Da Escrita à Fala - Da Fala à Escrita

Isabel Trancoso, Luís Oliveira, João Neto
INESC

M. Céu Viana
CLUL

1. Introdução

Numa exposição dedicada ao tema "A Escrita" que sentido fará incluir também "A Fala"? Poderíamos encher muitas páginas comparando as duas formas de comunicação, mas não é esse o nosso propósito. Pretendemos, em vez disso, discutir o problema da conversão (automática) entre ambas. Na literatura de língua inglesa, a conversão entre fala/escrita e escrita/fala é tipicamente descrita por duas siglas - ASR (Automatic Speech Recognition) e TTS (Text-to-Speech) - que traduziremos aqui por reconhecimento e síntese de fala, respectivamente.

Fazer um computador reconhecer fala e sintetizá-la já não é ficção científica, mas ainda estamos muito longe de poder conversar com um computador como o HAL, tal como na visão de Kubrick e Clarke em 2001 Odisseia no Espaço. De facto, as capacidades do HAL excediam muito aquilo que se costuma englobar em reconhecimento e síntese de fala: ele era capaz de a "compreender", o que implica muito mais que uma simples conversão fala-texto, e era capaz de a gerar a partir de conceitos, o que também excede claramente a mera conversão texto-fala.

Não podemos, no entanto, deixar aqui a ideia de que a conversão fala/texto ou vice-versa é um problema simples e resolvido. Muito pelo contrário, trata-se de uma área de trabalho bastante complexa e, sobretudo, fortemente interdisciplinar. Envolve conhecimentos de engenharia (processamento de sinais, aprendizagem automática, técnicas de busca, etc.) e de linguística (fonética, fonologia, prosódia, morfologia, sintaxe, ...), mas não só. A interacção Homem-Máquina através da fala coloca desafios enormes que mal começámos a conquistar.

Começaremos por explicar para que servem o reconhecimento e a síntese de fala e descrever como são feitos, mencionando os principais componentes. Apresentaremos em seguida as demonstrações seleccionadas para figurar nesta exposição. Elas fazem parte do trabalho de há vários anos de uma equipa multi-disciplinar que engloba muitos outros colegas para quem vai, naturalmente, o nosso agradecimento. Finalizaremos com um breve resumo das perspectivas futuras nesta área.

2. Domínios de aplicação

Numa era em que palavras como *multimédia* e *multimodal* já entraram no léxico de todos os dias, não é de estranhar a importância crescente da fala como modalidade de interacção com o computador. A fala é, de facto, a modalidade preferível em situações de ocupação dos olhos e/ou mãos, sempre que seja impossível utilizar teclados, ratos ou écrãs, e também em casos de deficiência (sobretudo visual e auditiva). Muitas vezes, porém, a utilização da fala não é imperativa, mas sim vantajosa, o que acontece muitas vezes quando se deseja utilizar língua natural.

Os domínios de aplicação são assim numerosos, com especial ênfase para os serviços de telecomunicações, as aplicações de escritório, o controlo industrial, as aplicações domésticas / lúdicas / educativas, as aplicações militares, e a ajuda a pessoas portadoras de alguns tipos de deficiência.

Os dois primeiros domínios citados são sem dúvida aqueles em que encontramos com maior facilidade já hoje em dia aplicações do reconhecimento e da síntese de fala. A automatização total ou parcial de serviços de informação telefónica, a marcação oral

de números de telefone, o acesso a bases de dados remotas (e.g., cotações da bolsa, horários de combóios e aviões, preenchimento de inquéritos, etc.), os sistemas de ditado automáticos, são exemplos de aplicações que já não fazem parte da ficção científica.

3. Descrição funcional

A complexidade de um sistema de reconhecimento de fala pode variar extraordinariamente dependendo de inúmeros factores tais como a dependência ou independência do orador, a dimensão do vocabulário, a capacidade de funcionar em ambientes mais ou menos ruidosos e o tipo de fala: lida ou espontânea, palavras isoladas ou fala contínua. Tomemos como exemplo um sistema razoavelmente complexo, utilizado numa aplicação de ditado, com um vocabulário extenso (p.e. 5000 palavras), desenhado para funcionar com qualquer orador num ambiente do tipo escritório. O diagrama de blocos típico de um sistema deste tipo está representado na Fig.1.

