4, 0.5]$ contient le plus de personnes.

Dans un premier temps, les distributions des ensembles de tous les panels sur les données issues des plateformes de *crowdsourcing* sont représentées sur la figure 4. Sur ces distributions, nous notons un petit saut sur l'intervalle $[0.4, 0.5]$ qui permet ainsi de déterminer le seuil d'expertise à prendre en compte pour discriminer les participants.

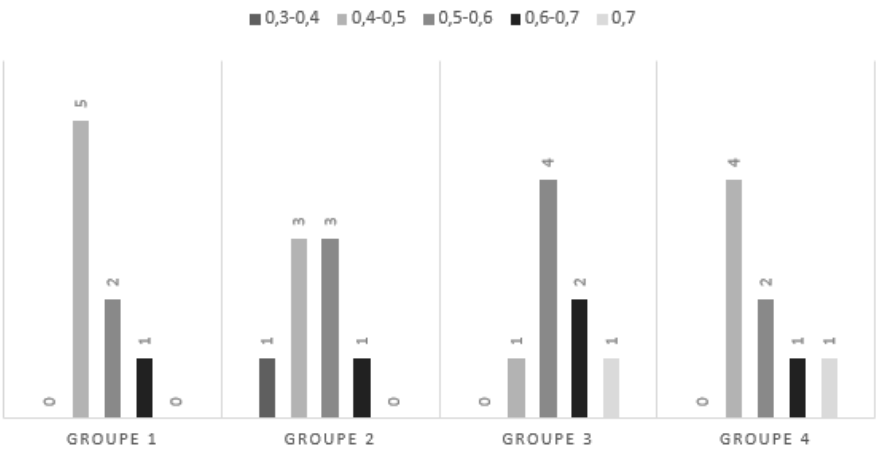


FIG. 3 – *Participants par intervalle d'expertise sur les données de laboratoire.*

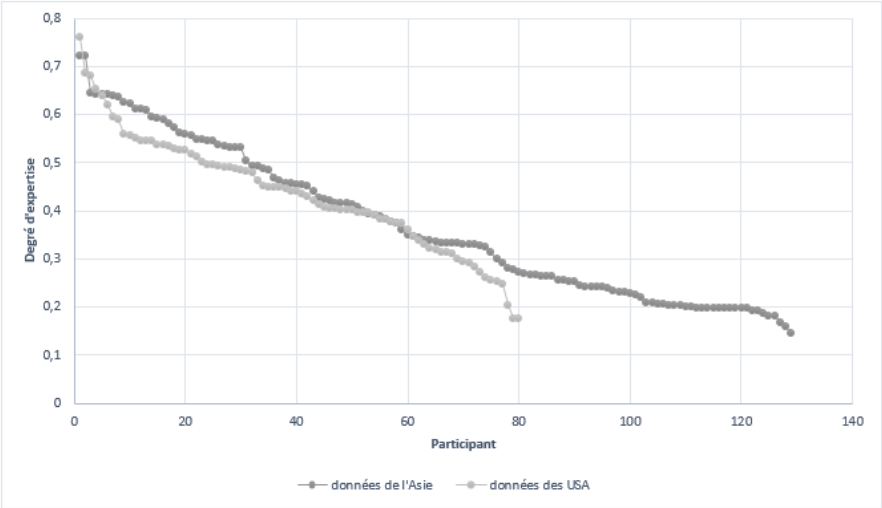


FIG. 4 – *Distribution des données*

Les degrés d'expertise varient dans un intervalle plus large que celui du laboratoire (de 0.1 à 0.7). Un premier facteur pouvant expliquer ces résultats serait le manque de sérieux chez un plus grand nombre de participants sur la plateforme de *crowdsourcing*. D'autre part, les conditions d'écoute (environnement sonore, casque ou haut-parleur(s), PC utilisé) sont variables d'un participant à l'autre, d'un *hit* à l'autre contrairement au laboratoire, et peuvent influer sur la qualité des réponses des participants. Dans ce travail, nous n'avons pas souhaité imposer des conditions d'expérimentation afin de placer les travailleurs dans un contexte familier. De plus, en comparant les deux distributions, on remarque une petite différence entre

les deux campagnes. Par exemple, l'intervalle d'expertise [0.1,0.2] est presque absent pour les données de la campagne américaine (2 personnes sur l'ensemble des panels *cf.* courbe grise) alors que pour l'Asie cet intervalle contient 19 participants (*cf.* courbe noire). D'autre part, pour l'Asie l'intervalle [0.2,0.3] contient le plus de participants alors que pour les USA, il s'agit de l'intervalle [0.4,0.5]. Les différences observées entre les deux campagnes peuvent s'expliquer à travers les différences culturelles entre les deux régions, et notamment une plus grande proximité culturelle avec les extraits musicaux choisis (musique occidentale) pour la campagne américaine.

