MEMÓRIAS
DA
ACADEMIA DAS CIÊNCIAS
DE
LISBOA

CLASSE DE CIÊNCIAS

A Fala: Como traçar e proteger o perfil de um falante

Isabel Trancoso

ACADEMIA DAS CIÊNCIAS
DE LISBOA

LISBOA • 2026

Título: A Fala: Como traçar e proteger o perfil de um falante
Edição: Academia das Ciências de Lisboa
Data de edição: 2026
DOI: https://doi.org/10.58164/bn0z-6x18

A Fala: Como traçar e proteger o perfil de um falante

Isabel Trancoso[1]

RESUMO

A fala codifica informação sobre atributos físicos do falante, como o sexo biológico, a faixa etária e até a altura, mas também sobre muitas outras características, como a emoção, o *stress* ou o sotaque, só para mencionar algumas. Além disso, também codifica pistas sobre uma infinidade de doenças, que vão para além das chamadas perturbações da fala e da linguagem, e incluem, por exemplo, doenças neurodegenerativas, mentais ou respiratórias. Hoje em dia, essa informação é extraída e codificada principalmente através de representações neuronais ou "embeddings" do falante, que podem ser subsequentemente aplicadas a muitas tarefas diferentes. Toda esta informação codificada, no entanto, pode não só ser extraída, mas também modificada ou ofuscada para reconstruir a fala. Esta comunicação aborda tanto a extração como a modificação de atributos do falante, apresentando a fala como veículo de "Informações de Identificação Pessoal" (PII) e enfatizando o seu potencial como biomarcador de saúde e as suas vulnerabilidades em termos de privacidade e segurança.

ABSTRACT

Speech encodes information about physical speaker attributes such as biological sex, age range, and even height, but also about many other characteristics such as emotion, stress or accent, just to name a few. Moreover, it also encodes cues about a plethora of diseases, which go beyond the so-called speech and language disorders, and include, for instance, neurodegenerative, psychiatric and respiratory diseases. Nowadays, this information is mostly extracted and encoded in embeddings or neural representations which may be subsequently applied to many different tasks. All this encoded information, however, can not only be extracted, but also modified or obfuscated when reconstructing speech. This talk covers both the extraction and modification of speaker attributes, presenting speech as Personal Identifiable Information, and emphasizing its potential as a health biomarker and its vulnerabilities in terms of privacy and security.

1. INTRODUÇÃO

A escolha do tópico para esta comunicação revelou-se muito difícil. Daí a escolha de não um, mas sim dois tópicos que me têm apaixonado ao longo de várias décadas. São de facto muito relacionados entre si: o primeiro tem a ver com a quantidade de informação que é passada quando alguém fala e que vai muito além do texto subjacente e da identidade do falante ou orador, como queiramos designar. E o segundo tem a ver com a forma como se consegue esconder ou proteger essa informação.

[1] Academia das Ciências de Lisboa, INESC-ID e Instituto Superior Técnico, Universidade de Lisboa.

Começaremos por rever que tipo de informação é possível extrair do sinal de fala, para além do texto (Figura 1). Podemos extrair atributos físicos como o género, a faixa etária e até traços físicos como a altura. Podemos obviamente identificar a língua, o sotaque, o estado emocional, traços de personalidade (p.e., introvertido, extrovertido, etc.), grau de educação, níveis de *stress*, intoxicação, sonolência ou carga cognitiva e entre muitos outros atributos, podemos também extrair pistas sobre doenças que afetam a fala. Num contexto multi-orador, podem também ser reveladas relações de hierarquia ou familiaridade, por exemplo.

Figura 1. Atributos de um falante.

Alguns atributos são de longo prazo como o género, ou melhor dizendo o sexo biológico, outros como as emoções são obviamente de curto prazo. Numa outra perspetiva, é também possível distinguir entre atributos controláveis pelo falante, outros apenas parcialmente e outros não controláveis de todo.

A extração de atributos tem sido uma das áreas em que o trabalho do grupo de investigação em que me integro mais se destacou, tendo ficado em primeiro lugar em competições internacionais de deteção de género e faixa etária e de deteção da língua nativa em sotaques estrangeiros. Mas é na área da deteção de doenças que afetam a fala que mais temos investido.

Este preâmbulo justifica a estruturação da comunicação em duas partes principais. A primeira tem como foco a extração de atributos que fazem da fala um potencial biomarcador de saúde. A segunda faz um apanhado ainda mais breve sobre privacidade no processamento da fala. Dado a ênfase no trabalho

desenvolvido pelo grupo de investigação, esta comunicação não pretende de todo fazer um sumário do estado da arte nestes tópicos, sendo necessariamente bastante polarizada, como aliás é evidenciado pelo grande número de referências do trabalho do grupo. Para uma leitura mais aprofundada sobre a primeira e segunda partes, aconselhamos, por exemplo, Singh (2019) e Nautch *et al.* (2019), respetivamente.

2. A FALA COMO BIOMARCADOR DE SAÚDE

O leque de doenças que afetam a fala vai muito além das perturbações articulatórias ou motoras, como, por exemplo, a Gaguez ou o Sigmatismo. Abrange doenças que afetam o aparelho respiratório (p.e., Apneia Obstrutiva do Sono, Asma, COVID-19), perturbações psiquiátricas (p.e., Depressão, Ansiedade, Doença Bipolar, espectro do Autismo) e também doenças neurodegenerativas (p.e., Doença de Parkinson, Alzheimer, Huntington, Esclerose Lateral Amiotrófica). Para compreender como estas doenças afetam a fala, tornando-a um biomarcador de saúde, basta rever o mecanismo de produção da fala. É o que faremos muito brevemente na primeira secção deste capítulo (2.1). Nas secções seguintes (2.2 e 2.3), abordar-se-ão respetivamente os vários tipos de tarefas de fala mais habituais em aplicações clínicas e as características com significado clínico que podemos extrair do sinal de fala. O foco na explicabilidade dos modelos de classificação de doenças a partir da fala é o tema da secção seguinte (2.4). Mas a prioridade na explicabilidade não nos deverá fazer descartar modelos de IA pré-treinados, muito complexos e poderosos. A última secção deste capítulo (2.5) pretende exemplificar como é possível capitalizar nesses modelos, não para a deteção direta de doenças na fala, mas sim para dela extrair características interpretáveis.

2.1 Mecanismo de produção de fala

A produção de fala é um mecanismo muito complexo que envolve processos cognitivos e ações musculares, que regulam a respiração, a fonação e a articulação. O sinal de fala vai assim intrinsecamente conter pistas sobre qualquer problema que afete tanto os processos cognitivos como as tais ações musculares.

Ao nível dos primeiros, são sobretudo as características de cariz linguístico que poderão revelar pistas sobre doenças neurodegenerativas como a doença de Alzheimer. Por exemplo, uma menor diversidade lexical ou uma menor coerência são características típicas de um orador com esta doença. Já o maior uso de pronomes na 1.ª pessoa poderá, por sua vez, ser um dos vários indicadores de depressão.

