

UNIVERSIDADE DO PORTO
FACULDADE DE ENGENHARIA

1.45

Departamento de Engenharia Electrotécnica e de Computadores

**Estudo e Desenvolvimento de Técnicas e Ferramentas de Análise de
Fala. Estudo e Integração de Técnicas de recolha de sinais de
Compressão e de Codificação da Voz.**

(*Relatório de Estágio* PRODEP)

Orlando José Lourenço da Costa

Porto, 03 de Dezembro de 1993



4

621.3(047.2)/LCC 1972/0050
02 10 09

Porto, Dezembro de 1993

PRODEP

MEDIDA 4.3 - Estágios Profissionais para Bacharelados, Licenciaturas e Pós-Graduação

PARECER

Assunto: Estágio de (1.45) Orlando José Lourenço da Costa

Na qualidade de supervisores do estágio acima mencionado e em face do trabalho desenvolvido e relatado pelos estagiário pode concluir-se que:

1- O estagiário manifestou entusiasmo e dedicação na execução do trabalho que foi proposto com o título:

"Estudo e desenvolvimento de técnicas e ferramentas de análise da fala . Estudo e integração de técnicas de recolha de sinais, de Compressão e de Codificação da voz".

O trabalho consistiu fundamentalmente no estudo dos conceitos relacionados pelo tema. Particular ênfase foi dada à codificação sinusoidal da fala, tendo o estudo sido baseado em literatura relevante, de várias origens.

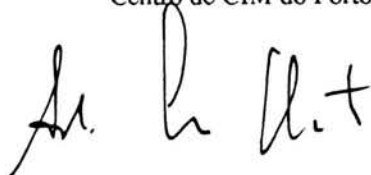
Nesta execução revelou apreciáveis qualidades de trabalho que mais desenvolveu. Do trabalho resultou um relatório de estudo bem elaborado e demonstrando uma apreciável aquisição de conhecimentos.

2- No cumprimento dos objectivos propostos, o estagiário revelou, para a fase curricular, bons conhecimentos científicos e técnicos, capacidade de aplicação desses conhecimentos e espírito de iniciativa. A metodologia seguida consistiu em pesquisa bibliográfica, definição dos grandes capítulos do estágio e estudo do material recolhido com perspectivação das técnicas associadas à análise bem como à codificação e compressão dos sinais de fala.

pelo que somos do parecer que foram cumpridos os objectivos que o estágio se propunha atingir.

