L'ambiguïté dans la représentation des émotions : état de l'art des bases de données multimodales

Hélène Tran*,**, Lisa Brelet**, Issam Falih*,
Xavier Goblet**, Engelbert Mephu Nguifo*

*Université Clermont-Auvergne, CNRS, Mines de Saint-Étienne,
Clermont-Auvergne-INP, LIMOS, 63000 Clermont-Ferrand, France
helene.tran@doctorant.uca.fr, {issam.falih, engelbert.mephu_nguifo}@uca.fr
**Jeolis Solutions, 63000 Clermont-Ferrand, France
{helene.tran, lisa.brelet, xavier.goblet}@lojelis.com

Résumé. La reconnaissance des émotions est une brique fondamentale dans l'octroi de l'intelligence émotionnelle aux machines. Les premiers modèles ont été conçus pour reconnaître les émotions fortement exprimées et facilement identifiables. Cependant, nous sommes rarement en proie à ce type d'émotions dans notre vie quotidienne. La plupart du temps, nous éprouvons une difficulté à identifier avec certitude notre propre émotion et celle d'autrui : c'est l'ambiguïté émotionnelle. Les bases de données, à la racine du développement des systèmes de reconnaissance, doivent permettre d'introduire l'ambiguïté dans la représentation émotionnelle. Ce papier résume les principales représentations émotionnelles et propose un état de l'art des bases de données multimodales pour la reconnaissance des émotions, avec une étude de leur positionnement sur la problématique. Le papier poursuit sur une discussion de la possibilité de représenter l'ambiguïté des émotions à partir des bases de données sélectionnées.

1 Introduction

Dans le développement du cerveau humain, le néocortex, partie du cerveau nous dotant d'une incroyable capacité intellectuelle, s'est progressivement formé sur le système limbique, partie plus primitive et siège de nos émotions (Goleman, 1995). Cette évolution anatomique sur des millions d'années nous offre une explication sur le rôle crucial des émotions dans notre prise de décision. En effet, sans elles nous ne serions que des machines raisonnant uniquement avec la logique formelle. Nous serions incapables de faire un choix dans notre propre intérêt, en accordant de la valeur (positive ou négative) à chaque option possible.

A l'inverse, les machines ont pour fondement la raison et la logique, auxquelles les chercheurs ambitionnent de les doter d'une intelligence émotionnelle pour les intégrer dans notre vie quotidienne. Les applications sont diverses : une interaction homme-machine plus naturelle, le développement de systèmes de recommandation, la personnalisation de contenu pour entretenir la motivation à une tâche ou même la création d'un système expert pour le suivi de la santé mentale. La première brique pour atteindre ce but est la reconnaissance des émotions.

Pantic et al. (2005) ont défini les cinq qualités que doit idéalement avoir un système de reconnaissance des émotions : multimodal, robuste et précis, générique, sensible aux dynamiques, et contextuel. Les recherches récentes (Sethu et al., 2019) montrent l'importance d'ajouter la représentation de l'ambiguïté émotionnelle dans ces critères. L'ambiguïté et l'incertitude, bien qu'elles soient intimement liées, sont deux notions distinctes : alors que l'incertitude correspond à ce qui n'est pas sûr d'être observé, l'ambiguïté désigne un caractère équivoque, où l'émotion observée peut prêter à confusion. L'ambiguïté émotionnelle comprend aussi l'observation de plusieurs émotions. L'apprentissage d'un modèle de reconnaissance reposant sur une base de données, celle-ci doit permettre d'inclure l'ambiguïté dans le modèle émotionnel final. Par conséquent, un état de l'art des bases de données de reconnaissance multimodale (expression faciale, voix, et transcription textuelle) les plus utilisées est proposé, avec un accent sur leur positionnement par rapport à cette problématique.

Le plan de ce papier est le suivant : après avoir mis en contexte la problématique de l'ambiguïté des émotions (Section 2), nous présenterons les modèles émotionnels existants et l'introduction de l'ambiguïté dans ceux-ci (Section 3). Un état de l'art des bases de données trimodales pour la reconnaissance des émotions sera ensuite proposé (Section 4), à partir duquel nous discuterons de la possibilité de représenter l'ambiguïté émotionnelle (Section 5).