A fala é também uma fonte de informação muito relevante para evidenciar problemas a nível do aparelho respiratório. Esse facto motivou até uma competição internacional cujo objetivo era detetar a frequência respiratória a partir da fala, tendo os resultados evidenciado a relação entre a informação que pode ser extraída do sinal de fala, e a captada por cintos respiratórios.

Os problemas de fonação ocorrem a nível das nossas cordas vocais que, numa pessoa saudável, abrem e fecham de uma forma quase periódica ao produzirmos sons como vogais. Desvios relevantes nessa quase periodicidade quer em termos da duração do ciclo (designados por *jitter*), quer em termos de amplitude (designados por *shimmer*) podem resultar de um pior controlo dos músculos que acionam as cordas e indiciar assim várias doenças (p.e., Parkinson).

Já os problemas de articulação, como o nome indica, ocorrem a nível do controlo dos articuladores (p.e., língua, maxilar), que movimentamos de forma a produzir diferentes tipos de som, quer vogais, quer consoantes. Problemas no controlo dos articuladores do trato vocal podem, por exemplo, afetar as ressonâncias deste tubo acústico, designadas por formantes. Problemas no controlo do véu palatino que abre e fecha a comunicação com o trato nasal, poderão por sua vez provocar diferentes graus de nasalação.

As pistas sobre saúde são muitas, como indicado no diagrama da Figura 2, extraído de Botelho *et al.* (2024). O diagrama inclui as doenças que afetam a fala e que estão mais estudadas e a forma como a fala é afetada em cada uma. O leque é muito vasto, desde a esquizofrenia à apneia obstrutiva do sono, passando pela depressão, demência, doença de Parkinson, etc… Como se pode constatar, algumas das pistas, como um menor débito de fala ou um maior jitter, são comuns a várias doenças, o que justifica uma abordagem holística ao tópico da fala como biomarcador de saúde.

Esta abordagem holística só é possível quando existirem corpora de fala paralelos para as várias doenças. A Figura, contudo, deixa de fora muitas outras

patologias cujo impacto na fala é conhecido, embora não tão divulgado, como por exemplo a diabetes (Elbéji *et al.*, 2024). O próprio ciclo menstrual das mulheres causa mudanças em características como sejam a frequência fundamental (Fischer *et al.*, 2011). Estes exemplos ilustram a importância da recolha de corpora de fala longitudinais, infelizmente muito pouco frequentes.

Apesar da ênfase em aplicações em saúde, há que frisar que muitas das características listadas poderão ser também importantes pistas para detetar por exemplo o estado emocional ou níveis de stress ou intoxicação, ou qualquer outro atributo mencionado.

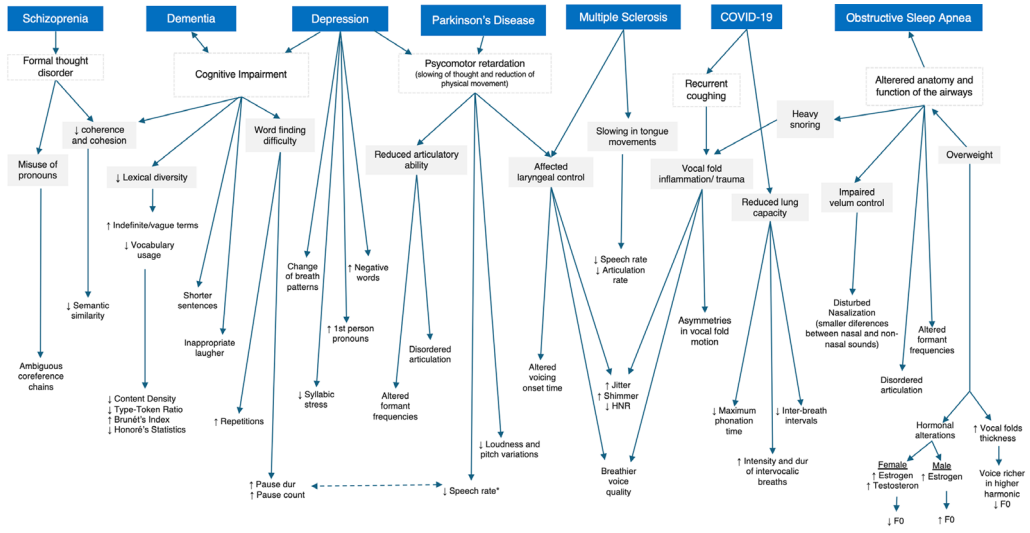

Figura 2. Doenças que afetam a fala. Extraída de Botelho *et al.* (2024).

2.2 Tipos de tarefa

A tarefa mais simples é pedir aos participantes que leiam parágrafos ou frases, ou que digam sequências de palavras (p.e., números até 10, dias da semana) ou mesmo que repitam palavras, frases ou sequências.

Se se pretender detetar eventuais problemas de fonação, a tarefa ideal será uma vogal sustentada (tipicamente /a/ ou /i/). Já para detetar problemas de articulação em adultos, é frequente recorrer-se a testes diadococinéticos (i.e, sequências de sílabas do tipo /pataka/). Para qualquer das tarefas acima citadas é sabido o que deveria ter sido dito, o que facilita a sua verificação automática.

O mesmo não acontece com tarefas que envolvam fala espontânea, tão importantes para detetar problemas a nível cognitivo. O exemplo de tarefa deste tipo mais apropriado num contexto de saúde seria talvez uma entrevista de carácter biográfico como num consultório. As entrevistas deste tipo, porém, levantam tantos problemas de reidentificação que estas bases de dados raramente estão publicamente disponíveis. Em alternativa, de forma a recolher fala espontânea, opta-se muitas vezes por pedir aos oradores que descrevam uma imagem ou recontem uma história.

Um segundo tipo de tarefa espontânea consiste em testes de fluência verbal, quer semânticos (p.e., dizer animais durante um minuto), quer fonéticos (p.e., dizer palavras começadas por "p"). Para obter automaticamente a transcrição de fala espontânea, há que usar sistemas de reconhecimento de fala, mas ao contrário do que se passa com fala lida, em que a taxa de erro é hoje em dia já muito baixa, para estas tarefas, sobretudo em testes de fluência, a taxa de erro pode exceder os 20%.

2.3 Classes de características

Os modelos de aprendizagem automática podem receber como entrada diretamente sinais de fala ou espectrogramas. Uma alternativa muito comum, porém, é dar-lhes como entrada características previamente extraídas. A tipologia que propomos para as características mais frequentemente utilizadas em aplicações clínicas distingue quatro classes.

Na classe de características que designamos de acústico-prosódicas, surgem como exemplos tanto as associadas ao trato vocal (p.e., frequências de formantes, declive espectral), como à qualidade vocal (p.e., frequência fundamental, *jitter*, *shimmer*, *harmonic-to-noise ratio* (HNR)).

