Capítulo 1

Introdução

A fala é o meio de comunicação humana por excelência e o que nos distingue de todos os outros seres vivos, permitindo trocarmos ideias, expressarmos opiniões ou revelarmos o nosso pensamento.

Num mundo a que hoje chamamos de *aldeia global*, a rede telefónica que suporta a transmissão de informação falada está a crescer, pelo que a representação com eficiência dos sinais de fala tem tomado cada vez maior importância. Por outro lado, embora sucessivamente mais dominados por sistemas automáticos, assiste-se à criação de interfaces desses sistemas mais cómodos para o homem, nomeadamente através da utilização da fala. O homem passa assim a não ser o único gerador ou destinatário da fala, podendo parte desta cadeia ser implementada por sistemas automáticos.

1.1 Processamento digital de fala

O objecto do processamento digital de sinais de fala é a representação digital dos sinais de fala, análise e extracção de características e o desenvolvimento de modelos de síntese. Todas estas ferramentas são cruciais na implementação de sistemas de comunicação falada, seja esta comunicação à distância ou comunicação homem-máquina. Tradicionalmente, estes sistemas estão divididos em sistemas de codificação, síntese, reconhecimento de sinais de fala e verificação e identificação do orador.

1.1.1 Codificação de fala

Uma das mais antigas aplicações do processamento digital de fala é a codificação. Por codificação entende-se a representação eficiente do sinal (tentando diminuir o débito binário na representação do sinal) com vista à sua transmissão ou armazenamento, mas mantendo a qualidade acima da exigida pela aplicação. O esforço de investigação realizado nesta área tem conduzido a diversas normas vocacionadas para trabalhar na rede telefónica pública, mas mais recentemente tem sido feito um esforço para as utilizar também em redes baseadas em protocolos transmitindo pacotes de informação.

No caso da transmissão de informação, o débito binário de codificação da fonte é um dos factores mais importantes na definição da largura de banda dos sistemas de comunicação. O armazenamento de grandes quantidades de informação para utilização posterior exige também a necessidade de reduzir o débito binário, já que este determina o espaço requerido na unidade de armazenamento. A necessidade de reduzir o débito binário mantém-se mesmo com o aumento da largura de banda dos canais de transmissão, possibilitando transmitir maior número de sinais no mesmo canal ou possibilitando

lidar com canais ruidosos que careçam de códigos de detecção e correcção de erros de grande redundância. Assiste-se ainda ao emergir de serviços multimédia com integração de voz, imagens e dados, que necessitam de racionalizar a distribuição do débito binário total por cada uma das aplicações.

1.1.2 Síntese de fala

Nos sistemas de síntese de fala a partir de texto, é gerado um sinal acústico de fala a partir de um texto fornecido pelo utilizador (eventualmente outro sistema automático). Uma das vantagens principais, senão mesmo a principal, dos sistemas de síntese de fala, prende-se com a necessidade dos sistemas com resposta automática através de voz diminuírem a quantidade de memória utilizada no armazenamento dos sinais respectivos. É ainda possível minimizar o esforço de adaptação do sistema a novas tarefas, não sendo necessário pré-gravar novas frases mas apenas inserir o texto correspondente. Uma das aplicações destes sistemas são os serviços automáticos via telefone, designadamente os de consulta a informação armazenada em bases de dados que se situem distantes dos utilizadores.

Os sistemas de síntese de fala a partir de texto são normalmente implementados dividindo o problema em dois: a conversão de texto em representação linguística (segmentos fonéticos, respectiva duração, contorno da energia e da frequência de vibração das cordas vocais); e a síntese da onda acústica a partir da representação linguística; No desenvolvimento destas duas tarefas, principalmente da primeira, são necessários conhecimentos profundos de linguística, resultando sempre sistemas dependentes da língua.

1.1.3 Reconhecimento de fala

Os sistemas de reconhecimento de fala desempenham o papel do ouvinte, convertendo uma onda acústica numa mensagem escrita ou equivalente, ou identificando um comando falado. A dimensão do vocabulário interpretado pelo reconhecedor é um parâmetro importante na sua complexidade e eficiência, podendo ser reconhecidas palavras de léxicos de pequenas dimensões (menos de 100 palavras), grandes dimensões (que podem chegar a 10000 palavras) e fala contínua, este último normalmente com constrições impostas pela tarefa a desempenhar e pela própria língua.

O reconhecimento envolve sempre uma fase de treino. Os sistemas treinados para serem utilizados com um orador específico têm um melhor desempenho que os sistemas para oradores múltiplos. Existem ainda sistemas capazes de se adaptar dinamicamente a um orador, melhorando o seu desempenho após a adaptação.

A conversão de fala para texto através do reconhecimento de fala contínua tem numerosas aplicações. É exemplo a legendagem automática em tempo real para ajuda a surdos, em programas televisivos ou mesmo no dia-a-dia. Estes sistemas podem ser previamente treinados para o locutor, melhorando o seu desempenho. Quando ligados em cadeia a um sistema de tradução e síntese de texto para fala, a conversão de fala para texto permite também a tradução automática entre línguas.

1.1.4 Verificação e identificação do orador

Um sistema de verificação do orador deve decidir se um orador é quem clama ser. Um sistema de identificação do orador deve decidir qual dos oradores, dentro de um conjunto de oradores, produziu determinada frase. Estes sistemas têm aplicações em situações em que é

necessário um controlo de acessos e a verificação de identidade, ou na ajuda de prova em alguns casos judiciais. A verificação de identidade (orador) em sistemas acedidos através da linha telefónica, como por exemplo os sistemas de *home banking*, complementa em termos de segurança a utilização de códigos numéricos digitados através do teclado do telefone, sem a introdução de *hardware* adicional.