Prof. Doutor António Maria da Rocha Quintas

Centro de CIM do Porto



Prof. Doutor Diamantino Rui da Silva Freitas

Faculdade de Engenharia da Universidade do Porto



1 - A COMUNICAÇÃO FALADA E O DESENVOLVIMENTO DA FALA	2
1.1 - MODELIZAÇÃO SINUSOIDAL DA FALA	3
2 - COMUNICAÇÃO FALADA	4
2.1 - PRODUÇÃO DA FALA	5
2.2 - CARACTERÍSTICAS ACÚSTICAS	8
3 - MODELIZAÇÃO ESPECTRAL	14
3.1 - MODELIZAÇÃO LOCAL	15
3.2 - MEDIDAS DE DISTORÇÃO	16
3.3 - MODELO HARMÓNICO GENERALIZADO	16
3.4 - CONCLUSÕES PARCIAIS RELATIVAS A ESTES ASSUNTOS	17
3.5 - ESTIMAÇÃO ESPECTRAL	17
3.6 - CONCLUSÕES PARCELAES DA RELAÇÃO COM O MÉTODO DE GAUSS-SEIDEL	18
3.7 - INTRODUÇÃO DA ESCALA DE TEMPO	19
4 - MODELIZAÇÃO SINUSOIDAL NO DOMÍNIO DO TEMPO	19
4.1 - MODELOS SINUSÓIDAIIS DA FALA	20
4.2 - ESTIMAÇÃO DE AMPLITUDES E FASES	20
4.3 - ESTIMAÇÃO POR AMOSTRAGEM ESPECTRAL	21
4.4 - MÉTODOS DE MÍNIMOS QUADRADOS ESTACIONÁRIO	22
4.5 - ESTIMAÇÃO QUASE-ÓPTIMA DE AMPLITUDES	22
4.6 - MÉTODO QUASE-ÓPTIMO MODIFICADO	22
4.7 - ESTIMAÇÃO DE SINAIS CONSTANTES	23
5 - REPRESENTAÇÃO DA FALA POR FUNÇÕES DE BANDA ESTREITA	23
5.1 - MODELIZAÇÃO DE FALA NÃO-VOCALIZADA COM FUNÇÕES DE BANDA ESTREITA	24
5.2 - CODIFICAÇÃO DE FALA NÃO VOCALIZADA COM FUNÇÕES DE BANDA ESTREITA	26
6 - CODIFICAÇÃO SINUSOIDAL	27
6.1 - CODIFICAÇÃO DE FASE	29
6.2 - ESTRUTURA DA FASE EM ZONAS NÃO-VOCALIZADAS	29
6.3 - DESENHO E ESTUDO DE UM CODIFICADOR DE 4.8 KB/S	30
6.3.1 - ESTRUTURA BÁSICA	30
6.3.2 - CODIFICAÇÃO DO MODELO SINUSOIDAL A RITMOS MÉDIOS E BAIXOS	31
6.3.3 - DESCRIÇÃO DO CODIFICADOR DE 4.8 KB/S	35
6.3.4 - COMPRIMENTO DO INTERVALO DE SÍNTESE	35
6.3.5 - CÁLCULO E CODIFICAÇÃO DA FREQUÊNCIA FUNDAMENTAL	35
6.3.6 - CÁLCULO DAS FREQUÊNCIAS DAS SINUSÓIDES	35
6.3.7 - CÁLCULO E CODIFICAÇÃO DE UM FILTRO DE LPC	36
6.3.8 - CÁLCULO DAS AMPLITUDES A PARTIR DA ENVOLVENTE ESPECTRAL	37
6.3.9 - ESTIMAÇÃO DE FASES	38
6.3.10 - SÍNTESE SINUSOIDAL	38
6.3.11 - ATRIBUIÇÃO DE BITS DO CODIFICADOR SINUSOIDAL	38
7 - CONCLUSÕES E RESUMO GERAL DO TRABALHO	39
8 - PARA FUTURO	46

1 - A COMUNICAÇÃO FALADA E O DESENVOLVIMENTO DA FALA

O desenvolvimento ocorrido nos últimos tempos, levou a uma necessária evolução nas telecomunicações e consequentemente numa área que lhe está por necessidade directamente adstrita, a do processamento da fala.

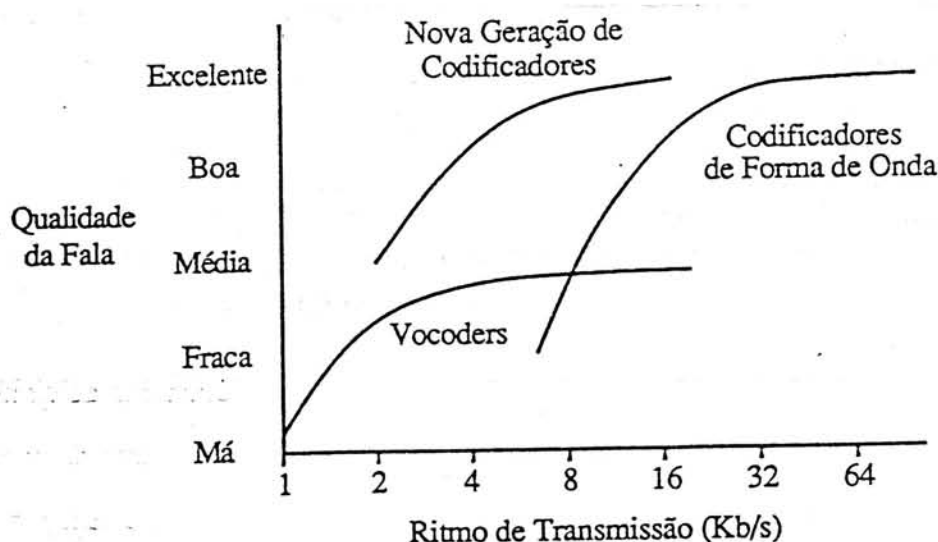
Por outro lado a evolução informática no campo da algoritmia e necessidades até de terapêutica de certas doenças levaram a uma investigação intensa no âmbito da comunicação por fala. A modelização e codificação e a correcção de erros, assumem, por tal, aspecto preponderante no estudo do desenvolvimento nesta área e que agora pretendo transcrever para este relatório.