Na classe de características relacionadas com o ritmo, enquadram-se as que captam informação temporal, quer a nível de palavras, quer a nível de (pseudo-) sílabas, incluindo também pausas e pausas preenchidas.

Há ferramentas que nos permitem extrair cerca de 6000 parâmetros destas duas classes para cada locução (p.e., openSMILE[2]). Um subconjunto muito utilizado em deteção de atributos a partir da fala é o EGEMAPS (Eyben *et al.*, 2016),

[2] https://www.audeering.com/research/opensmile/

também extraível com base nesta ferramenta. Para além de funcionais destas características, este subconjunto de 88 parâmetros inclui também coeficientes cepstrais na escala de Mel.

A classe seguinte congrega características de conteúdo linguístico como sejam: diversidade/riqueza do vocabulário (p.e., estatística de Honoré, índice de Brunet, *Type-Token-Ratio*), densidade de conteúdo ou de ideias (rácios de categorias gramaticais), marcadores discursivos, polaridade, repetições, pronomes na 1.ª pessoa, cadeias de coreferência (ambíguas ou não), frequência e idade de aquisição das palavras, etc.

Para testes de fluência verbal, é também importante incluir nesta classe as contagens de palavras corretas, repetições e intrusões (i.e., para um teste semântico de nomeação de animais, tudo o que não for animal). Para este tipo de tarefas, há ainda que incluir uma quarta classe, muito específica, que tem a ver com a organização cognitiva e contabiliza o número e dimensão de agrupamentos e transições entre eles.

Para todas estas características, clinicamente interpretáveis, é possível construir intervalos de referência, tal como é vulgar em análises clínicas. Este foi um dos desafios por nós abordados em Botelho *et al.* (2024). A Figura 3 mostra com um sombreado verde os intervalos de referência construídos com base num corpus extenso de pessoas saudáveis (Haulcy e Glass, 2021) para as características mais relevantes que podem ser extraídas em duas tarefas diferentes: vogais sustentadas do lado esquerdo (sexo feminino) e descrição de imagem (sexo masculino), do lado direito. Sobrepostos a estes intervalos, mostram-se nos diagramas de radar do lado esquerdo, os valores obtidos no corpus PC-Gita (Orozco-Arroyave *et al.*, 2014), composto por fala de doentes de Parkinson (a rosa) e controlos (a azul). O mesmo se passa nas figuras do lado direito obtidas com o corpus ADReSS (Luz *et al.*, 2020), com fala de doentes de Alzheimer (a rosa) e de controlos (a azul). Nestas figuras, são bem patentes os maiores desvios relativamente aos intervalos de referência de algumas das características da fala dos doentes de Parkinson e de Alzheimer. A título de exemplo, para os doentes de Parkinson, seis das características consideradas, todas relacionadas com *jitter, shimmer* e HNR, saem fora do intervalo de referência em mais de 20% das gravações de vogais sustentadas, enquanto que, para os controlos do mesmo corpus, isso ocorre em menos de 5% das gravações.

2.4 Modelos de classificação de doenças a partir da fala

Atualmente, é com algoritmos de aprendizagem automática que tipicamente se traça o perfil de um falante, detetando automaticamente os atributos mencionados. Por exemplo, se se alimentar um modelo com fala de pessoas saudáveis e fala de doentes de Parkinson, com a respetiva etiqueta para cada gravação, o modelo poderá ser treinado como classificador desta doença. Num classificador mais convencional, a primeira etapa é tipicamente a extração de múltiplas características a partir do sinal de fala. Atualmente, porém, é possível operar com modelos ponta-a-ponta (*end-to-end*) muito mais complexos, dando como entrada destes modelos diretamente formas de onda no tempo, ou representações do tipo espectrograma, que mostram simultaneamente o sinal nos domínios do tempo e da frequência e podem até ser modificadas para mapear também o mecanismo auditivo humano.

Mas o maior desafio sobretudo na era da IA (inteligência artificial) são os **dados** com que estes modelos são treinados. Muitos modelos de IA são de tal forma complexos, frequentemente com biliões de parâmetros, que exigem quantidade de dados tipicamente não praticáveis na área da saúde para garantir que os modelos têm capacidade de generalizar e não estão sobre treinados para aqueles dados. No caso da fala, isso implicaria a recolha de um grande número de oradores tanto afetados por uma dada doença, como saudáveis, o que está muito longe da realidade dos corpora de fala disponíveis. Por outro lado, a alta complexidade dos modelos faz com que se possam tornar caixas negras. O segundo desafio, tão importante na área da saúde, é assim garantir a **explicabilidade** dos resultados destes modelos.

Estes desafios levam a que se opte muitas vezes por classificadores muito mais simples (p.e., máquinas de suporte vetorial, árvores de decisão, etc.), tarefas de fala bem específicas e, dado o domínio, modelos que operem com características clinicamente interpretáveis. Todas as características acima citadas podem ser extraídas com métodos convencionais. Em alternativa, é também possível capitalizar em modelos de IA pré-treinados para a extração de características clinicamente interpretáveis, como veremos na próxima secção.

2.5 Extração de características interpretáveis com base em modelos pré-treinados

O número crescente de modelos pré-treinados quer de uma forma supervisionada, ou auto supervisionada, faz-nos questionar como tirar partido destes modelos para extrair informação da fala e/ou da respetiva transcrição.

O exemplo mais conhecido de modelo pré-treinado é sem dúvida o dos chamados LLMs (*Large Language Models*), usados para inúmeras aplicações. Falar das muitas famílias de LLMs (Claude, DeepSeek, Gemini, GPT, GroK, LaMDA, Llama, Mistral, PaLM, Phi, Qwen, etc.) sai completamente fora do âmbito desta comunicação, mas uma das suas aplicações mais frequentes é na correção e geração de textos. Isso significa que podem também ser aplicados às transcrições produzidas automaticamente por sistemas de reconhecimento de fala, de forma, por exemplo, a diagnosticar problemas cognitivos. Adiante veremos exemplos da sua utilização para a extração de características interpretáveis.

Um outro exemplo que considero marcante é o dos chamados *x-vectors* (Snyder *et al.*, 2017). Trata-se de um modelo supervisionado do tipo rede neuronal profunda ou DNN (*Deep Neural Network*) que começou por ser proposto para a tarefa de reconhecimento de oradores. Para treinar os *x-vectors*, há que dispor de fala de um grande número de oradores. A rede é treinada de forma a maximizar a probabilidade de cada orador do conjunto de treino, a qual constitui a camada de saída.

Os *x-vectors* derivam precisamente das camadas precedentes, constituindo assim uma representação latente. O progresso que conseguiram em 2018 foi espantoso — cerca de 3% de erro numa tarefa de identificação de oradores num corpus de mais de 6000. Conseguir encapsular toda a informação sobre um falante num vetor que tipicamente tem 512 coeficientes reais foi de facto um marco muito importante com repercussões em termos de privacidade, como veremos adiante. Mas os *x-vectors* não foram apenas um marco importante em reconhecimento do orador. De uma forma equivalente ao treino da rede para reconhecer oradores, é possível treinar redes para reconhecer, por exemplo, línguas ou sotaques (Snyder *et al.*, 2018). Mais ainda, é possível tirar partido do facto dessas representações ou *embeddings* de oradores codificarem intrinsecamente parâmetros que espelham características de patologias como Parkinson ou OSA (Botelho *et al.*, 2020) ou outros atributos atrás mencionados.

