# Capítulo 5

# Atributos dos Codificadores de Fala

Pode-se considerar que a barreira da distância na comunicação falada foi quebrada em 1876, quando da invenção do telefone por Alexander Graham Bell. Desde então, a importância da comunicação telefónica na sociedade não tem parado de crescer, aproximando as pessoas e tornando o mundo *mais pequeno*. A introdução da comunicação digital na década de 70 iniciou uma nova era na comunicação à distância, tendo a representação digital eficiente de sinais de fala se tornado uma área de grande importância.

Quando um sinal é transmitido, o número de bits que representa cada segundo de fala, ou seja o débito binário produzido, é um parâmetro importante na definição da largura de banda do canal de transmissão. Da mesma forma o advento das tecnologias

multimédia e a necessidade de armazenamento de grandes quantidades de informação para utilização posterior, exige a necessidade de reduzir o débito binário, já que este determina o espaço requerido na unidade de armazenamento. A redução do débito binário na representação de sinais de fala, mantendo a qualidade dos sinais sintetizados acima do exigido pela aplicação em causa, é o principal objectivo da codificação.

Nos critérios de escolha de um codificador de fala para determinada aplicação, existem alguns atributos que são decisivos, enquanto que outros ou não têm influência ou algum compromisso pode ser levado em consideração. Para além do débito binário produzido e da qualidade do sinal de saída, os atributos mais relevantes dos codificadores de fala são a complexidade dos algoritmos e a quantidade de memória necessária, o atraso introduzido e a sensibilidade a erros de canal. Seguidamente descreveremos de um modo sucinto cada uma destes atributos e abordaremos os principais compromissos envolvidos.

## 5.1 Débito binário

A primeira motivação da codificação de fala é a redução do débito binário com vista a uma transmissão ou a um armazenamento mais eficientes. A diminuição do débito binário leva desde logo a uma diminuição da qualidade, devido à necessidade de representação com menor precisão da informação a ser transmitida.

Para a denominada largura de banda telefónica, o sinal acústico é amostrado a 8000 amostras por segundo. Para que este sinal tenha uma alta qualidade e possa ser considerado como uma referência, é representado com uma resolução de 16 bits por amostra, em PCM com quantificação uniforme, resultando um débito binário de 128 kbit/s. É

este sinal, já na sua forma digital, que é processado de modo a gerar um conjunto de bits com um débito binário mais reduzido, para ser transmitido ou armazenado. Este conjunto de bits serve de entrada a um receptor, que constrói uma aproximação do sinal original em PCM e o converte posteriormente num sinal acústico.

Tradicionalmente, os codificadores de fala produzem débitos binários fixos, ideais para utilização em canais de transmissão não partilhados, já que não é necessário ter em conta critérios para definir, em cada instante, qual o débito binário atribuído. Contudo, hoje em dia, as redes de comunicação utilizam protocolos flexíveis, capazes de lidar com codificadores de débito binário variável (VBR - *Variable Bit Rate*). Um dos exemplos mais simples, de que resulta um codificador de débito binário variável, é a utilização de dois modos de funcionamento, um para zonas de silêncio ou de ruído de fundo e outro para sinais de fala [Vahatalo (99)].

Um débito binário variável pode ainda ser imposto externamente, por exemplo por sistemas de multiplexagem de vários codificadores num mesmo canal de transmissão. Para responder a um maior número de utilizadores, o sistema pede aos codificadores que entrem em modos de funcionamento de menor débito binário (codificadores multimodo), embora diminuindo a qualidade dos sinais reproduzidos, sendo no entanto sempre mantida uma qualidade mínima aceitável. Quando o tráfego é baixo, cada codificador pode funcionar com o débito máximo possível, aumentando a qualidade. Com maior flexibilidade, os codificadores embebidos [Wassell (88)] são caracterizados por ser possível extrair bits pré-determinados entre o emissor e o receptor, sem conhecimento do emissor.

## 5.2 Qualidade

A largura de banda dos sinais de entrada marca desde logo a qualidade da fala. Os sinais de fala têm uma banda perceptualmente importante até cerca dos 10 kHz, embora esta seja normalmente limitada entre os 300 e os 3400 Hz para a *banda telefónica*. Está também normalizada a banda dos 50 aos 7000 Hz, sendo o sinal amostrado a 16000 amostras por segundo, que denominaremos por *banda larga*. Em relação à banda telefónica, a diminuição para 50 Hz no inicio da banda faz aumentar a naturalidade dos sinais produzidos e a extensão das altas frequências até aos 7 kHz produz uma maior inteligibilidade, tornando-se mais fácil de distinguir entre sons fricativos, por exemplo entre o /*s*/ e /*f*/. Esta banda é utilizada em aplicações multimédia, em teleconferência e no videotelefone com transmissão baseada em Rede Digital com Integração de Serviços (RDIS).