2 Contexte théorique

Pour communiquer, nous formons essentiellement un message, le plus souvent oral, avec nos propres mots pour transmettre à autrui nos idées, nos valeurs, ou encore nos émotions. Néanmoins, nous véhiculons également des informations à travers d'autres signes physiques tels que la prosodie (e.g., timbre de la voix, rythme d'élocution, etc.), l'expression faciale (e.g., lèvres pincées, sourcils levés, etc.), et la posture du corps (e.g., bras croisés, poings serrés, etc.). Mehrabian (1971) dans son livre *Silent Messages* affirme même que le contenu verbal représente seulement 7% de l'information reçue par autrui, tandis que 38% provient de la prosodie et 55% de l'expression faciale et de la gestuelle. Ainsi, lorsque nous étudions la communication, et plus particulièrement celle des émotions, il est important de s'intéresser et de distinguer les trois formes de communication présentées, à savoir : verbale (i.e., les mots formulés), para-verbale (i.e., prosodie) et non-verbale (i.e., expression faciale et posture du corps). En conséquence, un système de reconnaissance des émotions doit, dans la mesure du possible, intégrer les signaux issus des trois types de communication.

Selon le modèle de communication de Shannon et Weaver (1949), la communication serait constituée de trois éléments : la source du message, le message, et le destinataire du message. Dans une communication entre machines, le message de la source est encodé par un émetteur (e.g., caméra), puis transmis à travers un canal de communication susceptible d'être bruité (e.g. réseau internet), et enfin décodé par un récepteur (e.g., écran) permettant au destinataire de prendre connaissance du message. Ces étapes peuvent conduire à une perte d'information, due à la résolution de la caméra par exemple, et influer sur la qualité de reconnaissance du système. Dans le cadre d'une communication humaine, la source du message se rapporte à la personne qui s'exprime, le message correspond à ce qui est transmis verbalement et non-verbalement par celle-ci, et le destinataire se réfère à l'individu qui reçoit et interprète le message.

Si nous étudions la communication sous le prisme des émotions, l'identification, l'expression et la reconnaissance de celles-ci peuvent parfois s'avérer délicates et biaisées. La no-

tion d'ambiguïté des émotions traduit cette complexité tridimensionnelle et se retrouve aux trois niveaux du modèle de communication de Shannon et Weaver (1949). Premièrement, à la source du message, une ambiguïté peut résider dans les **émotions ressenties** par la personne qui s'exprime. Il est en effet difficile d'identifier sa propre émotion avec certitude, cette dernière relevant souvent de l'inconscient. Deuxièmement, au niveau du contenu du message, une ambiguïté peut exister dans les **émotions exprimées**. Comme nous l'avons décrit, l'expression d'une émotion peut emprunter plusieurs canaux (i.e., verbal, para-verbal, et non-verbal), complexifiant ainsi le message transmis. Troisièmement, au niveau du destinataire, une ambiguïté peut être constatée dans les **émotions perçues**. Différentes interprétations d'un même message émotionnel peuvent être observées. L'absence fréquente d'accord unanime entre annotateurs sur l'émotion observée, notamment lorsque celle-ci n'est pas intense, montre la différence de perception des émotions et rappelle que l'émotion reste une quantité subjective. Le modèle de lentille fonctionnelle de Brunswick apporte une explication aux différences de perception (Scherer, 2003) : chaque personne porte un jugement sur l'émotion observée, en accordant plus ou moins d'importance à chaque signal de communication présent. Mathématiquement, l'émotion perçue peut se concevoir comme une combinaison linéaire des signaux observés. Un poids de confiance est attribué à chaque signal et varie d'une personne à une autre selon sa propre expérience affective. Ainsi, l'expression faciale aura plus de poids dans l'évaluation émotionnelle d'un premier annotateur, un autre sera plus attentif au style d'élocution.

A l'image de l'homme, la tâche attribuée au modèle pour identifier l'émotion ressentie à partir de l'émotion observée n'est pas aisée, ces trois niveaux d'ambiguïté devant être levés. Estimer les émotions ressentie et exprimée nécessite une étude du profil psychologique de l'individu, où il faudra notamment identifier les facteurs psychologiques ayant un impact sur la reconnaissance des émotions, tels que l'alexithymie (difficulté à identifier et réguler ses propres émotions). Cependant, il n'existe à notre connaissance aucune base de données permettant de faire cette étude. Par conséquent, un objectif actuellement atteignable serait de développer un système automatique capable de reconnaître une représentation ambiguë de l'émotion perçue. Deux sous-niveaux d'ambiguïté issus de la perception émotionnelle ont été identifiés :

— L'**ambiguïté d'identification**, qui concerne directement la représentation finale des émotions (émotion ambiguë ou pas de certitude sur l'émotion observée).
— L'**ambiguïté de perception entre annotateurs**, liée à la différence de perception des émotions, où deux personnes peuvent être en désaccord sur l'émotion perçue.