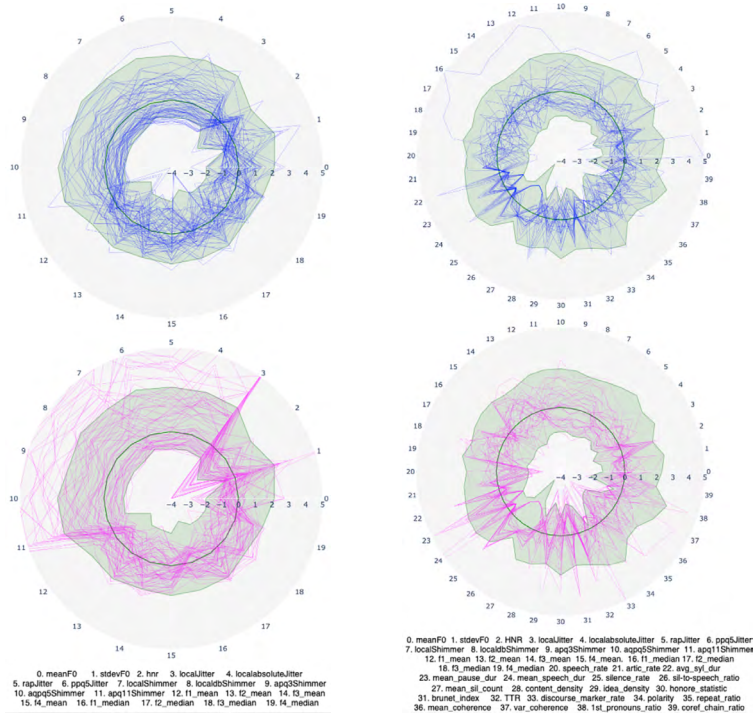

Figura 3. Intervalos de referência (sombreado a verde) para duas tarefas distintas: vogais sustentadas (sexo feminino, à esquerda) e descrição de imagens (sexo masculino, à direita). Sobrepostos a estes intervalos, mostram-se nos diagramas de radar do lado esquerdo, os valores obtidos para falantes dos respetivos sexos, num corpus de fala de doentes de Parkinson (a rosa) e controlos (a azul). O mesmo se passa nas figuras do lado direito obtidas com um corpus de fala de doentes de Alzheimer (a rosa) e de controlos (a azul). Extraída de Botelho *et al.* (2024).

Um terceiro exemplo pretende ilustrar o potencial de modelos auto-supervisionados. Trata-se de um modelo com base em *transformers*, o wav2vec2 lançado pela Meta em 2020 (Baevski *et al.*, 2020). Num paralelo com os LLMs, o modelo wav2vec é treinado para predizer unidades de fala em partes do sinal de áudio que foram mascaradas. O modelo aprende unidades básicas de 25ms, mais curtas que fones, para conseguir aprender representações em contexto de mais alto nível. Como este conjunto de unidade é finito, o modelo aprende a focar-se nos fatores mais importantes que representam o sinal de fala. O modelo aprende tanto de fala gravada como de texto não correspondente, o que diminui imenso a necessidade de transcrições. Estes modelos pré-treinados com uma imensidão de dados de uma forma não supervisionada, podem depois ser ajustados (*fine-tuned*) com uma quantidade muitíssimo menor de dados de uma dada tarefa. De facto, o wav2vec é uma peça fundamental de um reconhecedor recentemente lançado

pela Meta, treinado com mais de 120 mil horas de 1690 línguas (Pratap *et al.*, 2024). Treinando uma rede generativa adversarial com um gerador e um discriminador, ensina o modelo a reconhecer palavras na gravação. Estas representações de unidades em contexto podem ser usadas não só em reconhecimento de fala, mas também em reconhecimento de emoções ou de vários outros atributos.

A par dos exemplos de modelos pré-treinados apresentados, poderíamos ter elencado muitos outros, também com grande sucesso em aplicações de fala, como por exemplo, o WavLM (Chen *et al.*, 2021) e o Hubert (Hsu *et al.*, 2021). As duas subsecções seguintes trazem exemplos de aplicação dos modelos pré-treinados acima citados como extratores de características que fazem da fala um potencial biomarcador de saúde.

2.5.1 Extração de características clinicamente interpretáveis a partir de descrições de imagens

Foi recentemente mostrado que a análise de certas características linguísticas do discurso pode antecipar o diagnóstico da doença de Alzheimer em mais de 5 anos (Eyigoz *et al.*, 2020). Esta correlação motivou um grande número de estudos sobre o discurso de pessoas com declínio cognitivo e doença de Alzheimer. Esse foi também o objetivo do estudo de Botelho *et al.* (2023). Numa tentativa de conciliar modelos do tipo caixa negra com explicabilidade, o trabalho mostrou como é possível capitalizar em modelos LLM para a extração de características linguísticas associadas a declínio cognitivo. O trabalho focou-se em quatro destas características: coerência textual, diversidade lexical, dificuldade em encontrar palavras e comprimento da frase.

A extração de características foi realizada sobre transcrições automáticas numa tarefa de descrição de imagens. Foram testados vários sistemas de reconhecimento de fala, vários LLMs e várias técnicas de *prompting* para cada LLM. A Figura 4 mostra um exemplo de transcrição e valores das quatro características obtidos para a combinação de testes com melhores resultados, envolvendo o reconhecedor Whisper (Radford *et al.*, 2022) e o LLM Mistral (Jiang *et al.*, 2023). Inclui também as distribuições de valores das mesmas características para doentes de Alzheimer e controlos, sendo bem patentes os desvios entre ambas, particularmente para a diversidade lexical e coerência textual.

Usando estas quatro características de alto nível como entrada de vários classificadores, obtiveram-se valores de exatidão superiores a 80% no corpus ADReSS, mostrando que, embora os LLMs não tenham sido treinados para detetar a doença de Alzheimer, são bastante eficientes na deteção de características interpretáveis do discurso que são vulgares nesta doença.

O estudo tem muitas limitações, sendo a principal a necessidade de ser replicado em corpora muito mais significativos e contendo tanto doentes de Alzheimer como falantes com declínio cognitivo. Mas apesar destas limitações, pensamos que o uso de LLMs como extratores de características interpretáveis é uma direção muito promissora para uma maior explicabilidade no uso da fala como biomarcador de saúde.