Devido à quantificação, o sinal reconstruído não é igual ao sinal original. Nos codificadores em que uma diminuição do ruído de quantificação leva a uma aproximação entre os sinais de entrada e de saída, é comum encontrar como medida de qualidade a relação sinal-ruído de quantificação (SNR), sendo normalmente expressa em decibéis (dB). Esta medida tem o inconveniente de ser dominada pelas zonas de maior energia. Uma vez que as zonas de baixa energia são também perceptualmente relevantes, uma medida melhor obtém-se calculando a SNR em tramas de dimensão de 10 a 20 ms e, no final, determinando a média ao longo de todo o sinal de fala. Esta medida é referida como SNR *segmental*.

Quando se impõe um débito binário reduzido, os codificadores apenas preservam as propriedades mais relevantes do sinal original,

sem o reproduzir de um modo detalhado. É o caso da preservação da envolvente espectral e da estrutura fina do espectro, perceptualmente mais importantes que a definição da fase. Para este tipo de codificadores, as medidas de qualidade mais utilizadas são subjectivas e expressas em termos da distorção introduzida, da naturalidade, da inteligibilidade e da "reconhecibilidade" do orador. Quanto mais baixo for o débito binário e consequentemente maior a distorção, maior importância tomam os critérios perceptuais. No entanto, a qualidade torna-se dependente das características do sinal de entrada, sendo difícil antecipar, em todas as situações, a qualidade final do sinal sintetizado.

As considerações de carácter perceptual, nomeadamente as que se referem ao mascaramento auditivo, têm um papel relevante no desenvolvimento de codificadores. O mascaramento auditivo descreve a observação de que um estímulo acústico claramente percebido quando apresentado isoladamente, torna-se imperceptível quando apresentado na presença de outro estímulo. A principal causa do mascaramento é a resolução limitada do ouvido humano ao nível do estímulo. Isto quer dizer que dois estímulos acústicos que difiram apenas no nível, tornam-se indistinguíveis desde que a diferença de nível entre os estímulos seja menor que determinado valor. Uma descrição pormenorizada sobre mascaramento auditivo pode ser encontrada no capítulo 11 [Veldhuis (95)] do livro de [Kleijn (95)].

A qualidade perceptual aferida através de testes subjectivos é pois um processo comum para a validação de codificadores. Seguidamente descreve-se alguns dos métodos principais deste tipo de avaliação. Uma descrição pormenorizada sobre avaliação da qualidade de sinais sintetizados pode ser encontrada no capítulo 13 [Kroon (95-a)] do livro de [Kleijn (95)].

### 5.2.1 MOS

Uma das metodologias mais importantes e mais utilizadas para aferição da qualidade perceptual é a pontuação por média da opinião (MOS - *Mean Opinion Score*), recomendação P.800 do ITU-T[1] (*International Telecommunication Union - Telecommunication Standardisation Sector*). Nesta, os ouvintes são confrontados com frases processadas através do codificador em teste, sendo-lhes pedido que classifiquem a sua qualidade através de uma escala de 5 pontos (1-5), a que corresponde uma qualidade desde a *má* à *excelente* (*má*, *fraca*, *razoável*, *boa*, *excelente*[2]). Do valor médio das respostas obtém-se a classificação final em termos perceptuais. Exemplos de resultados do teste MOS para diversas normas de codificação serão apresentados nas figuras 7.1 e 7.18.

### 5.2.2 Inteligibilidade

Na avaliação da inteligibilidade, um dos testes mais comuns é o DRT (*Diagnostic Rhyme Test*) [Deller (93)], em que é presente a um ouvinte uma palavra, de um par de palavras, que variam apenas numa consoante. Estas palavras, denominadas de estímulos, têm mm Inglês normalmente uma estrutura /C-V-C/ (consoante-vogal-consoant) (*e.g.*, *veal-feel*), tendo o ouvinte que marcar qual das palavras foi entendida. O valor final do DRT é obtido através de:

$$DRT\% = \frac{N_{correctas} - N_{incorrectas}}{N_{total}} \times 100, \qquad (5.1)$$

[1] Anteriormente a 1993 denominava-se CCITT. O ITU não pode produzir formalmente normas mas apenas recomendações, que expressam um acordo entre um sector alargado da industria das telecomunicações.
[2] A escala MOS será referenciada no texto em *itálico*.

