Reconnaissance automatique des expressions émotionnelles dans la parole naturelle : le cas des patients dans des appels d'urgence à Besançon et à Lausanne

Dongjun Wei[1], Mohamed Embarki[1], Oussama Barakat[2], Coralie Vaucherey[1,2], Thibaut Desmettres[2,3], Tania Marx[2,4] et Stephan Robert[5]
[1]ELLIADD EA 4661, université de Franche-Comté, Besançon [2]Laboratoire de Nanomédecine, Imagerie et Thérapeutique EA 4662, université de Franche-Comté, Besançon [3]Hopitaux universitaires Genève, université de Genève[4] Centre Hospitalier Régional Universitaire de Besançon [5]Haute Ecole d'Ingénierie et de Gestion du Canton de Vaud, Yverdon-les-Bains

La **reconnaissance automatique de la parole** (ASR) a fait des progrès significatifs à la fois en termes d'approches et de modèles, la reconnaissance du locuteur a aussi franchi des étapes importantes. La parole véhicule aussi bien des informations linguistiques que des informations non linguistiques, comme les **émotions**. Hormis le lexique qui peut être rattaché aisément à un type d'émotion, la **prosodie** a été décrite comme source importante dans l'expression des émotions. La **reconnaissance automatique des émotions** a été développée pour diverses applications. Un besoin croissant de systèmes de reconnaissance des émotions se fait sentir dans le domaine de la **régulation médicale d'urgence** afin de pouvoir identifier rapidement la gravité de la situation, orienter vers la bonne filière de soins et/ou engager les moyens de transport adéquats.

Matériel et méthodes

32 appels téléphoniques de patients aux services des urgences du **CHRU de Besançon et du CHUV de Lausanne**. Seuls les extraits correspondant aux **4 émotions** à polarité négative ont été retenus. Les extraits ont été d'abord segmentés et étiquetés; ensuite leurs **propriétés acoustiques sont mesurées manuellement** sous Praat; finalement, **406 énoncés** retenues correspondent aux **émotions vocales non linguistiques** (Tableau 1).

	Angoisse		Colère		Embarras		Tristesse		
	H.	F.	H.	F.	H.	F.	H.	F.	
France + Suisse	342 (197	145)	5 (2	3)	43 (38	5)	16 (13	3)	406
France	214 (127	87)	5 (2	3)	41 (36	5)	14 (12	2)	274
Suisse	128 (70	58)	0 (0	0)	2 (2	0)	2 (1	1)	132

Table 1: Nombre d'émotions par type d'émotion, genre et pays.

L'**étiquetage** des émotions prosodiques est largement basé sur les **sentiments subjectifs** des étiqueteurs, mais pour mieux distinguer les émotions des patients, une **explication précise** leur a été fournie préalablement. Les **mesures de F0** ont ciblé les **4 points d'inflexion** de chaque énoncé (**début, fin, maximum et minimum**) et la **F0 moyenne**, il en est de même des **mesures d'intensité**. L'**écart de F0** est converti en **demi-ton**. Les mesures de durée ont ciblé la vitesse d'articulation et la vitesse de parole.

Résultats

Une comparaison est proposée entre : même genre du même pays; même genre de différents pays / genre différent du même pays.

		Fr. F0Moy.	Su. F0Moy.	EDT 5
An.	F.	231±32	232±30	0
	H.	151±20	142±24	1
	EDT 1	7	8	
Co.	F.	226±22		
	H.	143±10		
	EDT 2	8		
Em.	F.	224±53		
	H.	133±19	108±19	4
	EDT 3	10		
Tr.	F.	232±40	331	6
	H.	131±16	101	5
	EDT 4	10	20	

Tableau 2: F0 Moyenne (Hz), écart-type (±) et écart en demi-tons=EDT des extraits chez des patients français et suisses dans 4 émotions.

		Fr. IM	Su. IM	EI5
An.	F.	74±3	75±7	1
	H.	71±5	72±5	1
	EI1	3	3	
Co.	F.	74±3		
	H.	73±4		
	EI2	1		
Em.	F.	71±4		
	H.	70±6	73±2	3
	EI3	1		
Tr.	F.	74±3	80	6
	H.	74±2	69	5
	EI4	0	11	

Tableau 3 : Intensité moyenne (dB)=IM., écart-type (±) et écarts d'intensité (dB)=EI pour les patients français et suisses dans les 4 émotions.