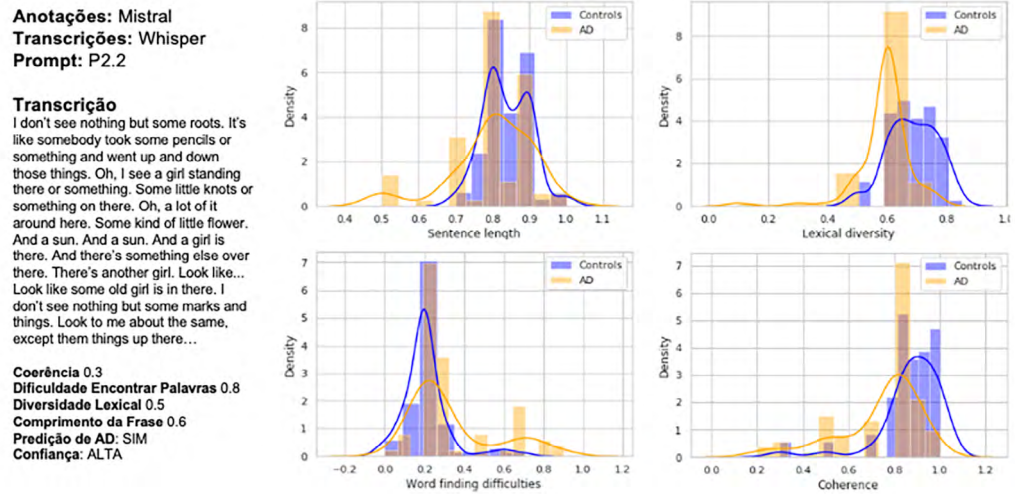

Figura 4. Distribuição das características de alto nível em doentes de Alzheimer e controlos. Extraída de Botelho *et al.* (2023).

2.5.2 Extração de características clinicamente interpretáveis a partir de testes de fluência verbal

A estratégia de usar modelos pré-treinados como extratores de características interpretáveis foi também adotada para uma tarefa de fala muito menos explorada — os testes de fluência verbal — aplicada à deteção de psicose e doença bipolar. O trabalho de Ponte *et al.* (2026) propõe uma arquitetura com 4 grandes blocos (Figura 5). O primeiro é obviamente o módulo de processamento de áudio, em que o reconhecedor de fala desempenha o papel central, complementado com

um detetor de atividade vocal e um detetor de pausas preenchidas, tão frequentes nesta tarefa. A transcrição resultante é a entrada do segundo módulo cujo objetivo é distinguir entre palavras-alvo (tipicamente, animais em testes semânticos, ou palavras começadas por uma dada letra, como "p", em testes fonéticos), repetições e intrusões (p.e., *Já não me lembro de mais animais*). Mais uma vez, os LLMs provaram ser um classificador muito eficiente para esta distinção, com taxas de erro da ordem dos 2%.

Segue-se a extração de características acústico-prosódicas, linguísticas e relacionadas com a dinâmica das respostas (p. e. débito de fala, contagens de pausas preenchidas). Na primeira classe, foram também incluídas características tipicamente relacionadas com emoções, tais como valência e arousal, extraídas com base num modelo wav2vec pré-treinado para a deteção de emoções (Wagner *et al.*, 2023). Uma quarta classe é particularmente relevante para este tipo de testes — as características de organização cognitiva. Aqui, mais uma vez, modelos pré-treinados como o WavLM (Chen *et al.*, 2022) e os próprios LLM mostraram-se muito eficientes na identificação de agrupamentos de palavras sucessivas. O módulo final é o classificador. Os melhores resultados foram obtidos com um classificador hierárquico que começa por distinguir entre fala de sujeitos saudáveis ou não e só depois distingue entre participantes com psicose ou doença bipolar. Com máquinas de suporte vetorial, conseguem-se taxas de 90%, que são de facto muito promissoras, especialmente se tivermos em conta que a duração total dos dois testes é de 2 minutos de fala.

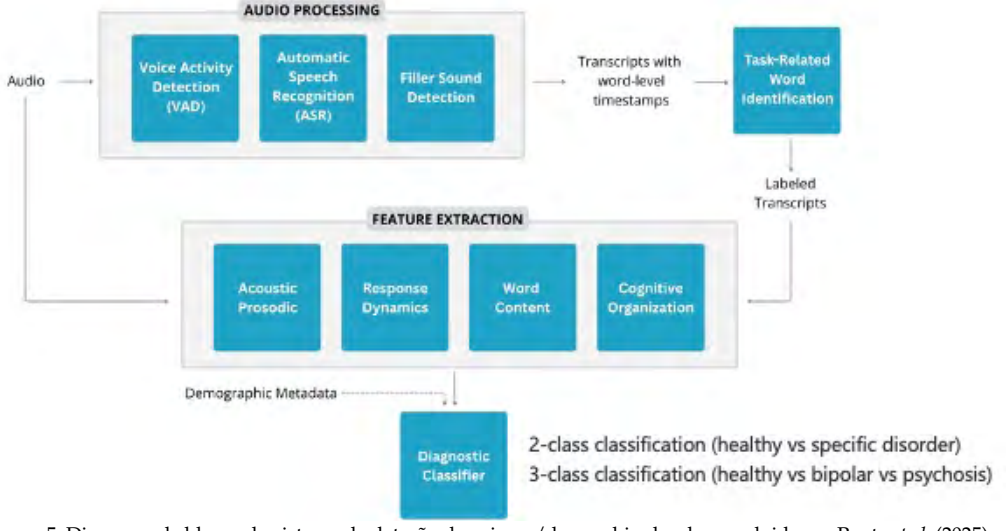

Figura 5. Diagrama de blocos do sistema de deteção de psicose/doença bipolar desenvolvido em Ponte *et al.* (2025).

Esta mesma arquitetura aplicada à deteção de doença de Alzheimer consegue valores de 84% numa tarefa binária de deteção saudável/Alzheimer. Quer em tarefas de descrição de imagens, quer de fluência verbal, ao incluirmos também participantes com declínio cognitivo, a deteção vai falhar muito mais em casos de fronteira, pelo que juntar outras tarefas e delas extrair mais características interpretáveis faz todo o sentido se quisermos de facto usar a fala como janela para a saúde mental.

3. COMO PROTEGER O PERFIL DE UM FALANTE

O facto de se poder extrair tantas pistas a partir da fala pode também constituir uma grande ameaça à privacidade. Na Tabela 1, elaborada pelo Prof. Tom Backstrom, coordenador da rede doutoral PSST (Privacy for Smart Speech Technology[3]) de que fazemos parte, listam-se vários exemplos práticos de ameaças à privacidade.

