# Détection des émotions à partir d'indices lexicaux, dialogiques et prosodiques dans le dialogue oral

L. Devillers [(1)], I. Vasilescu[(2)]

[(1)] LIMSI-CNRS, BP133, 91 403 Orsay Cedex, France, [(2)] LTCI-ENST, 46, rue Barrault, 75013
devil@limsi.fr, vasilesc@tsi.enst.fr

## ABSTRACT

This paper deals with emotion detection in spoken dialogs. Detecting emotions in the context of automated call center services can be helpful for the management of the human-computer dialogs, enabling dynamic modification of the dialog strategy according to the user behaviour and influencing the final outcome. In the present study, we make use of an Agent/Client dialog corpus recorded in a Stock Exchange Center in the framework of the Amities project. In our corpus recorded in real-life conditions the manifestation of emotion is complex, i.e. shaded emotions occur since the interlocutors attempt to control the expression of their internal attitude. Firstly, we aim at validating appropriate emotion labels for automated call center services and at validating them via perceptual tests. Secondly, we focus on multi-level detection cues, i.e. lexical and prosodic, as speaker employs complex strategies to manifest their emotions.

In this paper we describe four studies. The first study describes the annotation methodology and the strategy adopted to validate the emotion labels. The second study focuses on emotion detection with lexical cues, whereas the third one concerns the role of prosodic cues in emotion detection. In the fourth, we discuss the correlation between the emotion labels and the dialogic acts. The final aim of the study is to provide a complex detection model including several levels of information.

## 1. INTRODUCTION

Modéliser et détecter les émotions indiquant des troubles dans la communication peut être un moyen d'améliorer les systèmes de dialogue homme-machine. En effet, détecter des émotions peut permettre de suivre l'évolution des interactions, de modifier dynamiquement les stratégies dialogiques et donc de contribuer au succès de la communication. Notre objectif est de trouver des indices robustes à différents niveaux linguistiques: prosodique, lexical et dialogique, pour identifier les émotions dans des échanges verbaux.

Selon Scherer [10], le premier problème dans l'analyse des émotions est lié à la difficulté d'isoler les facteurs qui en sont responsables, dans la mesure où la manifestation des émotions est dépendante de la personnalité, des attitudes, de l'état d'esprit etc. des individus. Dans les interactions verbales spontanées, les émotions sont rarement manifestées à l'état pur et sous une forme primaire. En revanche, on retrouve dans ce type d'interactions des émotions plus mesurées, souvent combinées entre elles, qui sont difficiles à extraire, décrire et détecter. Afin de contrôler cette variabilité inhérente au domaine, la plupart des études consacrées à l'analyse des émotions dans la parole fait référence à un nombre minimal d'émotions dites fondamentales (colère, peur, tristesse, joie), voire à une opposition entre émotions négatives vs positives [2, 9] ou encore à une parole produite dans des conditions de stress ou non [8].

De plus, les comportements émotionnels sont fortement dépendants des corpus utilisés. La majorité des travaux ont pour l'instant porté sur des corpus artificiels (acteurs, Magiciens d'Oz) ou le niveau sémantique et lexical est contrôlé et les marqueurs d'émotions se retrouvent essentiellement au niveau prosodique. Par conséquent, il est souvent difficile de transposer les résultats obtenus sur des corpus artificiels à des corpus réels. En effet, la réalité langagière est beaucoup plus complexe et l'analyse des manifestations émotionnelles dans la parole spontanée se doit de considérer le but final de la démarche qui est d'intégrer ses résultats dans une application dialogique réelle. Les travaux que nous menons ont comme objectif final de réaliser un modèle de détection automatique des émotions multi-niveaux en intégrant des indices dialogiques, lexicaux et prosodiques. Ces recherches sont menées dans le cadre du projet IST AMITIES (*Automated Multilingual Interaction with Information and Services*) [1] et font appel à un corpus de dialogues réels entre clients et agents enregistrés dans un centre de transactions boursières.