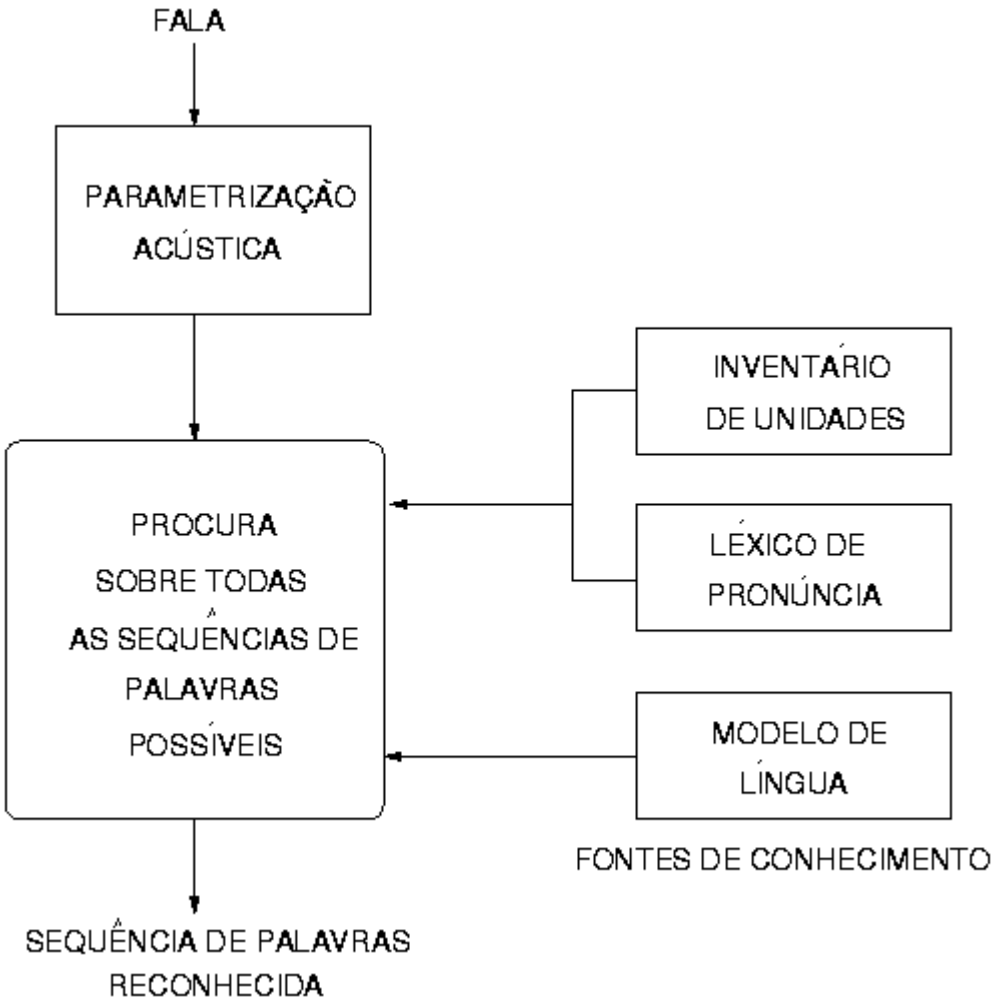

Fig. 1 – Diagrama de blocos de um sistema ASR.

O primeiro bloco no processamento do sinal de fala consiste na extracção de parâmetros acústicos relevantes para o reconhecimento da sequência de palavras faladas. Esta extracção pretende eliminar informação redundante existente no sinal de

fala, que tenha por exemplo a ver com as características vocais do orador. Tipicamente, extrai-se um conjunto de parâmetros relativos à distribuição espectral cem vezes por segundo. Esta sequência de conjuntos de parâmetros constitui uma das entradas do bloco que, do ponto de vista computacional, é o mais complexo. Este bloco tem como função procurar de entre todas as sequências de palavras possíveis a que foi dita com maior probabilidade. Para isso precisa de duas fontes de conhecimento adicionais: uma que indique qual a probabilidade da sequência de parâmetros extraída para cada uma das palavras possíveis, e outra que indique qual a probabilidade de uma dada palavra se seguir às anteriores já reconhecidas. No cálculo da primeira são tipicamente envolvidos modelos para todos os fones possíveis da língua, juntamente com um léxico de pronúncia que associa a cada palavra o seu desdobramento em sequência de fones. No cálculo da segunda, tipicamente, usam-se modelos de bigramas e trigramas (indicando a probabilidade de uma palavras dada a anterior ou as duas anteriores, respectivamente). Métodos como os modelos de Markov não observáveis e as redes neuronais estão na base dos sistemas mais bem sucedidos nesta área.