Nous retenons comme seuils dans une première analyse, les seuils de 0.4 et 0.5 qui sont proches des sauts dans les distributions (4). Les participants retenus sont ceux ayant un degré d'expertise supérieur au seuil considéré. La moyenne de leurs réponses sera prise en compte pour l'évaluation de la qualité audio. Nous comparons ainsi les données issues des deux campagnes sur les plateformes de *crowdsourcing* avec celles obtenues en laboratoire selon ces deux seuils.

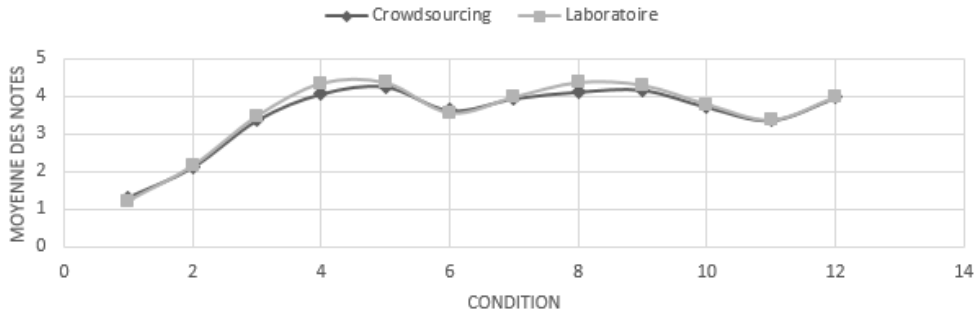


FIG. 5 – *Comparaison des courbes laboratoire/crowdsourcing des données fusionnées pour un seuil d'expertise de 0.4.*

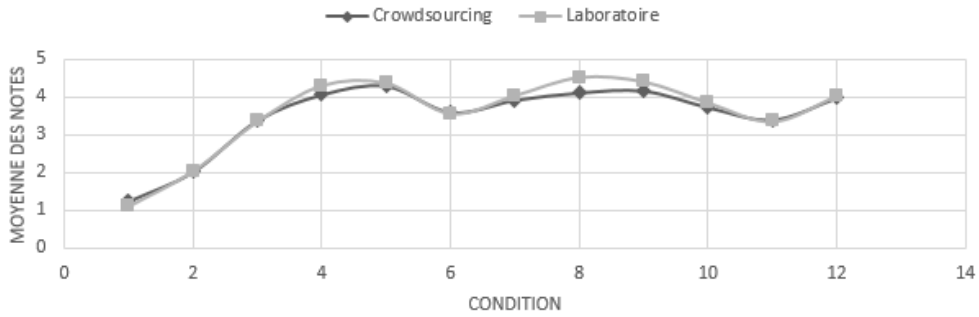


FIG. 6 – *Comparaison des courbes laboratoire/crowdsourcing des données fusionnées pour un seuil d'expertise de 0.5.*

Nous remarquons que les courbes obtenues (laboratoire et *crowdsourcing*) pour le premier seuil 0.4 (5) sont plus proches les unes des autres que pour un seuil à 0.5. Néanmoins, les

courbes obtenues pour le deuxième seuil 0.5 (6) montrent que les participants qualifiés d'experts selon ce critère ont bien réussi à différencier les 5 premières conditions de référence (MNRU) de qualité connue.

Idéalement, la courbe sur les cinq premiers morceaux devrait être une droite car elle correspond aux MNRUs. Cependant, ces résultats sont expliqués par des comportements habituels que l'on retrouve ici sur les données de laboratoire et des plateformes. La proximité des deux courbes montre l'intérêt de réaliser ce type d'évaluation sur des plateformes de *crowdsourcing*, une fois que les participants les plus experts aient été sélectionnés.