Dans les paragraphes suivants, le corpus, la stratégie d'annotation ainsi que les tests perceptifs adoptés sont décrits. Le troisième paragraphe est consacré à la détection lexicale des émotions. Les indices prosodiques sont présentés dans le quatrième paragraphe. Les annotations d'émotions sont également corrélées avec les annotations dialogiques (section 5).

Enfin, le dernier paragraphe de cet article présente des conclusions et perspectives.

## 2. DESCRIPTION DU CORPUS ET PROTOCOLE D'ANNOTATIONS

Le corpus utilisé comporte environ en 5000 tours de parole (100 clients dont 8 femmes, 4 agents dont 1 femme) extraits d'appels à un centre de transactions boursières. Ces enregistrements ont été effectués dans le cadre du projet Amities pour une étude de développement d'un centre de routage d'appels. Le service de transactions peut être atteint *via* une connexion *Internet* ou directement en appelant un agent. Les appels couvrent une large palette de manifestations conversationnelles possibles en terme de sujet, longueur et mode de phrases, et enfin caractéristiques des locuteurs. En majorité, les appels sont dus à des problèmes de connexion au service *Internet,* cependant certains clients préfèrent une interaction avec un agent humain. Les sujets des dialogues portent sur des demandes d'informations générales (cotations, taux des commissions,…), passages d'ordres (achat, vente, statut), demandes de conseils, confirmations de transaction, problèmes de connexion *Internet*, etc. Le nombre de tours de parole par dialogue est en moyenne de 50, le nombre moyen de mots d'un tour de parole étant de l'ordre de 9 mots.

L'annotation avec des étiquettes émotionnelles est sujette à subjectivité et nécessite un protocole d'annotation rigoureux afin d'assurer la cohérence des annotations. Cette variabilité est encore plus marquée dans les corpus de données enregistrées dans des conditions réelles. Pour cette étude, les étiquettes émotions sont portées par les tours de parole.

Le protocole d'annotation [6] nécessite plusieurs étapes de traitement :

- la sélection d'une liste d'étiquettes d'émotions appropriées, une première phase d'annotation par au moins deux annotateurs avec une mesure d'accord inter-annotation,
- une validation perceptive des choix des classes et de l'annotation elle-même à partir d'un sous-ensemble de tours de parole tirés aléatoirement dans le corpus,
- enfin, au vue des résultats des tests et de l'inter-annotation, une révision des étiquettes choisies peut s'avérer nécessaire ainsi qu'une ré-annotation pour les cas ambigus.

Dans notre étude, nous avons considéré à la fois des émotions et des comportements/attitudes dépendants de la tâche. Ainsi, deux émotions négatives parmi les quatre primaires ont été retenues, la Colère et la Peur ainsi que des comportements comme la Satisfaction et l'Excuse (gêne) qui étaient fréquents dans ce corpus. L'étiquette Peur figure, pour ce corpus, un état d'inquiétude voir d'anxiété. Enfin, l'état Neutre de référence correspond à l'évolution normale du dialogue.

Deux annotateurs ont indépendamment écouté les dialogues et étiqueté chaque tour de parole. 2,7% du corpus a été annoté de façon ambiguë dans le choix des étiquettes. Le coefficient Kappa mesure la fiabilité des annotations entre annotateurs, il est de 0,8 pour ce corpus. L'ambiguïté concerne notamment l'étiquette Neutre versus une autre émotion. Ces cas ont été désambiguïsés par un troisième annotateur.

Afin de valider les annotations, deux tests perceptifs (l'un avec écoute du signal, l'autre sans) ont été menés auprès de quarante sujets (vingt par test) sur un sous-ensemble de quarante tours de parole présentés hors contexte dialogique, huit phrases étant tirées aléatoirement pour chacune des cinq émotions.