1.2 Objectivos deste texto

As metodologias de processamento digital de sinais de fala são de grande importância num curso de Licenciatura em *Engenharia de Sistemas de Telecomunicações e Electrónica* ou em *Engenharia Informática e Computadores*. O processamento de sinais de fala específica para o Português Europeu (P.E.), como para qualquer outra língua, inibem a utilização de determinados produtos sem adaptações à língua, tais como os sistemas de reconhecimento ou os sistemas de síntese com conversão de texto para fala. Sendo o mercado português desta tecnologia reduzido, as empresas mundiais têm tendência a retrair o investimento, deixando às empresas portuguesas maior espaço de manobra nesta área tecnológica. É com base nestas considerações que é proposta a disciplina de opção de Processamento Digital de Fala, abarcando as vertentes acima mencionadas.

Este texto tem como objectivo dotar os alunos de uma introdução ao processamento digital de sinais de fala. Ao longo do texto são referenciados artigos científicos que devem ser lidos caso o leitor queira aprofundar o assunto em causa. Como livros complementares, nas áreas da codificação e síntese propõe-se o livro "*Speech Coding and Synthesis*", dos autores W. Kleijn e K. Paliwal, Elsevier, de 1995, e na área do reconhecimento o livro "*Fundamentals of Speech Recognition*", dos autores L. Rabiner e B. Juang, Prentice

Hall *Signal Processing Series*, de 1993. Embora já com alguns anos, estes livros são ainda referências fundamentais na apresentação dos conceitos básicos do processamento digital de fala.

O texto está dividido em três partes principais, para além desta introdução. A primeira parte é intitulada de *Produção e Modelação de Fala* e correspondente aos capítulos 2, 3 e 4. No capítulo 2 discute-se brevemente a produção da fala e caracterizam-se os diferentes tipos de sons, restringindo aos sons do Português Europeu. No capítulo 3 são apresentados alguns dos métodos mais comuns de análise e extracção de características de sinais de fala, dando ênfase especial ao método da predição linear. Estes métodos são essencialmente utilizados na codificação e reconhecimento, mas também na síntese. No capítulo 4 são apresentados modelos de síntese de sinais de fala, com base nas características estimadas a partir das técnicas apresentadas no capítulo anterior. Estes modelos são utilizados quer na codificação quer na síntese de sinais de fala a partir de texto.

A segunda parte, intitulada de *Codificação de Fala*, compreende os capítulos 5, 6, 7. No capítulo 5 apresentam-se os principais atributos dos codificadores e os métodos para a sua aferição, discutindo-se os principais compromissos envolvidos. O bloco correspondente ao emissor inclui normalmente um dos métodos de análise e extracção de características descritos no capítulo 3 sobre análise de fala, enquanto que o bloco correspondente ao receptor emprega um dos modelos de síntese descritos no capítulo 4 sobre modelos de síntese. Os parâmetros a ser transmitidos entre o emissor e o receptor deverão ser quantificados de um modo eficiente, pelo que as metodologias de quantificação são descritas no capítulo 6. Finalmente no capítulo 7 apresentam-se, numa perspectiva histórica, os principais métodos de codificação e respectivos codificadores normalizados.

A terceira parte, intitulada *Reconhecimento de Fala*, não está ainda redigida mas corresponderá aos capítulos 8, 9, 10 e 11. O capitulo 8 descreverá os princípios gerais do reconhecimento. Os capítulos 9 e 10 descreverão duas das metodologias mais importantes de reconhecimento, respectivamente o DTW (*Dynamic Time Warping*) e os modelos de Markov não observáveis (HMM - *Hidden Markov Models*). Sobre este último é apresentada uma versão preliminar. Por fim, o capítulo 11 descreverá as metodologias de verificação e identificação do orador.

De referir que alguns dos assuntos tratados são leccionados em disciplinas anteriores, mas são apresentados neste texto com o duplo objectivo de os rever e enquadrar nas aplicações de processamento digital de fala.

Sendo a disciplina de processamento digital de fala vocacionada para o processamento directo do sinal de fala, não descreveremos a conversão de texto em representação linguística, até porque esta faz recurso a conhecimentos profundos de fonética e linguística. Outros exemplos de sistemas que não processam directamente o sinal de fala, desviando-se do contexto da disciplina, são o controlo de diálogo em comunicação homem-máquina, a tradução de fala natural em interrogações a uma base de dados ou a conversão da resposta de uma base de dados em fala natural para ser sintetizada.

Este texto está acessível na *Internet*, na página[1] da disciplina de Processamento Digital de Fala, Secção de Comunicações e Processamento de Sinais, Departamento de Engenharia de Electrónica e Telecomunicações e de Computadores, do Instituto Superior de Engenharia de Lisboa. Aí também se encontram demonstrações sobre

[1] www.deec.isel.ipl.pt/comunicacoesep/disciplinas/pdf

aplicações do processamento de fala, algumas das quais referenciadas ao longo do texto. São também apresentados ponteiros para conferências, associações, laboratórios e jornais e indicadas páginas em que se encontra *software* livre de encargos sobre codificadores. Os leitores são encorajados a ouvir as demonstrações e a visitar as páginas referidas, de modo a terem uma melhor ideia do tipo de actividades desenvolvidas pela comunidade de processamento de fala.