Ameaça	**Exemplo**
Aumento abusivo de preços	Sinais de depressão ou outros problemas de saúde na voz dos utilizadores podem ser indevidamente explorados para provocar um aumento nos respetivos prémios de seguro. Sinais das emoções dos utilizadores podem ser explorados para oferecer produtos a preços mais altos.
Rastreamento, perseguição	A reidentificação da voz pode permitir detetar o mesmo utilizador em plataformas diferentes, desde redes sociais relacionadas com a profissão, a grupos de apoio online ou aplicações de namoro, por exemplo.
Extorsão, humilhação pública	Problemas de saúde e casos amorosos podem ser detetados na voz e usados para chantagem ou tornados públicos contra a vontade do utilizador.
Estereotipagem algorítmica	Sistemas de recomendação baseados em voz podem tornar-se tendenciosos em relação à idade, religião ou etnia, de formas que são quase impossíveis de monitorizar.
Assédio, investidas inapropriadas	Utilizadores em salas de chat ou realidade virtual podem ser automaticamente selecionados por género ou opiniões, tornando-os alvo de atenção indesejada e assédio.
Medo de monitorização	A sensação subjetiva de estar a ser continuamente monitorizado pode causar danos psicológicos. Também pode sufocar a expressão política, prejudicando as sociedades democráticas.

Tabela 1. Exemplos práticos de ameaças à privacidade (traduzida de Backstrom, 2025).

[3] https://psst-doctoralnetwork.eu/

Toda esta informação,que permite traçar o perfil de um orador a partir da fala, faz com que ela possa ser considerada como PII (*Personal Identifiable Information*), de acordo com o RGPD (Nautch *et al.*, 2019b). Mas a possibilidade de traçar o perfil de um orador não é a única vulnerabilidade da fala. Hoje em dia, o progresso em síntese de fala para qualquer voz é de tal ordem que se torna já difícil distinguir oradores verdadeiros de sintéticos. Chegámos à era dos *deep fakes*.

Neste capítulo, começaremos por rever de uma forma extremamente breve a evolução na área de síntese de fala e as crescentes ameaças em termos de segurança levantadas pela facilidade em falsificar oradores (Secção 3.1). Estas ameaças de privacidade e segurança motivam um investimento exponencial na investigação em processamento da fala em servidores remotos, por forma a garantir que possa ser feito sem acesso ao sinal de fala original. A secção 3.2 apresenta os principais tipos de abordagem a este problema. A última secção deste capítulo apresenta como exemplo um dos métodos que desenvolvemos para esse fim.

3.1 Falsificação de oradores

Até há uns dez anos, os modelos de síntese de fala a partir de texto (*Text-to-Speech* – TTS) eram de tal forma limitados que se tornava bem difícil construir uma nova voz. O nosso primeiro sintetizador, o DIXI, desenvolvido no início dos anos 90 (Oliveira *et al.*, 1991), produzia uma única voz, masculina, robótica, obtida com base em modelos de sintetizadores de formantes (Klatt, 1980). Apesar da flexibilidade de controlo destes modelos paramétricos, a naturalidade da fala sintética era muito baixa. Mais tarde surgiram modelos do tipo concatenativo (Hunt & Black, 1996), representados no nosso grupo pelo DIXI+ (Oliveira *et al.*, 2001). A qualidade obtida por este tipo de modelos era bastante superior, mas, em contrapartida, estes modelos eram mais limitados no controlo da expressividade e gerar novas vozes exigia repositórios de muitas horas de gravação. A geração dos sintetizadores estatísticos paramétricos (Tokuda *et al.*, 2002) que se seguiu, e que foi representada no grupo por Paulo *et al.*, (2008), tentou combinar as vantagens dos anteriores modelos, mas as trajetórias geradas para os parâmetros eram demasiadamente suavizadas, o que afetava a naturalidade da fala sintética.

Quando as DNNs se começaram a generalizar, há uns dez anos atrás, o salto qualitativo em síntese de fala foi enorme, produzindo-se pela primeira vez fala de alta qualidade e expressividade, com base em modelos que aprendem diretamente de corpora de fala muito extensos. Alguns marcos importantes da primeira geração de sintetizadores neuronais são o WaveNet (Van Den Oord *et al.*, 2016), o Tacotron (Wang *et al.*, 2017) e o Deep Voice (Arik *et al.*, 2017). As arquiteturas foram evoluindo, desde modelos do tipo *encoder-decoder* combinados com *vocoders* até modelos ponta-a-ponta, que integram o próprio *vocoder* na sua arquitetura.

Mas já nos primeiros modelos, a possibilidade de separar representações linguísticas (extraídas do texto) de representações do orador, permitia um fácil re-treino (*fine-tuning)* para novas vozes, com base num conjunto muito pequeno de frases faladas por um novo orador (Zhang *et al.*, 2020). Construir novas vozes passou assim a ser muito fácil e a fronteira entre modelos cuja entrada é texto (síntese) ou uma nova voz alvo (conversão) esbateu-se.

Atualmente, com modelos ponta-a-ponta como o XTTS (Casanova *et al.*, 2024), ou o F5-TTS (Chen *et al.*, 2024), entre muitos outros, consegue-se fala sintética soando quase como o original e sem necessidade de retreino (*zero-shot*). O progresso atingido por este último sintetizador com modelos de difusão é particularmente notório. Isto levanta uma série de problemas éticos. O desafio agora é como detetar falsificações. Daí o investimento crescente em técnicas de *anti-spoofing* (Dao *et al.*, 2025) (especialmente vocacionadas para impedir a verificação do orador usando fala sintética), marcas de água (Özer *et al.*, 2026), etc.

3.2 Preservação da privacidade em processamento remoto da fala

O grande desafio de processar remotamente a fala preservando a sua privacidade pode ser abordado com várias classes de métodos tal como esquematizado na Figura 5 (Teixeira, 2024). Na primeira classe, incluem-se métodos criptográficos do tipo cifra homomórfica, computação multipartidária segura (*Secure Multi-Party Computation*), ou *Limited Leakage Hashing*, por exemplo. Esta classe de métodos pode ser aplicada a tarefas em que se torna difícil separar a informação relacionada com o orador e com a tarefa e pode fornecer garantias formais de privacidade. As desvantagens em termos de custos de computação e comunicação, contudo, têm impedido a sua adoção, na prática.

Na segunda classe, incluem-se métodos que manipulam a fala para aumentar a sua privacidade. O objetivo desta manipulação poderá ser quer anonimizar uma dada locução, quer extrair de uma forma privada atributos do orador, sem acesso à sua fala em claro, quer ainda apenas esconder algum dos atributos do orador que, como vimos, estão embebidos em representações do tipo *x-vector*, por exemplo. Esta classe de métodos é tipicamente caracterizada por uma complexidade muito inferior à primeira classe, o que viabiliza a sua implementação nos dispositivos do utilizador, mas, em contrapartida, não oferece garantias formais de privacidade, apenas empíricas.

Estas duas grandes classes não esgotam a grande panóplia de métodos cujo alvo é a privacidade no processamento remoto da fala. Há que juntar-lhes, entre muitos outros, técnicas de privacidade diferencial, aprendizagem federada e enclaves seguros.

Figura 6. Processamento remoto de fala – classes de métodos.