55% des tours de parole sont majoritairement perçus avec la même étiquette émotion dans les deux conditions de test montrant l'importance des indices lexicaux. Dans la condition avec écoute du signal, les tours de parole portant des émotions négatives ont été correctement perçues à 75% par les sujets validant ainsi les étiquettes initiales. L'Excuse n'a pas posé de problème de reconnaissance, tandis que la Satisfaction a été globalement perçue comme Neutre c'est-à-dire un état normal de progression du dialogue. 13,2% du corpus a été annoté avec des étiquettes non neutres. Parmi ces tours de parole les émotions négatives sont 8 fois plus représentées chez les clients que chez les agents (2,10% pour les agents vs 16,7% pour les clients). Plus précisément, parmi les tours de parole étiquetés Colère, 7,5% appartiennent aux agents et 92,5% aux clients, alors que la proportion pour Peur est de 20% vs 80%. Finalement, l'Excuse caractérise surtout les tours de parole des agents. De manière générale, les agents produisent deux fois plus de tours de parole étiquetées Satisfaction ce qui confirme son caractère non marqué et proche du Neutre.

## 3. DÉTECTION LEXICALE DES ÉMOTIONS

Un système de détection des émotions basé sur un modèle markovien unigram a été développé [3].

L'émotion portée par une phrase inconnue $u$ est déterminée par le modèle E qui obtient la meilleure probabilité a posteriori P(u/E) :

$$\log P(u/E) = \frac{1}{L_u} \sum_{w \in u} tf(w,u) \log \frac{\lambda P(w/E) + (1-\lambda) P(w)}{P(w)}$$

où P(w/E) est la probabilité d'un mot w sachant le modèle d'émotion E, P(w) est la fréquence d'un mot dans le modèle général obtenu sur l'ensemble du corpus d'entraînement, tf(w,u) représente la fréquence d'un mot dans la phrase, et $L_u$ est la longueur de la phrase en nombre de mots. Les procédures de normalisation

utilisées sont la lemmatisation et la composition de structures négatives, par exemple, « marche_pas ».

Ce système de détection des émotions fournit un taux de détection d'environ 70% pour les cinq émotions. Les résultats montrent que certaines émotions sont plus facilement détectables que d'autres, le meilleur score étant obtenu pour la Satisfaction et l'état Neutre et le moins bon pour la peur. La bonne détection de la Satisfaction peut être attribuée aux marques lexicales spécifiques comme, par exemple, « merci, d'accord ». Au contraire, l'expression de la Peur est plus syntaxique que lexicale à travers des répétitions et des re-formulations. Les performances du modèle augmentent de manière significative lorsqu'on considère deux classes principales d'émotions, Positives (Neutre/Excuse/Satisfaction) vs Négatives (Colère/Peur). Dans cette configuration les scores de détection atteignent 83% (Négatives) vs 87% (Positives) et un total de 85% de bonne détection.

## 4. INDICES PROSODIQUES CARACTÉRISANT LES ÉMOTIONS

Les paramètres prosodiques classiques tels que le débit, le contour mélodique et l'énergie ont été étudiés [4]. Dans cet article, nous allons présenter les paramètres relatifs aux variations du contour mélodique de la phrase (variation de F0).

Les mesures de F0 sont estimées sur les segments voisés à l'aide du logiciel PRAAT. Pour chaque tour de parole, le minimum, le maximum, la moyenne et la différence entre minimum et maximum (plageF0) sont calculés au niveau global de la phrase. Le maximum de variation de F0 entre deux segments consécutifs voisés (maxDF0) a également été calculé (niveau segmental). Nous avons finalement considéré les paramètres les plus distinctifs des émotions positives et négatives. Il s'agit des deux paramètres de variation de F0 à savoir la différence entre minF0 et maxF0 (i.e. plageF0) et le maxDF0.

**Paramètres F0 au niveau local (tour de parole) et global (dialogue)**

Les deux paramètres (plageF0 et maxDF0) ont été analysés selon deux points de vues : au niveau du tour de parole (indépendamment du locuteur) et au niveau du dialogue (dépendant du locuteur).