Por modelização de fala, pode-se entender o estudo de representações de sinais de fala de forma a pôr em evidencia características importantes para diversos tipos de processamento.

A caracterização de um sinal de fala, ou da fala em termos gerais, não se restringe à sua caracterização apenas em termos acústicos, mas também de forma precisa relativamente à sua emissão ou produção e recepção.

Para nós informáticos por necessidade, a codificação de um sinal de fala, numa sequência de bits, seja por ser enviada por um sistema de comunicação ou até para tratamento digital, assume para nós uma importância também fundamental, mais que não seja, pela componente de utilização prática que esta ideia incute. Para este processo utilizam-se os chamados codificadores (ex: codificadores de forma de onda e codificadores de fonte), que se devem adaptar às inúmeras restrições existentes neste processo de codificação, mais tarde falaremos nestes tipos e explicitaremos as suas características.

De seguida, exhibe-se um gráfico que apresenta as potencialidades de desempenho dos diversos codificadores.



De notar que o conceito de qualidade de fala, usado no gráfico em escala gradativa, deve ser entendido como a qualidade de comentário, com fala com largura de banda igual ou superior a 7Khz e sem distorções audíveis, (ex: fala usada em radiodifusão) o que se entende por uma questão de uniformidade. Qualidade telefónica, aplica-se a fala com largura de banda tipicamente de transmissão telefónica (0-3400Hz) e relação de sinal-ruído superior a 30 dB no intervalo de 200-300 Hz.

por vezes é também usada a designação qualidade alta e qualidade muito alta, para fala com pequenas distorções que não afectem a inteligibilidade nem naturalidade.

1.1 - MODELIZAÇÃO SINUSOIDAL DA FALA

A modelização sinusoidal de um sinal de fala, pode ser fruto de diversas interpretações, uma delas pode ser sinusoidal, que consiste em representar um sinal de fala, por intermédio da sobreposição de sinusóides com amplitudes e frequências variáveis ao longo do tempo. A origem desta técnica é de certa forma antiga e pretende-se por exemplo com a análise de Fourier e a teoria da estimação.

Existem várias motivações principais para a utilização de modelos, mas duas são de maior importância.

No modelização sinusoidal, o sinal de fala é decomposto na soma de sinusóides com frequências distintas que variam ao longo do tempo, estando o espectro centrado em torno da frequência média da sinusóide nesse intervalo, logo para cada intervalo o modelo sinusoidal pode ser entendido como uma representação do sinal de fala no domínio da frequência, já que produz uma decomposição do sinal em bandas de frequências distintas; *a modelização sinusoidal, goza assim em geral das vantagens da modelização no domínio da frequência, permitindo nomeadamente, um controlo fácil de distribuição em frequência do erro de modelização (noise shaping).*

Uma outra motivação reside na existência de zonas extensas do sinal de fala produzidas por vibração das cordas vocais, as chamadas zonas vocalizadas, que apresentam uma estrutura quase-periódica, estando a modelização sinusoidal, especialmente vocacionada para a modelização deste tipo de sons.

De notar, que se idealmente conseguíssemos ter um sinal de fala vocalizada periódico, este poderia ser representado, de uma forma exacta, através de um modelo sinusoidal estacionário, ou seja, pela sobreposição de sinusóides com amplitudes e frequências constantes ao longo do tempo; ou seja, uma verdadeira série de Fourier.

De notar também que embora o sinal de fala vocalizada não seja periódico, a sua estrutura altera-se habitualmente de forma suficientemente lenta para que possa

ainda ser bem aproximado pela soma de sinusóides, admitindo que as amplitudes e frequências variam lentamente ao longo do tempo.

O objectivo principal da modelização sinusoidal, é a representação de cada risca espectral por meio de uma sinusóide, admitindo que as amplitudes e frequências variam lentamente ao longo do tempo.