Avant d'étudier la représentation de l'ambiguïté émotionnelle dans les bases de données, les principales représentations émotionnelles et les deux types d'annotation sont présentées.

3 Représentations des émotions existantes

Les approches discrète et continue sont les deux familles de représentation des émotions. Il existe deux méthodes pour les annoter : absolues (évaluation indépendante de chaque échantillon) ou ordinales (évaluation basée sur des éléments de référence). La section 3.1 décrira les représentations discrète et continue, et la section 3.2 les annotations absolues et ordinales. Leurs limites vis-à-vis de la notion d'ambiguïté seront discutées dans la section 3.3. La section 3.4 présentera les principaux travaux effectués pour intégrer l'ambiguïté.

3.1 Représentation des émotions

Représentation discrète. Les émotions sont représentées par des états affectifs discrets. Cette approche constitue le moyen le plus naturel pour l'homme de définir ses propres émotions. Selon la perspective évolutionniste, il existe deux grandes catégories d'émotions : les émotions primaires et les émotions secondaires.

- **Émotions primaires** : il s'agit d'émotions innées et universelles, de courte durée et rapidement déclenchées en réaction à un stimulus émotionnel. La liste des émotions primaires varie fortement selon les disciplines et les auteurs. Celle qui reste la plus communément utilisée est celle d'Ekman (1992) où elles seraient au nombre de six : la peur, la colère, la joie, la tristesse, le dégoût et la surprise. Un autre exemple est la roue des émotions de Plutchik (2001). Elle comprend huit émotions primaires regroupées en quatre paires d'émotions opposées : la joie et la tristesse, la confiance et le dégoût, la peur et la colère, l'anticipation et la surprise.
- **Émotions secondaires** : elles découlent des sentiments issus des émotions primaires, sont acquises par l'apprentissage et par la confrontation à la réalité, et sont dépendantes de la culture. Par exemple, après avoir ressenti de la colère, nous pouvons ensuite éprouver de la honte ou de la culpabilité. Les émotions secondaires sont généralement des combinaisons d'émotions primaires (Nugier, 2009).

Les émotions primaires sont fréquemment choisies comme classes pour le développement de modèles de reconnaissance, évitant ainsi un possible chevauchement d'états émotionnels. Bien que l'approche discrète reste largement privilégiée dans la recherche du fait de son caractère intelligible, se limiter à la reconnaissance d'une émotion primaire ne suffit pas pour décrire tout le spectre des émotions.

Représentation continue. Les émotions sont placées dans un espace à trois dimensions. Une grande majorité de chercheurs travaillant sur le modèle continu s'accordent sur les deux premières dimensions : la valence et l'activation.

- La **valence** correspond au caractère plaisant de l'émotion. A titre d'exemple, le dégoût est une émotion désagréable et est donc associé à une valence négative. A l'opposé, une valence positive est reliée à une émotion plaisante telle que la joie.
- L'**activation** renvoie à l'intensité physiologique ressentie. La tristesse, qui entraîne un repli sur soi, est associée à une activation faible. L'activation de la colère est élevée car elle libère un afflux d'énergie dans le corps de l'individu.

Or, ces deux dimensions ne suffisent pas pour caractériser l'émotion. Par exemple, la colère et la peur ont toutes les deux une valence négative et une activation élevée, nécessitant ainsi une troisième dimension pour les différencier.

Toutefois, il n'existe pas de consensus sur celle-ci. Selon Russell et Mehrabian (1977), il pourrait s'agir du **contrôle**. Ainsi, un individu en colère se sent en contrôle de la situation (contrôle positif) tandis qu'une personne submergée par la tristesse semble voir la situation lui échapper (contrôle négatif). Le modèle résultant est nommé PAD (*Pleasure, Arousal, Dominance*) et serait suffisant selon les auteurs pour décrire toutes les émotions.

Outre une meilleure résolution temporelle, l'approche continue permet de représenter un large éventail d'états émotionnels et de gérer les variations de ceux-ci au cours du temps (Gunes et al., 2011). Du fait de ces avantages, cette approche suscite un intérêt grandissant en informatique affective.

3.2 Annotation des émotions

Annotation absolue. Cette méthode consiste à annoter chaque échantillon de manière indépendante, sans points de référence. Pour l'approche discrète, une émotion est choisie parmi une liste d'émotions préétablie. Dans le cas continu, une valeur absolue est donnée à chaque dimension de l'émotion à un instant donné.