A complexidade de um sistema de síntese de fala também pode variar significativamente conforme o domínio de aplicação. Se se tratar de um sistema capaz apenas de sintetizar um conjunto de mensagens ou palavras previamente gravadas, a complexidade é mínima. Na realidade, um tal sistema não deveria ser apelidado de *sintetizador*, embora a confusão seja muito comum. A designação de *sintetizador* é tipicamente reservada para um sistema capaz de sintetizar fala a partir de qualquer texto. Entre os dois extremos podemos encontrar muitas variantes de complexidade média, como, por exemplo, um sistema capaz de sintetizar todas as sequências de dígitos possíveis, com a entoação típica de um número de telefone ou de um cartão de crédito. O diagrama de blocos da Fig. 2 diz respeito a um verdadeiro sistema de síntese a partir de texto sem restrições de vocabulário.

A conversão entre o texto de entrada e a fala passa por duas principais etapas: o processamento linguístico e a geração da forma de onda. O primeiro módulo gera a partir do texto a sequência de fones correspondente, juntamente com a informação prosódica que especifica o ritmo e a entoação desejados. O segundo gera o sinal de fala correspondente a cada fone (ou mais propriamente cada sequência de fones). Como seria de esperar, o primeiro módulo é fortemente dependente da língua. Compreende submódulos de normalização do texto (números, datas, quantias, abreviaturas, etc.), análise morfo-sintáctica (e.g. nome *almoço* e verbo *almOço*), conversão de grafemas para fones e geração de informação prosódica. A utilização crescente de métodos de aprendizagem automática no desenvolvimento destes submódulos tem tido uma influência significativa na aceleração da construção de sintetizadores para novas línguas. O segundo módulo pode ser realizado recorrendo a técnicas substancialmente diferentes, com uma forte componente de processamento de sinal. A mais comum, hoje em dia, baseia-se na concatenação de unidades pré-gravadas. Muitos são os sistemas existentes que concatenam difones (parte final de um fone e parte inicial do seguinte) extraídos de um conjunto grande de palavras faladas pelo dono da *voz* do sintetizador. Cada vez mais, porém, se consegue tirar partido do crescente poder computacional e capacidade de armazenamento para construir sistemas baseados na concatenação de sequências de um número variável de fones extraídas automaticamente de bases de dados muito extensas.

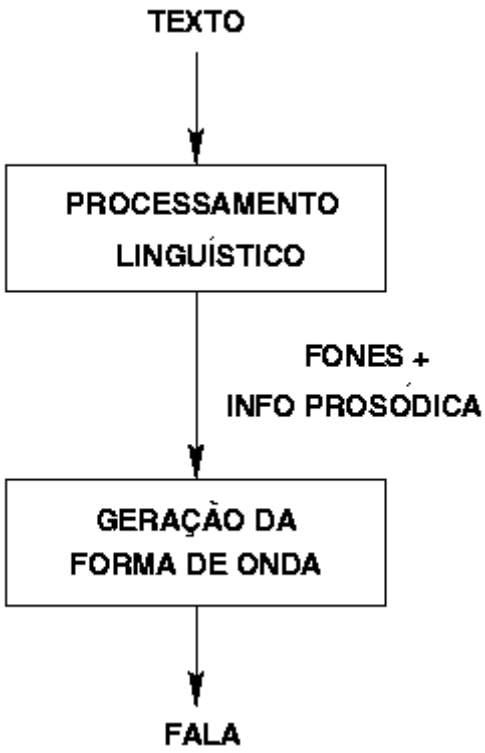

Fig. 2 – Diagram de blocos de um sistema TTS.

4. As actividades do INESC/CLUL neste domínio

As actividades do INESC (Instituto de Engenharia de Sistemas e Computadores) em processamento de fala datam da sua criação em 1980, sendo o ênfase principal na altura dado ao trabalho sobre codificação, que é completamente independente da língua em questão. Cerca de uma década depois, as actividades principais centravam-se já sobre reconhecimento e síntese de fala, no âmbito mais genérico do processamento computacional do Português falado. Para esta mudança, muito contribuíu o convénio estabelecido formalmente com o CLUL (Centro de Linguística da Universidade de Lisboa) e o intenso trabalho multidisciplinar que daí resultou.

