Desenvolvimento de um modelo computacional para reconhecimento e síntese de expressões emocionais auditivas utilizando *deep learning*

Ingryd Vanessa de Sá Teles Pereira[1], Mestranda em Engenharia da Computação - UPE (ivstp@ecomp.poli.br)
Alexandre Magno Andrade Maciel[2], Professor Doutor - UPE (amam@ecomp.poli.br)

Nos últimos 50 anos, o desenvolvimento de robôs, na indústria e na esfera doméstica (pessoal e de serviço), tem sido significativo e tem expectativa de um crescimento exponencial nos próximos anos (MINAMI, 2013). Esta situação levantou questões sobre como os robôs interagem com humanos e o quanto uma interação amigável seria benéfica. A interação entre seres humanos é modulada por contextos emocionais. Portanto, os elementos-chave da comunicação natural são as habilidades de perceber e expressar emoções (NUMMENMAA et al., 2014). Dado isto, *Voice User Interfaces* (VUI) ganharam cada vez mais relevância na interação entre humanos e interfaces robóticas. Para Charfuelan e Steiner (2013), a maneira mais eficiente de realizar uma comunicação entre humanos e robôs é usar síntese de fala com diferentes estilos emocionais.

Essa pesquisa visa construir um modelo computacional para reconhecimento e síntese de expressões emocionais auditivas para aplicação em ambientes robóticos. Representar emoções é um desafio, pois são muito subjetivas e não existe uma *ground truth*. Uma maneira de representar emoções é a partir de rótulos categóricos, no qual são definidos *labels* para as emoções. Outra maneira para representar emoções é identificar e atribuir valores a componentes nas quais as emoções são feitas. Por exemplo, o trabalho de Russel (1997) que propõe que a experiência emocional é descrita por 2 dimensões: *Valence* (mede o quão positiva ou negativa uma experiência foi) e *Arousal* (mede o quão ativa a experiência foi).

Essa pesquisa utilizou 2 bases de dados, ambas as bases são audiovisual emocionais, multimodais e *multispeaker's*. A base de dados SAVEE é anotada com rótulos categóricos (raiva, desgosto, medo, felicidade, neutro, tristeza e surpresa), e a base de dados IEMOCAP possui rótulos categóricos (raiva, felicidade, tristeza, neutralidade) e rótulos dimensionais (*Arousal* e *Valence*). Para o reconhecimento de emoções foi utilizado CNN (*Convolutional Neural Network*) e para a geração de voz emotiva será utilizado BEGAN (*Boundary Equilibrium Generative Adversarial Networks*).

Experimentos foram realizados com a base SAVEE, onde várias formas de representação acústica presentes na literatura foram selecionadas e também o áudio puro (sinal), e foi realizado o reconhecimento de emoções utilizando uma CNN. Os melhores resultados foram obtidos utilizando apenas o áudio puro, com uma média de 81,08% de acurácia. A principal conclusão tomada até o momento é de que a utilização de *deep learning* com dados brutos conseguem extrair características do áudio melhor do que as transformações acústicas. Após o desenvolvimento e validação dos algoritmos, o algoritmo será integrado a um ambiente robótico, e espera-se que ao fim, tenha-se um robô capaz de se comunicar de forma emotiva.

Palavras-chave: *Reconhecimento de emoção; Síntese de voz emotiva; deep learning*

Referências

CHARFUELAN, M.; STEINER, I. **Expressive speech synthesis in marytts using audiobook data andemotionml**. In: INTERSPEECH. [S.l.: s.n.], 2013. p. 1564–1568.

MINAMI, Y. **Executive summary**: world robotics 2012 industrial robots. Available online on http://www.worldrobotics.org, p. 8–18, 2013.

NUMMENMAA, L. et al. **Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks**. NeuroImage, Elsevier, v. 102, p. 498–509, 2014.

RUSSELL, J. A. **13-reading emotion from and into faces**: Resurrecting a dimensional-contextual perspective. The psychology of facial expression, New York: Cambridge University, p. 295–320, 1997.