Annotation ordinale. Il s'agit ici d'évaluer l'émotion discrète ou la dimension en la comparant avec des éléments de référence. Ceux-ci ont un ordre bien défini. Yannakakis et al. (2017) présente une argumentation complète sur la fiabilité et la validité de la méthode. Les approches les plus couramment employées pour ce type d'annotation sont :

— **Echelle de Likert** : très utilisée en recherche pour l'élaboration de questionnaires, elle permet de mesurer une quantité subjective en choisissant parmi plusieurs options ordonnées. Par exemple, le degré de présence d'une émotion primaire peut être déterminé par une échelle de type Likert en 4 points *absent, peu présent, présent, très présent*.
— ***Self-Assessment Manikins* (SAM)** : proposé par Bradley et Lang (1994) et principalement destiné aux enfants, des illustrations de forme humaine sont utilisées pour décrire les intensités de chaque dimension de la représentation continue.
— **Roue des émotions de Genève** : développée par Scherer (2005), elle a la particularité d'intégrer les deux types de représentation émotionnelle en plaçant différentes intensités d'émotions discrètes dans l'espace bidimensionnel des émotions.

3.3 Limites des représentations émotionnelles actuelles

Les émotions sont habituellement représentées de façon ponctuelle. Pour les annotations absolues, les échantillons sont associés à une seule émotion dans le cas discret. Dans le modèle continu, l'émotion est représentée par un point se déplaçant dans le temps dans l'espace multidimensionnel. Concernant les annotations ordinales, une seule intensité est associée à une émotion (modèle discret) ou à une dimension (modèle continu).

Choisir une représentation ponctuelle revient donc à être certain sur la nature de l'émotion perçue. L'ambiguïté émotionnelle n'y est pas prise en compte. En reprenant les deux sous-niveaux d'ambiguïté de la perception émotionnelle présentés à la fin de la section 2 :

— **Ambiguïté d'identification** : il n'est pas possible d'avoir deux émotions différentes, de définir une zone dans l'espace multidimensionnel plutôt qu'un point, ou de choisir deux options ordonnées.
— **Ambiguïté de perception entre annotateurs** : un conflit d'annotation est géré par élimination des échantillons problématiques ou par un vote à la majorité (dans le cas absolu/discret et ordinal) ou un calcul de la moyenne (dans le cas absolu/continu).

3.4 Tentatives d'intégration de l'ambiguïté émotionnelle

Certains chercheurs ont proposé des approches pour intégrer l'ambiguïté dans le modèle émotionnel. Sethu et al. (2019) a réalisé une étude complète sur l'introduction de l'ambiguïté dans la représentation des émotions. Un résumé des principales méthodes est ici proposé.

Annotation absolue. La représentation ambiguë diffère selon l'approche choisie :

— Représentation discrète : une seconde émotion peut être autorisée pour représenter l'émotion observée. En outre, Vidrascu et Devillers (2005) proposent la notion d'émotions majeure et mineure. Par extension, un profil émotionnel peut être établi où le niveau de présence de chaque émotion primaire est estimé (Mower et al., 2010).

— Représentation continue : au lieu de la définir par un point dans l'espace, l'émotion peut être modélisée par une fonction gaussienne en lui associant la moyenne et l'écart-type (Han et al., 2017). Dang et al. (2017) proposent de ne pas se restreindre à la distribution gaussienne en utilisant un modèle de mélange gaussien.

Annotation ordinale. Cette méthode implique le choix d'une option discrète parmi d'autres, que ce soit pour le modèle discret ou le modèle continu. Une solution est de permettre aux annotateurs de choisir plusieurs options plutôt qu'une. La diversité des annotations peut également être reflétée par une distribution des annotations sur les éléments de référence.

4 État de l'art des bases de données

Le choix de la base de données est l'étape sur laquelle repose le développement entier d'un système de reconnaissance des émotions. Il va avoir un rôle essentiel dans la définition du modèle émotionnel, c'est pourquoi ce choix doit se faire avec soin en accord avec les objectifs finaux. Pour répondre à cette problématique, un état de l'art des bases de données les plus utilisées en reconnaissance des émotions est proposé. Une attention particulière est donnée aux aspects qui vont permettre d'intégrer l'ambiguïté dans la représentation des émotions. Par ailleurs, l'expression faciale de l'individu, sa voix, et les mots formulés donnent des informations complémentaires, parfois contradictoires, sur les émotions : elles devront donc être prises en compte pour reconnaître efficacement l'émotion (cf. Section 2). Nous allons en conséquence nous intéresser uniquement aux bases de données contenant ces trois modalités. Toutes les bases de données sélectionnées pour ce papier (N=8) sont en anglais, hormis CMU-MOSEAS qui comprend le français comme langue d'élocution. Celles qui ont été retenues sont les suivantes :