Notemos também que existe uma necessidade particular de enfatizar a modelização das zonas vocalizadas do sinal de fala, o que se relaciona com o facto de ser necessária uma boa representação dessas zonas, tendo em vista a obtenção de uma melhor qualidade ao nível da fala sintética, dada a conexão existente com uma particular sensibilidade do ouvido humano a distorções de forma de onda em zonas quase periódicas do sinal de fala.

Nas zonas não vocalizadas e de transição, o ouvido é menos sensível a distorções na forma de onda, estando as funções de base sinusoidais, menos adaptadas à modelização deste tipo de sons.

2 - COMUNICAÇÃO FALADA

Qualquer trabalho que se relacione com a fala, deve referir obrigatoriamente os conceitos físicos e humanos ligados à comunicação falada, por isso mesmo, convém referir alguns conceitos relativos a esse aspecto.

O sinal de fala é um elo de uma cadeia complexa cujos principais constituintes são, o aparelho fonador (fonte), o canal ou meio físico por onde o sinal de propaga e os mecanismos de audição e percepção, o receptor.

Logo, em termos de análise de fala, tal pode ser feito ao nível do aparelho fonador, retirando partido das restrições existentes nos movimentos dos articuladores (língua, maxilar, véu palatino); ao nível do canal, através da descrição das características acústicas do sinal de fala; ao nível do sistema acústico, usando modelos do sistema auditivo periférico. O nosso alcance estende-se apenas, até ao sistema auditivo periférico, uma vez que o relacionamento da informação com o cérebro, ao longo das suas várias etapas, ainda não é conhecido, nem assume para nós uma importância fundamental.

Apesar de um aperfeiçoamento dos mecanismos de estudo da fala, passar por uma compreensão destes diferentes aspectos, na realidade em termos práticos, aquilo que se usa como objecto de estudo, é apenas a representação directa do sinal.

Apesar de as considerações seguintes, correrem o risco de poderem ser enfatizadas por um certo tom de simplicidade, a sua inclusão, prende-se com a necessidade de tornar este relatório, tanto quanto possível, acessível e plural .

2.1 - PRODUÇÃO DA FALA

A fala é produzida, pela passagem do ar expelido pelos pulmões através do aparelho fonador, o qual é constituído pela laringe, onde estão localizadas as cordas vocais e por um conjunto de cavidades designadas por tracto vocal. O fluxo de ar expelido pelos pulmões, dá origem a uma onda acústica que se propaga ao longo do tracto vocal e é radicada pela boca.

Na produção do sinal de fala, são normalmente considerados três aspectos:

- i) A excitação acústica do tracto vocal,
- ii) A propagação da onda acústica até à boca,
- ii) O efeito de radiação da boca.

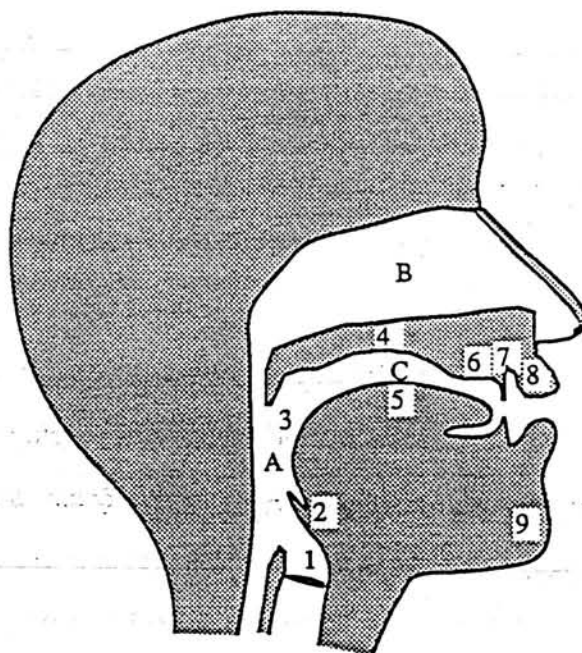
A excitação do tracto vocal é realizada tipicamente de três modos distintos:

- i) Por vibração das cordas vocais, interrompendo periodicamente o fluxo de ar que passa da traqueia para a faringe através da glote.
- ii) Por uma constrição do tracto vocal, provocando uma passagem turbulenta de ar no ponto de constrição, ou ainda
- iii) Por obstrução completa da passagem de ar, seguida de uma libertação repentina da pressão acumulada durante o tempo em que o tracto vocal esteve fechado.