3.3 Manipulação de atributos do orador

O trabalho do grupo de investigação nesta área focou-se inicialmente na primeira das classes de métodos (Portêlo *et al.*, 2013; Teixeira *et al.*, 2022), mas mais recentemente tem-se debruçado sobre a minimização da informação de alguns atributos do orador (Teixeira *et al.*, 2024). A arquitetura explorada foi um *vector-quantised variational autoencoder*, como esquematizado na Figura 6, que provou ter uma capacidade inerente para destrinçar informação ao nível do módulo de quantização. Como não se tem acesso a dados emparelhados com diferentes

valores do mesmo atributo, há que condicionar o descodificador com um classificador de atributo externo, tal que, aquando do teste, o descodificador possa usar essa informação para reconstruir a saída, permitindo que o atributo seja removido ou modificado. Para promover a tal separação entre a informação relacionada com o orador e a relacionada com a tarefa, foi usado um classificador adversarial e uma função de custo de informação mútua, que tem como função ensinar o modelo a remover informação, em vez de simplesmente a reajustar no espaço latente, o que a tornaria recuperável usando classificadores retreinados. Com esta arquitetura, conseguiu-se efetivamente manipular atributos como o sexo biológico ou a faixa etária.

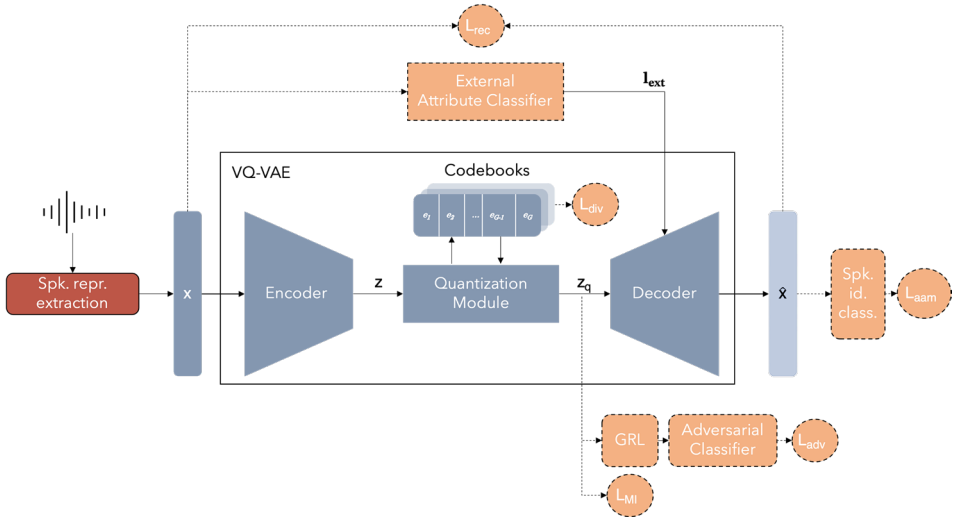

Figura 7. Manipulação de representações do orador. Extraída de Teixeira *et al.* (2024).

Embora o objetivo deste estudo não tenha sido a anonimização, poderá fornecer contribuições relevantes para o desenvolvimento de sistemas de anonimização em que seja possível controlar alguns dos atributos do orador. Este é um dos muitos desafios desta área de investigação, buscando sempre um compromisso entre utilidade e privacidade. Se por um lado, há que garantir a utilidade da fala anonimizada para a tarefa em questão, por outro lado, há que garantir que o orador original não é reconhecível, mas também que o método não é invertível. Um dos desafios que junta os dois tópicos focados nesta comunicação é a anonimização da fala para aplicações em saúde, dado que em todos os métodos de anonimização mais complexos se perdem muitas pistas sobre a saúde do orador original.

Para além da preocupação com a privacidade dos sinais de fala quando partilhados, há também uma preocupação crescente com a privacidade dos dados que são usados para treinar modelos. Há assim que estudar a forma de um dado modelo "esquecer" parte da informação com que foi treinado, por exemplo, para retirar a fala de um dado orador, a pedido deste. Este é também o tópico de um recente projeto exploratório no grupo (LeaF – *Learning to Forget*)[4].

Estes são apenas alguns dos temas em aberto entre muitos outros ligados à privacidade e segurança em fala. De facto, os progressos esperados nas tecnologias de fala denominadas como "inteligentes" (ou *smart speech technologies*) abrem a possibilidade de inúmeras ameaças à privacidade com consequências potencialmente muito nefastas a nível individual, societal, ético e económico. Daí a importância crescente de fazer face a ameaças emergentes.

4. CONCLUSÕES

À laia de conclusão, deixo duas mensagens que espero ter conseguido passar ao longo desta comunicação. Em primeiro lugar, a fala é uma modalidade ubíqua e não intrusiva que pode ser usada como biomarcador de muitos atributos do falante, nomeadamente de saúde. E neste domínio em particular, onde é tão importante a transparência dos resultados, podemos capitalizar em modelos de larga escala para extrair características interpretáveis. Quem sabe, talvez um dia a análise da fala poderá vir a ser tão comum como uma análise ao sangue para o diagnóstico de certas doenças.

Apresenta, porém, vulnerabilidades. Pode ser considerada como PII, o que exige progresso em termos de preservar a sua privacidade, com base em métodos criptográficos ou outras formas de anonimizar/encobrir atributos do falante; e pode ser falsificada, o que exige progresso em deteção de fala sintética/marcadores de água, etc. Temos assim pela frente um longo caminho a percorrer cheio de enormes desafios.

Estando à frente do Comité Científico de um projeto PRR intitulado CRAI[5] — Centro para a IA Responsável — não gostaria de terminar sem realçar a necessidade

[4] http://leaf.github.io

[5] https://centerforresponsible.ai/

de usar a IA de uma forma responsável ao processar fala. Nesta apresentação foquei-me apenas em alguns dos seus pilares: explicabilidade, privacidade e segurança. Mas há muitos outros tais como equidade, sustentabilidade, robustez, etc. também muito relevantes no contexto das tecnologias de fala.

A regulamentação europeia para o sector digital é densa e muito complexa. Só o AI Act tem180 considerandos, 113 artigos e 13 anexos, o que não o torna fácil para engenheiros ou médicos, ou muitas outras profissões em que a IA se tornou imprescindível. Daí a importância crescente de adotarmos no nosso trabalho os pilares da IA responsável.

Termino com um enorme agradecimento a todo o meu grupo de investigação (Human Language Technology@INESC-ID) e à Academia das Ciências de Lisboa, pela honra concedida.

COMUNICAÇÃO APRESENTADA À CLASSE DE CIÊNCIAS
NA SESSÃO DE 27 DE NOVEMBRO DE 2025

COMUNICAÇÃO RECEBIDA A 19 DE FEVEREIRO DE 2026

BIBLIOGRAFIA

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 195-204). *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v70/arik17a.html

Bäckström, T. (2025). Privacy in speech technology. *Proceedings of the IEEE, 113*(7), 668-692. https://doi.org/10.1109/JPROC.2025.3632102

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449-12460.