— **IEMOCAP** (Busso et al., 2008) : base de données encore très utilisée dans la recherche en informatique affective. Les émotions sont jouées en studio par des acteurs professionnels et sont représentées avec les deux approches discrète et continue. Elle fournit également des informations sur l'expression faciale et les mouvements de tête et de main, toutes obtenues à l'aide de marqueurs physiques.

— **SEMAINE** (McKeown et al., 2012) : des volontaires interagissent avec un personnage artificiel à qui un trait de caractère a été attribué (en colère, heureux, morose, sensible). Dans l'un des trois scénarios proposés, des opérateurs humains commandent ces personnages. Seul ce scénario a été retenu pour cet état de l'art car, en plus de posséder le plus grand nombre d'annotations, toutes ses sessions ont été transcrites.

— **CMU-MOSEI** (Zadeh et al., 2018) : rapidement incontournable dans la recherche, ses principaux atouts sont sa grande taille, une diversité de sujets traités et une multiplicité de locuteurs. Chaque émotion discrète est annotée sur son degré de présence à l'aide d'une échelle de type Likert en 4 points allant de 0 (= absence de l'émotion) à 3 (= une forte présence de l'émotion). Elle contient uniquement des monologues face caméra.

— **OMG-Emotions** (Barros et al., 2018) : elle a été conçue pour prendre en compte le contexte dans l'expression émotionnelle. Les émotions, représentées avec les deux approches discrète et continue, ont été annotées de façon graduelle (sans changement brusque) dans le temps. En outre, Sutherland et al. (2021) démontre que cette base de données offre une meilleure robustesse que IEMOCAP.
— **MELD** (Poria et al., 2018) : constituée d'extraits de la série télévisée *Friends*, il s'agit de l'une des plus larges bases impliquant plus de deux personnes dans une conversation.
— **SEWA** (Kossaifi et al., 2019) : deux personnes ordinaires de même culture discutent de publicités par appel vidéo. Les langues présentes sont l'anglais, l'allemand, le hongrois, le grec, le serbe et le chinois. Outre la reconnaissance des émotions, elle a été également créée pour l'analyse du comportement humain, incluant des études culturelles.
— **CMU-MOSEAS** (Zadeh et al., 2020) : large base de données en français, espagnol, portugais et allemand et développée par les mêmes auteurs que CMU-MOSEI. Elle possède les mêmes attributs que cette dernière, avec des annotations supplémentaires sur 12 traits de caractère tels que confiant et nerveux, ainsi qu'une variable binaire indiquant si une opinion est exprimée ou non dans la vidéo.
— **MuSe-CaR** (Stappen et al., 2020) : composée de critiques automobiles issues de la plateforme YouTube, les vidéos ont des caractéristiques dites *in-the-wild* (bruit ambiant, visage non face caméra, vocabulaire technique, etc). Elle est à notre connaissance la plus large base de données incluant les modalités visuelle, vocale, et textuelle.

Un état de l'art complet des bases de données de reconnaissance multimodale des émotions est proposé au lien suivant : `https://github.com/helenetran3/SOTA-Multimodal-Emotion-Recognition`. Il contient des informations utiles (e.g., taille, émotions, langues) pour pouvoir faire un choix. Le tableau 1 décrit le positionnement des huit bases de données présentées sur l'introduction de l'ambiguïté émotionnelle.

Ambiguïté d'identification. Elle permet de savoir s'il y a eu une tentative de représenter l'ambiguïté lors de l'identification finale des émotions.

— "*Modèle, Annotation*" : le modèle émotionnel est soit discret soit continu. L'annotation peut être soit absolue soit ordinale. Pour plus d'informations, se référer à la section 3.
— "*Annotation finale*" décrit l'annotation finale de chaque phrase après avoir agrégé les annotations. Pour les annotations absolues, cela peut être une ou plusieurs émotions par phrase pour le cas discret, un point ou une zone se déplaçant dans l'espace pour le cas continu. Pour les annotations ordinales, il peut s'agir d'une ou plusieurs options.