		Fr.		ED5	Su.		ED6
		VPM	VAM		VPM	VAM	
An.	F.	5±1	5±1	0	5±1	5±2	0
	H.	5±2	5±1	0	5±2	6±2	1
	ED1	0	0		0	1	
Co.	F.	5	6	1			
	H.	3	4	1			
	ED2	2	2				
Em.	F.	5±1	6±1	1			
	H.	5±1	5±1	0	4±2	4±2	0
	ED3	0	1				
Tr.	F.	5±2	5±1	0	4±2	4	0
	H.	4±2	5±	1	4±2	5	0
	ED4	1	0		0	1	

Tableau 4 : Vitesse d'articulation moyenne (syl/s) = VAM et la vitesse de parole moyenne (syl/s) = VPM, écart-type (±) et écarts de durée (syl/s) =ED chez des patients français et suisses dans 4 émotions.

		An.		EDT5	Co.		EDT6	Em.		EDT7	Tr.		EDT8
		F.	H.		F.	H.		F.	H.		F.	H.	
Déb	Fr.	224	142	8	228	142	8	220	133	9	263	138	11
	Su.	234	138	9					103		305	106	18
	EDT1	1	1						4		3	6	
Fin	Fr.	226	159	6	181	120	7	205	138	7	217	111	12
	Su.	224	141	8					118		447	89	28
	EDT2	0	2						3		13	4	
EDT 9	Fr.	0	2		4	3		1	1		3	4	
	Su.	8	0						2		7	3	
Max	Fr.	360	237	7	345	220	8	328	217	7	291	194	7
	Su.	323	209	8					142		463	115	24
	EDT3	2	2						7		8	9	
Min	Fr.	137	108	4	155	73	13	142	93	7	132	94	6
	Su.	162	99	9					88		254	80	20
	EDT4	2	1						1		11	3	
EDT 10	Fr.	17	14		14	19		15	15		14	13	
	Su.	12	13						8		10	6	

Tableau 5 : F0 Moyenne (Hz) et écart en demi-tons=DT à quatre points d'inflexion chez des patients français et suisses dans 4 émotions

		An.		EI5	Co.		EI6	Em.		EI7	Tr.		EI8
		F.	H.		F.	H.		F.	H.		F.	H.	
Déb	Fr.	67	67	0	65	66	1	67	68	1	70	69	1
	Su.	71	67	3					71		72	75	3
	EI1	4	0						3		2	6	
Fin	Fr.	62	62	0	64	63	1	55	61	6	75	64	11
	Su.	64	61	3					62		62	56	6
	EI2	2	1						1		13	8	
EI9	Fr.	5	5						7		5	5	
	Su.	7	6						8		10	9	
Max	Fr.	82	82	0	87	81	6	80	80	0	82	83	1
	Su.	84	81	3					82		88	78	11
	EI3	2	1						2		6	5	
Min	Fr.	28	28	0	25	32	7	26	27	1	72	29	43
	Su.	35	31	4					41		46	26	20
	EI4	7	7						13		26	3	
EI 10	Fr.	54	54						53		10	54	
	Su.	49	50						41		42	52	

Tableau 6 : l'intensité moyenne (dB) et l'écart d'intensité (dB)=EI à quatre points d'inflexion chez des patients français et suisses dans 4 émotions.

Conclusion

La **reconnaissance des émotions verbales naturelles** est plus complexe que la reconnaissance des émotions verbales non naturelles.

- Parmi F0 , l'intensité et la durée, les **modulations de F0** sont les plus représentatives du **changement d'émotion vocale linguistique**, et plus particulièrement les **changements de contour et de F0 moyenne**.
- L'**intensité** et la **durée** ne distinguent que **très faiblement** les émotions vocales linguistiques.
- Par conséquent, la conception du système de reconnaissance automatique des émotions vocales doit absolument tenir compte des F0 ainsi que de leurs modulations, l'extraction des données de F0 reste la propriété acoustique de base pour la reconnaissance des émotions vocales linguistiques.

Références

- COWIE, R., DOUGLAS-COWIE, E., TSAPATSOULIS, N., VOTSIS, G., KOLLIAS, S., FELLENZ, W., TAYLOR, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32-80.
- BÄNZIGER, T., GRANDJEAN, D., BERNARD, P. J., KLASMEYER, G., SCHERER, K. R., (2001). Prosodie de l'émotion : étude de l'encodage et du décodage. Cahiers de linguistique française 23, p. 11-37.
- SCHERER, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256.
- LEE, C., LUI, S. (2014). Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition. The Journal of the Acoustical Society of America 135, 2422.
- ALI, S.A., KHAN, A., BASHIR, N. (2015). Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus. International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.2, pp.54-59