Botelho, C., Teixeira, F., Rolland, T., Abad, A., & Trancoso, I. (2020). Pathological speech detection using x-vector embeddings. *arXiv*. https://arxiv.org/abs/2003.00864

Botelho, C., Abad, A., Schultz, T., & Trancoso, I. (2024). Speech as a biomarker for disease detection. *IEEE Access, 12*, 184487-184508. https://doi.org/10.1109/ACCESS.2024.3506433

Botelho, C., Mendonça, J., Pompili, A., Schultz, T., Abad, A., & Trancoso, I. (2024). Macro-descriptors for Alzheimer's disease detection using large language models. In *Proceedings of Interspeech 2024* (pp. 1975-1979). https://doi.org/10.21437/Interspeech.2024-1255

Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., & Weber, J. (2024). XTTS: A massively multilingual zero-shot text-to-speech model. In *Proceedings of Interspeech 2024* (pp. 4978-4982). https://doi.org/10.21437/Interspeech.2024-2016

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., & Chen, Z. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing, 16*(6), 1505-1518. https://doi.org/10.1109/JSTSP.2022.3188113

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., & Chen, X. (2025). F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (pp. 6255-6271). https://doi.org/10.18653/v1/2025.acl-long.313

Dao, A.-T., Rouvier, M., & Matrouf, D. (2024). ASVspoof 5 challenge: Advanced ResNet architectures for robust voice spoofing detection. In *Proceedings of the Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)* (pp. 163–169). https://doi.org/10.21437/ASVspoof.2024-24

Elbéji, A., Pizzimenti, M., Aguayo, G., Fischer, A., Ayadi, H., Mauvais-Jarvis, F., Riveline, R., Despotovic, V., & Fagherazzi, G. (2024). A voice-based algorithm can predict type 2 diabetes status in USA adults: Findings from the Colive Voice study. *PLOS Digital Health, 3*(12), e0000679. https://doi.org/10.1371/journal.pdig.0000679

Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., & Naylor, M. (2020). Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine, 28*, 100583. https://doi.org/10.1016/j.eclinm.2020.100583

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Shriberg, E., & Rodero, R. (2016). The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), pp. 190-202. https://doi.org/10.1109/TAFFC.2015.2457417

Fischer, J., Semple, S., Fickenscher, G., Jürgens, R., Kruse, E., Heistermann, M., & Amir, O. (2011). Do women's voices provide cues of the likelihood of ovulation? *PLoS ONE, 6*(9), e24490. https://doi.org/10.1371/journal.pone.0024490

Haulcy, R., & Glass, J. (2021). CLAC: A speech corpus of healthy English speakers. In *Proceedings of Interspeech 2021* (pp. 2966-2970). https://doi.org/10.21437/Interspeech.2021-1810

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451–3460.

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 373-376). https://doi.org/10.1109/ICASSP.1996.541110

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T.,

Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *arXiv*. https://doi.org/10.48550/arXiv.2310.06825

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*(3), 971-995. https://doi.org/10.1121/1.383940

Luz, S., Haider, F., Fuente, S. de la, Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In *Proceedings of Interspeech 2020* (pp. 2172-2176). https://doi.org/10.21437/Interspeech.2020-2571

Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M. A., Mtibaa, A., Abdelraheem, M. A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., Bonastre, J.-F., Raj, B., Trancoso, I., & Busch, C. (2019). Preserving privacy in speaker and speech characterisation. Computer Speech & Language, 58, 441-480. https://doi.org/10.1016/j.csl.2019.06.001

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., & Evans, N. (2019). The GDPR & speech data: Reflections of legal and technology communities. In *Proceedings of Interspeech 2019* (pp. 3695-3699). https://doi.org/10.21437/Interspeech.2019-2647

Oliveira, L. C., Viana, M. C., & Trancoso, I. M. (1991). DIXI: Portuguese text-to-speech system. In *Proceedings of Eurospeech 1991* (pp. 1239-1242). https://doi.org/10.21437/Eurospeech.1991-284

Oliveira, L. C., Viana, M. C., Mata, A. I., & Trancoso, I. (2001). *Progress report of project DIXI+: A Portuguese text-to-speech synthesizer for alternative and augmentative communication* (Technical report). FCT.

Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., González-Rátiva, M. C., & Nöth, E. (2014). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 342-347).

Özer, Y., Ge, W., Zhang, Z., Wang, X., & Yamagishi, J. (2026). Self voice conversion as an attack against neural audio watermarking. *arXiv*. https://doi.org/10.48550/arXiv.2601.20432

Paulo, S., Oliveira, L. C., Mendes, C., Figueira, L., Cassaca, R., Viana, C., & Moniz, H. (2008). DIXI: A generic text-to-speech system for European Portuguese. In Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (Eds.), *Computational processing of the Portuguese language* (Lecture Notes in Computer Science, Vol. 5190). Springer. https://doi.org/10.1007/978-3-540-85980-2_10

Ponte, M., Botelho, C. & Trancoso, I. (2026). Verbal Fluency Tasks as Diagnostic Tools for Speech--Affecting Disorders, In IEEE Melecon 2026.

Portêlo, J., Abad, A., Raj, B., & Trancoso, I. (2013) Secure binary embeddings of front-end factor analysis for privacy preserving speaker verification. Proc. Interspeech 2013, 2494-2498, https://doi.org/10.21437/Interspeech.2013-417.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., & Auli, M. (2024). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research, 25*, 1-52. https://doi.org/10.48550/arXiv.2305.13516

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2212.04356

Singh, R. (2019). *Profiling humans from their voice*. Springer.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Proceedings of Interspeech 2017* (pp. 999-1003). https://doi.org/10.21437/Interspeech.2017-620

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., & Khudanpur, S. (2018). Spoken language recognition using x-vectors. In *Proceedings of the Odyssey Speaker and Language Recognition Workshop* (pp. 105-111). https://doi.org/10.21437/Odyssey.2018-15

Teixeira, F., Abad, A., Raj, B., & Trancoso, I. (2022). Towards end-to-end private automatic speaker recognition. In *Proceedings of Interspeech 2022* (pp. 2798-2802). https://doi.org/10.21437/Interspeech.2022-10672

Teixeira, F., Abad, A., Raj, B., & Trancoso, I. (2024). Privacy-oriented manipulation of speaker representations. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.39862

Teixeira, F. (2024). *Privacy-preserving machine learning for remote speech processing* (Doctoral dissertation, Instituto Superior Técnico, Universidade de Lisboa).

Tokuda, K., Zen, H., & Black, A. W. (2002). An HMM-based speech synthesis system applied to English. In *Speech Synthesis Workshop*. https://doi.org/10.1109/WSS.2002.1224415

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv*. https://arxiv.org/abs/1609.03499

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*, 10745-10759. https://doi.org/10.1109/TPAMI.2023.3263585

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. (2017). Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv*. https://doi.org/10.48550/arXiv.1703.10135

Zhang, J.-X., Ling, Z.-H., & Dai, L.-R. (2020). Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28*, 540-552. https://doi.org/10.1109/TASLP.2019.2960721