Cinq bases de données proposent une représentation discrète, cinq autres l'approche continue. Toutes les bases de données excepté IEMOCAP version continue, CMU-MOSEI et CMU-MOSEAS ont choisi une annotation absolue. Elles se sont limitées à une émotion par phrase dans le cas discret et un point dans l'espace dans le cas continu. IEMOCAP version continue est annotée de façon ordinale, où un seul élément du SAM peut être choisi par dimension. Les deux bases du CMU ont adopté une représentation discrète avec une annotation ordinale. Contrairement à toutes les autres, plusieurs émotions peuvent être reconnues dans une même phrase, à l'aide de l'échelle de Likert qui estime le degré de présence de chaque émotion.

Ambiguïté de perception entre annotateurs. Les annotations collectées reflètent les différences de perception pouvant subsister entre les personnes observant l'émotion.

	Ambiguïté d'identification		**Ambiguïté de perception annot.**	
	Modèle, Annotation	Annotation finale	# annotations par phrase	Agrégation annotations
IEMOCAP (2008)	Discret, Absolue	1 émotion par phrase	3	Vote
	Continu, Ordinale	1 élément du SAM	2	Cote standard
SEMAINE (2012)	Continu, Absolue	Point dans l'espace	Entre 3 et 8	NC
CMU-MOSEI (2018)	Discret, Ordinale	Plusieurs émotions possibles	3	NC
OMG-Emotions (2018)	Discret, Absolue	1 émotion par phrase	5	Vote
	Continu, Absolue	Point dans l'espace	5	Evaluateur Estimateur Pondéré
MELD (2018)	Discret, Absolue	1 émotion par phrase	3	Vote
SEWA (2019)	Continu, Absolue	Point dans l'espace	Au moins 3	Distorsion Temporelle Canonique
CMU-MOSEAS (2020)	Discret, Ordinale	Plusieurs émotions possibles	3	NC
MuSe-CaR (2020)	Continu, Absolue	Point dans l'espace	5	Evaluateur Estimateur Pondéré

TAB. 1 – *Critères d'introduction de l'ambiguïté émotionnelle dans les bases de données trimodales (visuelle, vocale, textuelle) étudiées. Ambiguïté de perception annot. pour Ambiguïté de perception entre annotateurs. # pour Nombre. NC pour Non Communiqué.*

— "*# annotations par phrase*" est le nombre d'annotations par phrase
— "*Agrégation annotations*" est la méthode utilisée pour agréger les différentes annotations et former l'annotation finale

Le nombre d'annotations varie entre 2 et 8. Les méthodes d'agrégation des annotations diffèrent selon le modèle émotionnel choisi : le vote à la majorité est privilégié dans le cas discret, alors que les stratégies varient pour le cas continu. En effet, IEMOCAP choisit de faire une

simple normalisation avec la cote standard. OMG-Emotions et MuSe-CaR tiennent compte de la confiance accordée à chaque annotateur en utilisant la méthode de l'évaluateur estimateur pondéré. La base SEWA utilise la distorsion temporelle canonique pour agréger les annotations : les valeurs finales sont obtenues à l'aide d'un sous-espace dans lequel les annotations sont maximalement corrélées entre elles.

En résumé, les bases de données IEMOCAP, OMG-Emotions, MELD, CMU-MOSEI et CMU-MOSEAS ont choisi le modèle émotionnel discret : les trois premières attribuent une seule émotion par phrase tandis que les bases du CMU permettent de classifier les vidéos en plusieurs émotions avec chacun un degré de présence. Les annotations sont agrégées par vote. Trois annotations sont effectuées par phrase excepté pour OMG-Emotions où elles sont cinq.

Les bases de données IEMOCAP, SEMAINE, OMG-Emotions, SEWA et MuSe-CaR ont préféré le modèle continu. Toutes ont choisi une représentation ponctuelle de l'émotion dans l'espace multidimensionnel (point dans l'espace des émotions, excepté pour IEMOCAP avec un élément du SAM). IEMOCAP choisit une normalisation par la cote Z pour agréger les 2 annotations par phrase, SEWA la distorsion temporelle canonique pour au moins 3 annotations, et OMG-Emotions et MuSe-CaR l'évaluateur estimateur pondéré pour 5 annotations. SEMAINE, avec 3 à 8 annotations par phrase, n'a pas communiqué de méthode d'agrégation.