5 Conclusion et discussion

Dans ce travail nous avons proposé une approche originale de calcul d'expertise pour des participants à une évaluation subjective de qualité audio à travers des tests d'écoute. L'approche proposée est fondée sur une modélisation des notations des participants sur la forme de graphe. Tenant compte de données dont l'ordre de préférence attendu est connu, nous avons développé une mesure de comparaison de deux graphes. Ainsi, l'approche est fondée sur quatre critères à partir desquels quatre fonctions de masse ont été définies afin de tenir compte des imperfections possibles des réponses des participants. A partir de ces fonctions de masse, un degré d'expertise est calculé pour chaque participant permettant ainsi de ne considérer que les participants ayant un degré d'expertise suffisant.

Les résultats comparant des données issues de deux campagnes de *crowdsourcing* et de laboratoire montrent l'intérêt de réaliser de telles évaluations à partir de plateformes de *crowdsourcing*. Il est cependant nécessaire d'évaluer correctement un degré d'expertise afin de ne pas considérer toutes les réponses issues des plateformes de *crowdsourcing*. L'approche développée dans ce travail pour l'évaluation des degrés d'expertise permet bien d'écarter les participants sans réponses pertinentes pour la tâche d'évaluation de la qualité audio.

Références

Ben Rjab, A., M. Kharoune, Z. Miklos, et A. Martin (2016). Characterization of experts in crowdsourcing platforms. In *The 4th International Conference on Belief Functions*.

Ben Rjab, A., M. Kharoune, Z. Miklos, A. Martin, et B. Ben Yaghlane (2015). Caractérisation d'experts dans les plate-formes de crowdsourcing. In *24 ème Conférence sur la Logique Floue et ses Applications*.

Dawid, A. P. et A. M. Skene (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society 28*(1), 20–28.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, 325–339.

Essaid, A., A. Martin, G. Smits, et B. B. Yaghlane (2014). A distance-based decision in the credal level. In *Artificial Intelligence and Symbolic Computation - 12th International Conference, AISC 2014, Seville, Spain, December 11-13, 2014. Proceedings*, pp. 147–156.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine 14*(6), 1–4.

Ipeirotis, P. (2010). Worker evaluation in crowdsourcing : Gold data or multiple workers ?

Ipeirotis, P. G., F. Provost, et J. Wang (2010). Machine-learning for spammer detection in crowd-sourcing. In *HCOMP '10 Proceedings of the ACM SIGKDD Workshop on Human Computation*.

ITU (1996). Modulated noise reference unit (MNRU). Technical Report ITU-T P.810, International Telecommunication Union.

Jouili, S. (2011). *Indexation de masses de documents graphiques : approches structurelles*. Ph. D. thesis, Université Nancy II.

Jousselme, A.-L., D. Grenier, et É. Bossé (2001). A new distance between two bodies of evidence. *Information fusion 2*(2), 91–101.

Le, J., A. Edmonds, V. Hester, et L. Biewald (2010). Ensuring quality in crowdsourced search relevance evaluation : The effects of training question distribution. In *Workshop on Crowdsourcing for Search Evaluation*.

Philips, P. (2011). Enterprise crowdsourcing or : How i learned to stop worrying and trust the crowd.

Raykar, V. C. et S. Yu (2012). Annotation models for crowdsourced ordinal data. *Journal of Machine Learning Research 13*.

Raykar, V. C., S. Yu, L. H. Zhao, G. Hermosillo Valadez, C. Florin, L. Bogoni, et L. Moy (2010). Learning from crowds. *Journal of Machine Learning Research 11*, 1297–1322.

Shafer, G. et al. (1976). *A mathematical theory of evidence*, Volume 1. Princeton university press Princeton.

Smets, P. (1990). The combination of evidence in the transferable belief model. *12*, 447 – 458.

Smyth, P., U. Fayyad, M. Burl, P. Perona, et P. Baldi (1995). Inferring ground truth from subjective labelling of venus images. *Advances in Neural Information Processing Systems 7*, 1085–1092.

Summary

Crowdsourcing, a major economic issue, is the fact that the firm outsources internal task to the crowd. It is a form of digital subcontracting for the general public. The evaluation of the participants work quality is a major issue in crowdsourcing. Indeed, contributions must be controlled to ensure the effectiveness and relevance of the campaign. We are particularly interested in small, fast and not automatable tasks. Several methods have been proposed to solve this problem, but they are applicable when the "golden truth" is not always known. This work has the particularity to propose a method for calculating the degree of expertise in the presence of gold data in crowdsourcing. This method is based on the belief function theory and proposes a structuring of data using graphs. The proposed approach will be assessed and applied to the data.