MEMÓRIAS
DA
ACADEMIA DAS CIÊNCIAS
DE
LISBOA

CLASSE DE CIÊNCIAS

TOMO XLVII
Volume 2

# 4 Décadas a Falar com Computadores

ISABEL TRANCOSO

ACADEMIA DAS CIÊNCIAS
DE LISBOA

LISBOA • 2020

# 4 Décadas *a Falar* com Computadores

Isabel Trancoso[1]

Esta apresentação pretende ser não só uma retrospectiva sobre a investigação em tecnologias da fala nos últimos quarenta anos, e a panóplia de aplicações que o progresso recente nestas tecnologias potencia, mas também uma reflexão sobre o futuro desta área.

A retrospectiva cobre apenas duas tecnologias cruciais – a conversão de fala para texto e a conversão de texto para fala – distinguindo em qualquer delas quatro gerações. A primeira geração de sistemas de conversão de fala para texto (mais conhecida como reconhecimento automático de fala, ou ASR – *Automatic Speech Recognition*), cobre a década de 60 (e anteriores), em que as abordagens eram sobretudo heurísticas. Na segunda, ao longo da década de 70, predominaram as abordagens baseadas em reconhecimento de padrões. A terceira geração estendeu-se pelas três décadas seguintes, com um sem número de abordagens estatísticas em que predominaram os modelos de Markov não observáveis. A quarta geração surgiu cerca de 2010 e é hoje em dia completamente dominada pelas abordagens de aprendizagem automática baseadas em *deep learning*.

A primeira geração dos sistemas de síntese de fala a partir de texto (ou TTS – *Text-to-Speech Synthesis*) pode considerar-se como abrangendo as décadas de 60 a 80, com o predomínio dos sistemas de síntese por regra. A segunda geração cobre as décadas de 80 e 90, e é dominada pelas abordagens de síntese por concatenação que, do ponto de vista comercial, são ainda muito importantes hoje em dia. A terceira geração surgiu na primeira década deste século numa tentativa de conciliar modelos estatísticos e paramétricos. Tal como nos sistemas de reconhecimento de fala, também em síntese as abordagens baseadas em deep learning marcam a quarta geração, cujo início podemos novamente situar na presente década. Apesar da enorme complexidade destes sistemas, sobretudo os do tipo end-to-end, é importante não os encarar como "caixas negras", mas antes procurar captar o que eles estão internamente a aprender. Ao longo das quatro gerações temos podido observar interessantes mudanças de paradigma que conduziram a reduções cada vez maiores da taxa de erro dos reconhecedores de fala que em certas con- dições se aproximam do comportamento humano. Também na síntese da fala, as tecnologias da quarta geração têm vindo a permitir progressos de tal ordem que se corre o risco de deixar de ser possível distinguir entre a fala real e sintética de uma pessoa, o que pode colocar questões de autenticidade graves, por exemplo, na esfera judiciária.

Todo este progresso permitiu nesta última década uma verdadeira explosão de aplicações destas tecnologias da fala, com um grande destaque para os assistentes virtuais (incluindo os robóticos), embora ainda existam muitos problemas de usabilidade. As aplicações destas tecnologias a outras áreas como a da saúde e a do ensino da língua são bem menos conhecidas, mas com um enorme impacto potencial.

[1] INESC-ID / Instituto Superior Técnico, Universidade de Lisboa

A reflexão sobre o futuro desta área deixa no ar uma pergunta: "Chegámos lá?", em que se questiona quão perto ou longe estamos do desempenho de um humano, e o tipo de abordagens que melhor tentam lá chegar, salientando no entanto as diferenças entre os progressos alcançados em tarefas do tipo perceptivo e cognitivo, e alertando para os perigos da má utilização destas tecnologias.

(Resumo da comunicação apresentada à Classe de Ciências
na sessão de 11 de janeiro de 2018)