5 Discussion

L'objectif de cet état de l'art est d'analyser le positionnement des bases de données trimodales (visuelle, vocale, et textuelle) sur la représentation de l'ambiguïté autour des émotions. Deux formes d'ambiguïté ont été analysées : l'identification de l'émotion perçue et les différences de perception. D'une part, seules les bases du CMU ont introduit l'ambiguïté d'identification de l'émotion en permettant à chaque annotateur de choisir plusieurs émotions primaires avec un certain degré de présence. Les autres ont représenté l'émotion de façon ponctuelle (une émotion par phrase pour le cas discret, un point dans l'espace ou un élément du SAM pour le cas continu). D'autre part, pour gérer les conflits d'annotation, les bases de données avec des émotions discrètes utilisent un vote à la majorité. Les différences de perception sont mieux considérées par celles adoptant l'approche continue, soit en prenant en compte la confiance accordée à l'annotateur (Evaluateur Estimateur Pondéré), soit en construisant un sous-espace dans lequel les annotations sont corrélées au maximum (Distorsion Temporelle Canonique).

L'ambiguïté d'identification des émotions peut être prise en compte uniquement lorsque le modèle émotionnel choisie tient lui-même compte de l'ambiguïté. C'est notamment le cas pour les bases du CMU. En revanche, si la méthode choisie pour gérer les conflits d'annotation est jugée insuffisante pour représenter l'ambiguïté de perception entre annotateurs, alors un travail peut être effectué à partir des annotations fournies. Une interrogation qui en découle est de savoir si le nombre d'annotations est assez grand pour refléter la diversité de perception.

Finalement, l'étude proposée se concentre uniquement sur les bases de données comprenant les modalités visuelle, vocale, et textuelle et prenant l'anglais ou le français comme langue d'élocution. Cet état de l'art ne reflète donc pas nécessairement le positionnement global de toutes les bases de données existantes sur l'ambiguïté autour de l'émotion perçue. Toutefois, nous observons que les bases de données récentes tentent d'inclure l'ambiguïté d'identification des émotions (bases du CMU) dans le modèle émotionnel final et d'améliorer la représenta-

tion de l'ambiguïté de perception entre annotateurs (cas continu). Cela souligne une prise de conscience grandissante du problème de l'ambiguïté émotionnelle dans la recherche affective.

6 Conclusion et perspectives

Lorsque nous interagissons, nous émettons des signaux physiques indiquant à autrui l'émotion que nous ressentons. L'expression de l'émotion varie d'un individu à un autre et selon les situations sociales. La perception de l'émotion diffère également selon les personnes, selon leur propre expérience affective. Nous pouvons même éprouver des difficultés à identifier l'émotion de notre interlocuteur. La communication humaine génère donc une ambiguïté quant à l'expression et la perception de l'émotion, et cela doit être reflété par les modèles de représentation émotionnelle. C'est dans cette optique qu'un état de l'art sur les bases de données multimodales, à la racine du développement d'un tel système, a été effectué. Seules les bases comprenant les modalités visuelle, vocale, et textuelle ont été sélectionnées. La reconnaissance de l'émotion ressentie et l'émotion exprimée nécessitant le profil psychologique du locuteur, l'émotion perçue a été uniquement étudiée dans la suite.

Deux formes d'ambiguïté ont été déterminées autour de l'émotion perçue : l'identification de l'émotion et la différence de perception entre annotateurs. L'état de l'art proposé se base sur plusieurs critères déterminant le positionnement des bases de données sur ces deux types d'ambiguïté. A l'heure actuelle, seules les bases CMU-MOSEI et CMU-MOSEAS ont tenté de lever l'ambiguïté d'identification en permettant l'annotation de plusieurs émotions primaires pour une même phrase. Quant à l'ambiguïté de perception entre annotateurs, les méthodes d'agrégation des annotations sont plus élaborées dans le cas continu que le cas discret. Toutefois, la représentation émotionnelle finale reste ponctuelle dans les bases de données adoptant l'approche continue, la variabilité de perception étant uniquement contenue dans ce point.

Finalement, les bases de données les plus récentes ont tenté d'inclure l'ambiguïté dans leur modèle émotionnel, dénotant une prise de conscience du problème. Cela reste toutefois timide : il convient donc de représenter davantage l'ambiguïté dans une nouvelle base de données, à l'aide des travaux présentés dans la section 3.4. Une étude pourrait également être réalisée pour tenter de reconnaître cette fois l'émotion ressentie. La nouvelle base de données devra décrire le profil psychologique de l'individu, en identifiant des facteurs qui influencent l'expression des émotions tels que l'aisance face caméra et l'alexithymie (difficulté à nommer les émotions).