Les analyses menées dans les deux conditions montrent une forte corrélation entre, les deux paramètres retenus et, les émotions négatives (Peur, Colère), comparées à l'état Neutre (Table 1, 2).

**Table 1 :** Valeurs moyennes des paramètres F0 pour les 5 émotions sur le corpus global (5000 tours de parole). Symboles: Satif=satisfaction

| Variations F0 (niveau de la phrase) | | | | | |
|---|---|---|---|---|---|
| **Etiquettes** | **Col** | **Peu** | **Exc** | **Sat** | **Neu** |
| *Nb phrases* | 253 | 192 | 51 | 167 | 4295 |
| *PlageF0 (Hz)* | **220** | **228** | 201 | 174 | 171 |
| *MaxDF0 (Hz)* | **129** | **127** | 97 | 91 | 81 |

**Table 2 *:*** Valeurs moyennes des paramètres F0 pour les 5 émotions sur le corpus global (5000 tours de parole). Symboles : Ind = Indices

| Variation F0 (au niveau du dialogue) | |
|---|---|
| **Règles** | **%dial** |
| ***R1*** *: ind(Peur) & ind(Colére)>ind(Neutre)* | 61% |
| ***R2*** *: inds(Colère)>ind(Neutre)* | 75% |
| ***R3*** *: ind(Peur)>ind(Neutre)* | 68% |

Pour la Table 2, les 3 règles considérées sont : R1: locuteurs où les 2 paramètres pour les 2 émotions négatives présentent des valeurs supérieures au Neutre ; R2 : % locuteurs où les deux paramètres pour Colère présentent des valeurs supérieures au Neutre ; R3 : % locuteurs où les deux paramètres pour Peur présentent des valeurs supérieures au Neutre. Les interactions dans des conditions réelles présentent des manifestations émotionnelles très complexes qui font appel à des marqueurs relevant de niveaux linguistiques différents.

**Paramètres F0 & discrimination entre Peur et Colère**

Une analyse plus fine a été menée afin de distinguer les deux émotions négatives révélées par le corpus, Colère et Peur [5]. A cette fin, deux variables ont été prises en compte, le locuteur (agent/client) et le genre (homme/femme). Les deux paramètres plageF0 et maxDF0 ont été considérés dans cette perspective et à nouveau aux deux niveaux, du tour de parole et du dialogue. La corrélation des émotions négatives avec la variable locuteur montre des manifestations différentes selon le statut du locuteur dans le dialogue. Ainsi, la Peur a des manifestations plus importantes à travers la magnitude des deux paramètres F0 chez les clients que chez les agents. Parmi les clients, cette observation concerne notamment les locuteurs masculins. Plus généralement, la prise en compte de la variable genre permet de mettre en évidence des valeurs plus hautes des paramètres F0 pour Colère et Peur par rapport au

Neutre chez les locuteurs clients masculins. Les clients féminins présentent des valeurs plus modérées globalement et une plus haute variation F0 pour les tours de parole étiquetés Peur. Cependant, il n'est pas possible de généraliser ces comportements, étant donné que les classes de locuteurs hommes/femmes ne sont pas équilibrées.

**Table 3 :** Valeurs moyennes pour les paramètres prosodiques corrélés avec trois émotions et en fonction du genre (5000 tours de parole).

| Variation F0 inter-agent | | | |
|---|---|---|---|
| **Etiquettes** | **Col** | **Peu** | **Neu** |
| *Agent1(homme)-plageF0 (Hz)* | **207** | 87 | 117 |
| *Agent1(homme)-maxDF0 (Hz)* | **111** | 43 | 60 |
| *Agent2(homme)-plageF0 (Hz)* | **122** | 65 | 102 |
| *Agent2(homme)-maxDF0 (Hz)* | **76** | 22 | 50 |
| *Agent3(homme)-plageF0 (Hz)* | **141** | **166** | 121 |
| *Agent3(homme)-maxDF0 (Hz)* | **96** | **104** | 56 |
| *Agent4(homme)-plageF0 (Hz)* | 132 | **193** | 125 |
| *Agent4(homme)-maxDF0 (Hz)* | 95 | 105 | 56 |