em que $N_{total}$ é o número de pares de palavras testadas, $N_{correctas}$ o número de respostas correctas e $N_{incorrectas}$ o número de respostas incorrectas. São normalmente utilizados 96 pares de palavras. Analisando os resultados deste teste, é possível conduzir os esforços de investigação no sentido de se perceber porque é que alguns dos segmentos fonéticos não são sistematicamente entendidos pelos ouvintes ou porque é que outros têm tendência a se confundirem. O DRT, sendo um teste de resposta fechada com apenas duas alternativas e limitado a palavras com sentido, produz uma taxa de inteligibilidade que pode estar sobrestimada. O MRT (*Modified Rhyme Test*) é um teste do tipo do DRT, mas em que o número de alternativas sobe (*e.g.*, 6) [EAGLES (95)], embora não sejam utilizados normalmente tantos estímulos como no DRT. Mantém-se no entanto a limitação de palavras com sentido, sendo outra limitação, comum ao DRT, a utilização exclusiva de consoantes como segmentos em teste. Repare-se contudo que são as consoantes que contribuem fundamentalmente para a perda da inteligibilidade.

O projecto SAM [SAM (92)] propõe um método de avaliação da inteligibilidade de sistemas de síntese de fala, limitado também à avaliação das consoantes, em posição inicial (/CV/), central (/VCV/) ou final (/VC/) de palavra, combinadas com 3 vogais, /i/, /u/ e /a/. Este teste é de resposta aberta, isto é, os ouvintes escolhem a resposta de entre todas as consoantes, sendo a única restrição a imposta pela própria língua. Uma das vantagens deste teste em relação ao DRT e ao MRT é a de que, sendo um teste de resposta aberta com palavras sem sentido, produz resultados comparáveis entre línguas. Naturalmente este tipo de teste pode ser facilmente estendido de modo a avaliar as vogais e pode ser utilizado para a avaliação de codificadores [Ribeiro (2000)].

### 5.2.3 Reconhecibilidade do orador

A reconhecibilidade do orador foi um dos critérios de selecção na escolha do codificador para a nova norma a 2400 bit/s [McCree (96)], pelo Consórcio de Processamento Digital de Voz do DoD [Schmidt-Nielson (95)(96)]. O teste baseia-se no julgamento de pares de frases como sendo do mesmo orador ou de oradores diferentes, utilizando 10 oradores de cada género. Foram efectuadas duas experiências. Na primeira experiência, a primeira frase do par não é processada, enquanto que a segunda é processada por cada codificador em teste. A segunda frase é sempre processada. Esta experiência testa a capacidade que o codificador tem de preservar as características originais do orador. Na segunda experiência, ambas as frases do par são processadas pelo codificador em teste. Esta experiência testa a capacidade que o codificador tem de captar características suficientes para se distinguir entre diferentes oradores, mesmo que a voz produzida seja alterada em relação à voz original.

Em ambas as experiências, um conjunto de ouvintes avaliam os pares de frases produzidas pelo mesmo orador ou por oradores diferentes. Após ouvir cada par, o ouvinte é inquirido sobre a sua opinião quanto a ser ou não ser o mesmo orador. Seguidamente, pontuam o par em relação à distância entre as duas vozes, numa escala de 5 pontos. Na primeira experiência, os melhores codificadores têm os valores mais baixos de distância entre oradores para os pares de frases realmente do mesmo orador. Na segunda experiência, os melhores codificadores tem valores mais elevados de distância entre oradores para os pares de frases realmente de oradores diferentes. Uma adaptação desta metodologia, juntamente com testes de inteligibilidade, foi efectuada por Ribeiro e Trancoso [Ribeiro (2000)] para aferir um codificador fonético para o Português Europeu.

## 5.3 Atraso

O atraso em codificação de fala é definido como o tempo máximo que medeia entre o instante em que uma amostra é apresentada ao emissor e aquele em que a amostra correspondente é gerada pelo receptor. Este tempo é medido estando o receptor ligado directamente ao emissor, retirando portanto a contribuição dos equipamentos de emissão e recepção e o tempo de propagação do sinal, mas não o tempo de transmissão de cada bit.

Embora o atraso não seja importante em aplicações de armazenamento, na conversação bidireccional, como por exemplo na comunicação telefónica, o atraso pode tornar-se maçador e mesmo afectar a naturalidade da conversação. Limites para este atraso poderão ir, nos casos mais permissivos, até cerca de 400 ms. Restrições mais severas são aplicadas quando as redes de comunicações não incluem canceladores de eco, pois o atraso é notado pelo próprio orador.