Remerciements

Ce travail de recherche est effectué dans le cadre d'un contrat CIFRE. Nous souhaitons vivement remercier l'ANRT pour leur soutien financier, Baraa Mohamad et Yoren Gaffary pour leur participation aux discussions autour du sujet, les membres du groupe de recherche Miners du LIMOS pour leur soutien, ainsi que les relecteurs pour leur retour constructif.

Références

Barros, P., N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, et S. Wermter (2018). The OMG-Emotion Behavior Dataset. In *2018 International Joint Conference on Neural*

Networks (IJCNN), pp. 1–7. IEEE.

Bradley, M. M. et P. J. Lang (1994). Measuring Emotion : the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry 25*(1), 49–59.

Busso, C., M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, et S. S. Narayanan (2008). IEMOCAP : Interactive emotional dyadic motion capture database. *Language Resources and Evaluation 42*(4), 335–359.

Dang, T., V. Sethu, J. Epps, et E. Ambikairajah (2017). An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression. In *INTERSPEECH*, pp. 1248–1252.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition & Emotion 6*(3-4), 169–200.

Goleman, D. (1995). *Emotional Intelligence*. Bantam Books.

Gunes, H., B. Schuller, M. Pantic, et R. Cowie (2011). Emotion representation, analysis and synthesis in continuous space : A survey. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 827–834. IEEE.

Han, J., Z. Zhang, M. Schmitt, M. Pantic, et B. Schuller (2017). From Hard to Soft : Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 890–897.

Kossaifi, J., R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, et M. Pantic (2019). SEWA DB : A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1022–1040.

McKeown, G., M. Valstar, R. Cowie, M. Pantic, et M. Schroder (2012). The SEMAINE Database : Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing 3*(1), 5–17.

Mehrabian, A. (1971). *Silent Messages*. Wadsworth Publishing Company, Inc., Belmont, CA.

Mower, E., M. J. Matarić, et S. Narayanan (2010). A Framework for Automatic Human Emotion Classification Using Emotion Profiles. *IEEE Transactions on Audio, Speech, and Language Processing 19*(5), 1057–1070.

Nugier, A. (2009). Histoire et grands courants de recherche sur les émotions. *Revue électronique de Psychologie Sociale 4*(4), 8–14.

Pantic, M., N. Sebe, J. F. Cohn, et T. Huang (2005). Affective Multimodal Human-Computer Interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 669–676.

Plutchik, R. (2001). The Nature of Emotions : Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist 89*(4), 344–350.

Poria, S., D. Hazarika, N. Majumder, G. Naik, E. Cambria, et R. Mihalcea (2018). MELD : A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *arXiv preprint arXiv :1810.02508*.

Russell, J. A. et A. Mehrabian (1977). Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality 11*(3), 273–294.

Scherer, K. R. (2003). Vocal communication of emotion : A review of research paradigms. *Speech Communication 40*(1), 227–256.

Scherer, K. R. (2005). What are emotions ? And how can they be measured ? *Social Science Information 44*(4), 695–729.

Sethu, V., E. M. Provost, J. Epps, C. Busso, N. Cummins, et S. Narayanan (2019). The Ambiguous World of Emotion Representation. *arXiv preprint arXiv :1909.00360*.

Shannon, C. et W. Weaver (1949). The mathematical theory of communication. *University of Illinois Press*.

Stappen, L., A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, et al. (2020). MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pp. 35–44.

Sutherland, A., S. Magg, C. Weber, et S. Wermter (2021). Analyzing the Influence of Dataset Composition for Emotion Recognition. *arXiv preprint arXiv:2103.03700*.

Vidrascu, L. et L. Devillers (2005). Real-life Emotion Representation and Detection in Call Centers Data. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 739–746. Springer.

Yannakakis, G. N., R. Cowie, et C. Busso (2017). The Ordinal Nature of Emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 248–255. IEEE.

Zadeh, A. B., Y. S. Cao, S. Hessner, P. P. Liang, S. Poria, et L.-P. Morency (2020). CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1801–1812.

Zadeh, A. B., P. P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, et L.-P. Morency (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246.

Summary

Emotion recognition is a core element in the development of emotional intelligence in machines. Early models are designed to recognise full-blown emotions which are easily identifiable. However, we rarely experience these types of emotion in our daily lives. Identifying our own and others' emotions with confidence is often difficult: this is called emotional ambiguity. Databases are the building blocks of the development of emotion recognition systems, hence they should introduce ambiguity in emotional representation. This paper summarises the main emotional representations and proposes a literature review of the most used multimodal databases in emotion recognition, with a study of their position on the problem. The paper further discusses the possibility of representing emotion ambiguity from the selected databases.