Des différences dans l'amplitude des deux paramètres F0 sont à noter également chez les agents. Ces différences ne suivent pas systématiquement la variable genre mais révèlent plutôt des spécificités de comportement émotionnel dépendant des types de locuteur. Ainsi, l'agent masculin (agent 3) présente le plus de variation pour les deux paramètres et les deux émotions, tandis que le seul agent féminin (agent 4) manifeste plutôt une variation des paramètres lorsqu'il s'agit de la Peur, ce qui s'avère un comportement inverse par rapport aux agents 1+2 lesquels présentent plus d'amplitude des paramètres lorsqu'il s'agit de la Colère. Plus intéressant, le comportement des agents semble influencer celui des clients. Ces derniers montrent donc des variations contraires des deux paramètres F0 pour les deux émotions par rapport à leurs interlocuteurs agents (i.e. agents 1+2, 4). En revanche, lorsque les manifestations émotionnelles sont globalement hautes chez l'agent (agent 3), les clients réagissent similairement. Cette analyse nous permet d'observer que les émotions ont des manifestations complexes qui dépendent à la fois du thème dialogique et des comportements émotionnels respectifs des interlocuteurs. Ainsi, on peut noter une interdépendance des comportements des agents et des clients.

## 5. CORRÉLATION AVEC LES ACTES DE DIALOGUES

Les annotations émotionnelles ont été corrélées avec les actes dialogiques (adaptés d'après DAMSL standard dialogs acts annotation) [7]. La corrélation montre que les émotions négatives Colère et Peur sont susceptibles de générer plus fréquemment certains actes de dialogue comme les Ré-assertions, et Répétitions etc., tandis que les émotions positives comme la Satisfaction et le Neutre sont corrélées avec des actes dialogiques comme l'Acceptation.

## 6. CONCLUSIONS ET PERSPECTIVES

Les résultats montrent que les émotions ont des manifestations à plusieurs niveaux linguistiques et prosodiques. Le but de ce travail est de proposer un modèle de détection qui prendra en compte tous ces niveaux. Les différents indices trouvés sont actuellement testés sur d'autres corpus de données réelles afin de juger de leur robustesse.

## BIBLIOGRAPHIE


[1] AMITIES : http://www.dcs.shef.ac.uk/nlp/amities

[2] A. Batliner, et al., (2003), "How to find trouble in communication", Speech Communication 2003.

[3] L. Devillers,, I. Vasilescu, L. Lamel, (2003), "Emotion Detection in a task-oriented Dialog Corpus", IEEE ICME 2003, Baltimore.

[4] L. Devillers, I., Vasilescu, (2003), "Prosodic cues for emotion characterization in real-life spoken dialogs", Eurospeech,, Genève.

[5] L. Devillers, I. Vasilescu, (2004), "Anger and Fear in recorded conversations", Speech prosody, Nara.

[6] L. Devillers, I. Vasilescu, C. Mathon, (2003), "Prosodic cues for perceptual emotion detection in task-oriented Human-Human corpus", ICPhs 2003, Barcelone..

[7] L. Devillers, S. Rosset, H. Maynard, L. Lamel, (2002), "Annotations for Dynamic Diagnosis of the Dialog State", LREC 2002, Las Palmas.

[8] R. Fernandez, R. Picard, (2003), "Modeling, Drivers' Speech Under Stress", Speech Communication 2003.

[9] C. Lee, N. Narayanan, R. Pieraccini, (2001), "Recognition of Negative Emotions from the Speech Signal", ASRU 2001.

[10] K. Scherer, (2003), "Vocal communication of emotion: A review of research paradigms", Speech Communication 2003.