Como é ilustrado na figura 5.1, o atraso de codificação pode ser dividido em três parcelas: (1) tempo de espera para que todas as amostras de uma trama estejam presentes, a que corresponde naturalmente uma trama de atraso; (2) tempo de processamento da trama, que para aplicações em tempo real é limitado a uma trama; (3) tempo de transmissão dos bits correspondentes a uma trama, que assumindo o débito binário mais baixo possível, corresponde também a uma trama. O receptor, normalmente com menor complexidade que o emissor, deverá ser capaz de descodificar imediatamente as primeiras amostras de cada trama.

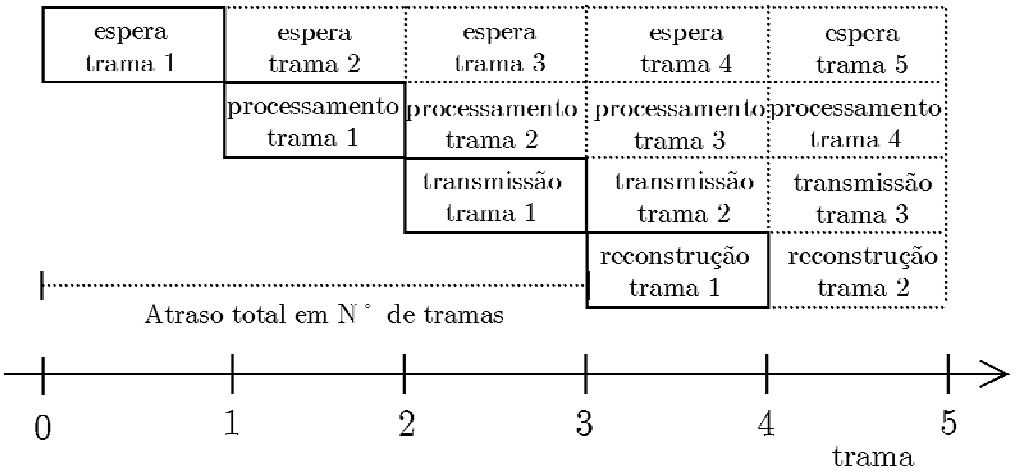


Figura 5.1
Diagrama temporal dos diversos blocos responsáveis pelo atraso num codificador.

O atraso total corresponde a três tramas, impondo apenas estas restrições. No entanto, muitos codificadores necessitam de informação posterior (*look ahead*) para poderem processar uma trama (*e.g.*, janela de análise LPC ou de *F*0), pelo que este valor deve ser interpretado como um valor mínimo. Ao atraso conjunto da dimensão da trama e da informação posterior necessária denomina-se de atraso do algoritmo. Para muito baixo débito binário os codificadores também tiram partido da correlação inter-tramas originada pela variação lenta do tracto vocal, tornando o atraso ainda maior.

## 5.4 Complexidade e memória necessária

Quanto maior complexidade apresentar o algoritmo de codificação e maior quantidade de memória for necessária, mais os sistemas serão dispendiosos, volumosos e com maior consumo de energia. O primeiro codificador a ser normalizado, a recomendação G.711 do ITU-T a 64 kbit/s, que data de 1972, era então implementado directamente no conversor analógico-digital, devido à sua baixa complexidade e à ausência de necessidade de memória. Com

a vulgarização dos processadores digitais de sinal (DSP - *Digital Signal Processor*) e o aumento da complexidade dos codificadores, os sinais passaram a ser quantificados uniformemente, tipicamente com 12 a 16 bits por amostra, e só depois codificados a débitos binários mais baixos. A complexidade é normalmente aferida através do número de MIPS (milhões de instruções por segundo) ou MFLOPS (milhões de operações em virgula flutuante por segundo) necessários para processar os algoritmos de codificação, enquanto que a memória necessária é medida em número de *bytes*.

Devido ao seu baixo preço e consumo, os DSP de vírgula fixa de 16 bits são uma boa plataforma para a implementação de codificadores. Como principal desvantagem apresentam uma maior dificuldade de programação em relação aos de vírgula flutuante. Os DSP de vírgula flutuante de 32 bits são uma alternativa em relação aos de vírgula fixa, mas o consumo mais elevado poderá inibir a sua utilização em sistemas móveis, em que a autonomia é um critério importante de projecto. A utilização destes processadores pode no entanto ser rentabilizada se o processador for partilhado com outros módulos do sistema, como por exemplo por filtros, *modems* ou sistemas multimédia. Todas estas preocupações têm vindo a ser postas em causa com os avanços da microelectrónica e a disseminação destes produtos, com a consequente baixa dos preços.