Estes modos de excitação, vão corresponder a três modos fundamentais de sons com características próprias, *os sons vocalizados, fricativos e oclusivos*. Contudo, a excitação do tracto vocal, pode ser combinada em qualquer dos outros dois tipos de excitação, dando origem dessa forma a sons de excitação mista, designados habitualmente por: *fricativos, vozeados e e sons oclusivos vocalizados*.

Todos os sons produzidos sem vibração das cordas vocais, são designados por

não-vocalizados.



Secção transversal do tracto vocal: 1- cordas vocais;

2- epiglote; 3- véu palatino; 4- palato; 5- língua; 6- zona alveolar; 7- dentes;

8- lábios; 9- maxilar inferior; A- faringe; B- cavidade nasal; C- cavidade oral;

Os sons vocalizados ao serem produzidos por vibração das cordas vocais, através da abertura e fecho da glote, são usualmente quase-periódicos, desde que a excitação global persista de forma quase estacionária ao longo do tempo. A frequência de vibração das cordas vocais, não permanece invariante no tempo, nem o tracto vocal assume sempre a mesma configuração, no entanto, estas variações ocorrem de forma lenta quando comparadas com o período de vibração das cordas vocais, motivo pelo qual a fala produzida tem estrutura quase-periódica. Nos sons vocalizados a frequência de vibração das cordas vocais, designada frequência fundamental ou frequência de Pitch e o intervalo entre impulsos da excitação do tracto vocal, têm características diferentes. Como exemplo, os sons fricativos, a passagem turbulenta de

ar através da constrição funciona como uma fonte de ruído de banda larga que excita o tracto vocal no ponto de constrição. Nos sons oclusivos, o tracto vocal é excitado de forma abrupta por um impulso de pressão que produz uma resposta transitória.

O tracto vocal é habitualmente modelizado como um tubo acústico de secção não uniforme com frequências de ressonância dependentes das posições dos articuladores (língua, lábios, véu palatino, maxilar, etc).

Ao propagar-se ao longo do tracto vocal, a onda acústica sofre uma operação de filtragem por um filtro linear e variante no tempo, que pode ser caracterizado através da sua resposta impulsional $\eta_T(t, \tau)$ (ver figura seguinte).

A estimação do filtro do tracto vocal pode ser feita por integração das equações de propagação ou por análise do sinal de fala. A este modelo de produção da fala, em que se separa a excitação acústica da operação de filtragem realizada pelo tracto vocal, alguns autores chamam modelo de engenharia. As frequências de ressonância do filtro do tracto vocal, são usualmente designadas de *Formantes*.



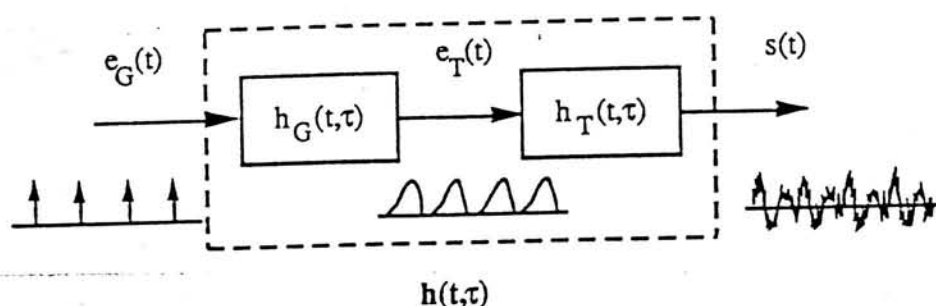
Modelo de engenharia de produção da fala

Em zonas vocalizadas, a excitação do filtro, do tracto vocal é constituída por uma sequência de impulsos glotais com início nos instantes de abertura da glote. A forma dos impulsos, depende de vários factores como sejam: a pressão sub-glotal, massa das cordas vocais, a elasticidade, etc; factores que variam ao longo do tempo.