De entre os principais produtos desta investigação relevantes para a presente exposição, salientamos os sistemas **AUDIMUS** e **DIXI**. O primeiro é um sistema de reconhecimento de fala de vocabulário extenso, independente do orador, para aplicações de ditado automático. O sistema, baseado num modelo híbrido redes neuronais - modelos de Markov, é bastante flexível, podendo o seu desempenho variar com a complexidade dos módulos envolvidos. Isto é, a taxa de erros de recohecimento obtida com a versão mais complexa é significativamente menor que a da versão mais simples. O desempenho pode assim ser ajustado às restrições do ponto de vista computacional e de funcionamento em tempo real. Tem também a capacidade de se adaptar ao orador e poder ser treinado com modelos de língua específicos de determinados domínios (ex: direito, medicina, etc.).

O segundo é um sistema de síntese de fala a partir de texto. A primeira versão que data de 1991 é um sistema de síntese por regra, baseado num modelo de formantes. Apesar da qualidade limitada desta primeira versão, ela foi julgada suficiente para apoiar as crianças do Centro de Paralisia Cerebral Calouste Gulbenkian. A aplicação desenvolvida para esse fim, o EDIXI, consiste num editor de texto que conjuga as capacidades de síntese com as de aceleração de escrita, bastante importantes em

situações de deficiência motora e de fala. A versão actual é baseada na concatenação de difones, demonstrando uma qualidade e inteligibilidade muito superiores.
De entre os marcos importantes da nossa actividade nesta área, salienta-se também o sistema vocal de informação telefónica (SVIT) desenvolvido para a Portugal Telecom. Coube ao INESC a síntese automática dos números de telefone. A elevada qualidade e naturalidade obtidas podem ser confirmadas ligando para o serviço 118. Trata-se, é claro, de um sistema de síntese de segmentos pré-gravados, embora bastante elaborado. Em termos de reconhecimento salienta-se também o desenvolvimento de protótipos de demonstração de reconhecedores através da rede telefónica para dígitos isolados, dígitos ligados, números naturais, nomes de terras, pessoas e companhias, etc.. É também de mencionar o protótipo de identificação da língua falada treinado e testado com 6 línguas europeias.
Uma parte muito significativa da nossa actividade não transparece das demonstrações apresentadas. De facto, o desenvolvimento de sistemas de reconhecimento e síntese só é possível com base em recursos linguísticos muito vastos e que não existiam de todo para o Português falado há uma década atrás. Temos, portanto, investido fortemente na criação de corpora (ou bases de dados) de fala e texto, e de léxicos de pronúncia.

5. Perspectivas para o futuro

Apesar de todo este esforço, estamos ainda longe de atingir o nível de qualidade obtido por sistemas de reconhecimento e síntese para línguas como o Inglês, por exemplo. O número e a dimensão das equipas que trabalham nesta língua e no Português falado não são de todo comparáveis, nem existe para a nossa língua, o mesmo volume de estudos prévios. Muitos dos principais avanços obtidos a nível mundial nesta área nos últimos anos, têm-se devido à existência de melhores meios computacionais e a métodos estatísticos aplicados a corpora de dimensões sucessivamente maiores. Contudo, apesar do progresso recente, não estamos de todo perante problemas resolvidos.
Em termos de reconhecimento, ainda estamos muito longe de conseguir lidar com vocabulários de grandes dimensões (p.e., superior a 500.000 palavras) e de atingir a robustez desejada face à enorme variabilidade de estilos de fala, ambientes adversos e canais de transmissão (sobretudo via telemóvel). Uma área de aplicação que conjuga todos estes desafios é o reconhecimento de notícias televisivas, onde estamos actualmente a dar os primeiros passos, num projecto internacional em colaboração com a RTP.
Em termos de síntese, temos computadores capazes de "ler", mas incapazes de "conversar", de imitar vários estilos de fala, diferentes qualidades de voz, personalidades e emoções, de interpretar formatos de texto variados (e.g. tabelas), etc.
Hoje em dia, uma parte muito significativa da investigação em processamento de fala centra-se no desenvolvimento de sistemas de diálogo falado robustos, combinando não só os avanços em síntese, reconhecimento e compreensão da fala, mas também processamento da língua natural, interfaces ergonómicas, etc. A necessidade de investirmos no desenvolvimento de sistemas multimedia multilingues é também inquestionável. Uma das aplicações que melhor reune todos estes desafios, é, por exemplo, o acesso a websites remotos através do telefone, falando numa língua que pode ser diferente da do website consultado e envolvendo assim também técnicas de tradução automática.
Torna-se difícil imaginar todas as futuras aplicações do reconhecimento e da síntese de fala e que permitirão a interacção com o computador em qualquer sítio onde esteja o utilizador. Afinal, a fala tem sobre a escrita uma grande vantagem: é ubíqua.