## 5.5 Sensibilidade a erros de canal

Na transmissão do sinal codificado, este fica sujeito a erros introduzidos pelo canal, que podem ser de dois tipos: erros aleatórios independentes, causados pelo ruído estacionário, e erros em rajada, limitados temporalmente, causados por interferências electromagnéticas nas imediações do canal. Os codificadores devem ter

metodologias para lidar com ambos os tipos de erro, de modo a que a qualidade do sinal de saída seja afectada o mínimo possível. Estas metodologias envolvem a introdução de códigos de codificação de canal com capacidade de detecção e correcção de erros, aumentando o débito binário total.

Em relação à detecção e correcção de erros aleatórios independentes, os parâmetros transmitidos podem ser classificados em função do impacto que os erros respectivos provocam na qualidade final, sendo este critério usado para adaptar a robustez dos códigos utilizados. Um outro método de minimizar os erros aleatórios é o da utilização de algoritmos de atribuição de índice (IA - *Index Assignment*). Assumindo que a probabilidade de que sejam afectados dois bits de um mesmo parâmetro é muito baixa, estes algoritmos atribuem às palavras de código com menor distância (*e.g.* euclidiana), índices com menor distância Hamming. Caso haja o erro de um bit na recepção do código do índice, a palavra de código descodificada não é radicalmente diferente da palavra de código correcta.

O número de bits atribuídos à detecção e correcção pode variar temporalmente, em função das flutuações nas condições do canal, sendo necessário adaptar o débito binário do codificador, para que o débito binário total não ultrapasse o máximo permitido no canal. Nesta situação, devem-se utilizar codificadores embebidos ou codificadores multimodo. O sistema multi-débito adaptativo (AMR - *Adaptive Multi-Rate*) é um exemplo de um codificador multimodo, que pode comutar entre 8 modos, a que correspondem 8 débitos binários entre os 4,75 kbit/s e os 12,2 kbit/s, estando o débito binário continuamente a ser adaptado às condições do canal de rádio [Ekudden (99)]. Este sistema foi seleccionado em Outubro de 1998

para funcionar na rede Europeia de telefones móveis digitais (GSM[3] - *Global System for Mobile Communications*).

Em relação aos erros em rajada, extremamente nocivos pois podem afectar seriamente o sinal de saída, os códigos de detecção de erros permitem aceitar ou rejeitar completamente uma trama. Caso uma trama seja rejeitada, o receptor substitui-a pela anterior, eventualmente modificada de modo a que o sinal desvaneça em potência e as características do sinal convirjam para ruído branco. Uma estratégia complementar é a de alterar a sequência de bits que é entregue ao canal (*scrambling*), de modo a que os bits afectados em rajada não pertençam a um mesmo parâmetro, sendo nomeadamente possível corrigir alguns destes erros através dos códigos de codificação de canal. Uma outra vantagem desta técnica é a de tornar a sequência de bits pseudo aleatória, melhorando a resposta em frequência do sistema de modulação digital utilizado.

## 5.6 Conclusões

Neste capítulo apresentámos os atributos principais dos codificadores de fala e os métodos mais comuns para a sua aferição. O débito binário é medido em bits por segundo. A qualidade dos sinais sintetizados pode ser medida objectivamente através da SNR e da SNR segmental, ou subjectivamente através do MOS, testes de inteligibilidade tal como o DRT e testes de reconhecibilidade do orador. O atraso é medido como o tempo máximo que decorre entre a apresentação de uma amostra à entrada do emissor e o tempo que essa amostra é gerada no receptor. A complexidade é medida em MIPS ou MFLOPS e a memória necessária em número de *bytes*. Finalmente a

[3] Inicialmente era acrónimo de *Groupe Speciale Mobile*

sensibilidade na presença de erros de canal é medida função da degradação na qualidade do sinal sintetizado.

O *melhor codificador* é um conceito inexistente em codificação de fala. Para determinada aplicação devem ser levados em conta os atributos mais relevantes, muito possivelmente à custa de um pior desempenho em relação aos outros. Por exemplo, a diminuição do débito binário só é possível correlacionando a informação entre amostras, o que leva a um aumento da complexidade e atraso, e possivelmente à diminuição da qualidade. A complexidade pode ser diminuída utilizando algoritmos subóptimos, mas estes levam a uma degradação da qualidade. A diminuição do atraso pode ser conseguida diminuindo o tamanho das tramas de análise, mas o consequente aumento de número de tramas por segundo leva ao aumento do débito binário.