Uma abordagem de aprendizagem semi-supervisionada para a percepção automática de personalidade, baseada em pistas acústico-prosódicas em domínios com poucos recursos

Rubén Solera-Ureña[1], Helena Moniz[1,2,5], Fernando Batista[1,3], Vera Cabarrão[1,2], Anna Pompili[1,4], Ramón Fernández-Astudillo[1], Isabel Trancoso[1,4]

1 Laboratório de Sistemas de Língua Falada, INESC-ID Lisboa, Lisboa, Portugal
2 FLUL/CLUL, Universidade de Lisboa, Lisboa, Portugal
3 Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
4 Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
5 Unbabel Lda, Portugal

Resumo

A análise automática da personalidade tem vindo a ganhar grande atenção nos últimos anos, por ser esta uma dimensão fundamental nas interações homem-máquina. No entanto, o desenvolvimento desta tecnologia nalguns domínios específicos, como a avaliação da personalidade em crianças, tem sido afetado pela escassez de dados ou pela reduzida dimensão das bases de dados de fala disponíveis para o treino de modelos robustos de personalidade.

Com o objetivo de resolver alguns dos problemas neste tipo de cenários, investiga-se aqui a aplicação duma abordagem de treino semi-supervisionado que faz uso de dados heterogéneos. Apresenta-se assim uma configuração experimental cuja principal caraterística é a disparidade das bases de dados aqui usadas em relação à idade e a língua dos falantes. Designadamente, empregam-se dois conjuntos de treino distintos. Em primeiro lugar, um pequeno conjunto de treino, usado no Interspeech 2012 Personality Sub-challenge [1], que contêm 1 hora e 40 minutos de gravações de **adultos franceses** etiquetadas para os cinco traços de personalidade do **modelo *Big-Five*** (OCEAN: ***O****penness*, ***C****onscientiousness*, ***E****xtroversion*, ***A****greeableness* e ***N****euroticism*), designado *SPC corpus*. Em segundo lugar, um conjunto de treino não etiquetado (desenhado para tarefas de reconhecimento automático de fala) com 20 horas de gravações de **crianças portuguesas**, *CNG corpus* [2]. Como conjunto de teste, usa-se uma base de dados de fala de crianças portuguesas etiquetada em termos de personalidade, *GoN corpus* [3]. Com base nesta configuração, investigamos uma abordagem de supervisão fraca na qual um modelo SVM (*support vector machine*) treinado inicialmente com o conjunto de dados etiquetados *SPC corpus* é refinado (re-treinado) de forma iterativa, adicionando dados não etiquetados do *CNG corpus*. Também investigamos representações baseadas em conhecimento linguístico prévio (do inglês "knowledge-based") sobre pistas acústicas e prosódicas para a tarefa de percepção de personalidade (*KB-features*), comparadas posteriormente a dois tipos de caraterísticas usadas habitualmente em paralinguística computacional (*openSMILE* [4] e *eGeMAPS* [5]), de modo a estabelecer comparações com resultados obtidos na literatura.

A Tabela 1 mostra os resultados sobre a base de dados *GoN corpus* (segmentos completos) do modelo inicial, treinado apenas com dados etiquetados. Os resultados são apresentados com as métricas *unweighted average recall* (UAR) e *accuracy* (Acc). As taxas de desempenho são razoáveis para a percepção de *Openness*, *Extroversion* e *Agreeableness*, os traços de personalidade mais evidentes para as crianças no contexto em que foram feitas as gravações do *GoN corpus.* Também pode comprovar-se a viabilidade das caraterísticas *KB-feature*s, sendo o seu número (41) muito menor do que as *openSMILE* (6125).

A Figura 1 mostra os resultados para *Openness* obtidos sobre o conjunto *GoN corpus* pelos modelos semi-supervisionados re-treinados de forma iterativa com dados não etiquetados (*CNG corpus*). Apresentam-se resultados sobre três versões distintas do *GoN corpus*, as quais constam de segmentos de fala completos, segmentos de 20 segundos e segmentos de 10 segundos de

duração, respetivamente. A figura ilustra como a incorporação progressiva de dados não etiquetados nas sucessivas iterações produz uma melhoria do desempenho do sistema.

Como conclusão, estes resultados apontam os potenciais da aplicação de abordagens de aprendizagem semi-supervisionada sobre bases de dados heterogéneas (i. e., língua e faixa etária dos falantes) para superar a falta de dados etiquetados em domínios com poucos recursos, como a tarefa de percepção da personalidade em crianças. Comprova-se a existência de pistas acústicas e prosódicas partilhadas por falantes com distintas línguas e idades, que são captadas pelas caraterísticas propostas, o que permite a percepção de traços de personalidade em crianças portuguesas, usando modelos treinados com gravações de adultos em francês.

Palavras chave: paralinguística computacional, percepção automática de personalidade, OCEAN, distintas línguas, faixas etárias diferentes, pistas acústico-prosódicas.

Agradecimentos: este trabalho foi financiado pela Fundação para a Ciência e a Tecnologia (FCT), referência UID/CEC/50021/2013, pelos projetos INSIDE (referência CMUP-ERI/HCI/0051/2013) e BioVisualSpeech (referência CMUP-ERI/TIC/0033/2014) no âmbito do protocolo Universidade Carnegie Mellon/Portugal, bem como pelo financiamento das bolsas de doutoramento SFRH/BD/96492/2013 e SFRH/BD/97187/2013 e de pós-doutoramento SFRH/PBD/95849/2013.

Bibliografia

[1] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, "The INTERSPEECH 2012 Speaker Trait Challenge," in Proc. of Interspeech 2012, Portland, OR, USA, Sep. 2012.

[2] A. Hämäläinen, F.M. Pinto, S. Rodrigues, A. Júdice, S.M. Silva, A. Calado, M.S. Dias, "A Multimodal Educational Game for 3-10-Year-Old Children: Collecting and Automatically Recognising European Portuguese Children's Speech," in Workshop on Speech and Language Technology in Education, Grenoble, France, Aug. 2013.

[3] J. Campos, P. Oliveira, A. Paiva, "Looking for Conflict: Gaze Dynamics in a Dyadic Mixed-Motive Game," Autonomous Agents and Multi-Agent Systems, vol. 30, no. 1, pp. 112–135, 2016.

[4] F. Eyben, F. Weninger, F. Gross, B. Schuller, "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor," in Proc. of the 21st ACM International Conference on Multimedia, New York, NY, USA, Oct. 2013, pp. 835–838.

[5] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, K.P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Trans. on Affective Computing, vol. 7, no. 2, pp. 190–202, 2016.

TRAIT	NORM.	openSMILE			eGeMAPS			eGeMAPS+KB-feats.		
		C	TEST		C	TEST		C	TEST	
			UAR	Acc		UAR	Acc		UAR	Acc
O	NORM.	3E-1	61.7	63.6	1E-1	65.8	68.2	1E-5	55.0	59.1
	STAND.	1E-7	70.0	72.7	3E-4	60.8	63.6	1E-4	70.8	72.7
C	NORM.	3E-5	38.3	36.4	3E-1	43.3	40.9	1E-5	50.0	45.5
	STAND.	3E-4	42.5	40.9	3E-2	37.5	36.4	3E-5	33.3	31.8
E	NORM.	3E-3	67.9	59.1	1E-1	71.4	63.6	1E-4	71.4	63.6
	STAND.	1E-4	71.4	63.6	1E-2	65.2	59.1	3E-3	68.8	63.6
A	NORM.	3E-4	64.3	68.2	1E-5	59.8	59.1	3E-4	59.8	59.1
	STAND.	3E-8	64.3	68.2	1E-5	59.8	59.1	3E-4	57.1	59.1
N	NORM.	3E-5	27.5	27.3	3E-6	27.5	27.3	1E0	26.7	27.3
	STAND.	3E-4	31.7	31.8	1E-7	27.5	27.3	1E-7	27.5	27.3

Tabela 1: Resultados do modelo supervisionado inicial sobre a base de dados *GoN corpus* –segmentos completos.

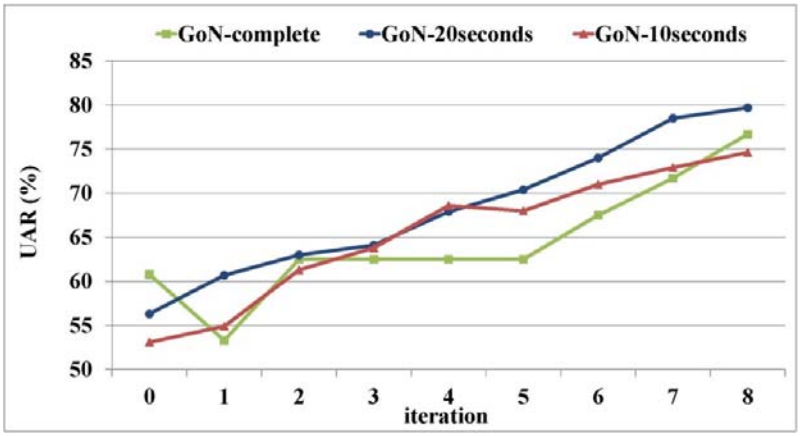

Figura 1: Resultados para Openness dos modelos semi-supervisionados sobre três versões da base de dados *GoN corpus* –segmentos completos, segmentos de 20 segundos e segmentos de 10 segundos.