A excitação glotal, é por isso, modelizada frequentemente através de um filtro linear variante no tempo, excitado por um trem de impulsos de Dirac localizados nos instantes de abertura da glote.

A acção de filtros que modelizam os impulsos da glote e o tracto vocal pode ser combinada num único filtro, tendo como entrada o trem de impulsos de Dirac e como saída o sinal de fala vocalizada, (ver figura seguinte).

Um caso particular desta formulação, é o modelo clássico de produção de fala, usado em Vocoder, em que a excitação é modelizada por um trem periódico de impulsos, em segmentos com tipicamente 20 ms, sendo o filtro estimado a partir do sinal de fala, usando técnicas de predição linear ou de análise cepstral.



Modelo de engenharia para fala vozeada

2.2 - CARACTERÍSTICAS ACÚSTICAS

A estrutura quase-periódica dos sons vocalizados é habitualmente caracterizada pelo período fundamental ou pelo seu inverso, a frequência fundamental. No entanto, nem sempre é fácil estimar o período fundamental de um som vocalizado. Pois é de notar que a estimação do período fundamental a partir da função de autocorrelação pode conduzir a estimativas erradas.

Para além das situações em que há dificuldade na estimação do período fundamental do sinal de fala, mas em que existe uma periodicidade bem definida, há casos mais complexos em que se torna difícil decidir qual o período fundamental, ou até, se faz sentido defini-lo.

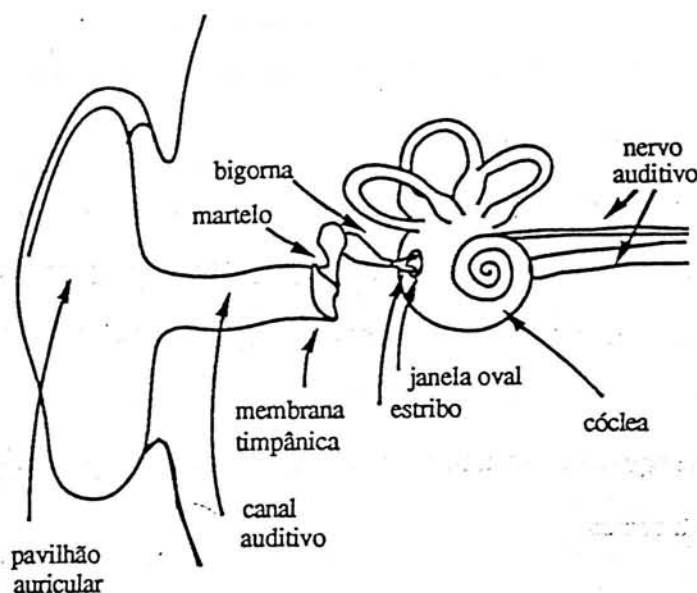
2.3 - SISTEMA AUDITIVO PERIFÉRICO

O sistema auditivo periférico é constituído por órgãos de audição utilizados para converter o sinal acústico em impulsos eléctricos que são enviados para o cérebro através do sistema nervoso (ver a figura seguinte).

Habitualmente, o sistema auditivo periférico é dividido em três regiões:

- i) Ouvido externo, constituído pelo pavilhão auricular e pelo canal auditivo;
- ii) Ouvido médio, constituído pela membrana timpânica e pelos ossículos; e
- iii) Ouvido interno, constituído pela cóclea.

Grande parte das transformações operadas no estímulo acústico pelo sistema auditivo são realizadas ao nível do ouvido interno pelo que o seu estudo tem conhecido nas últimas décadas avanços muito importantes.



Constituição do sistema auditivo periférico

No que respeita ao ouvido externo, a sua principal função parece estar relacionada com a localização de fontes acústicas funcionando também como filtro passa-banda de baixa selectividade tendo como entrada a pressão do sinal acústico que incide no pavilhão auricular e como saída, a pressão exercida na membrana timpânica. A característica passa-banda, é em parte explicada pelo funcionamento do canal auditivo, que actua como um ressoador de $1/4$ de comprimento de onda, tendo uma primeira ressonância perto de 2,5 KHz. No entanto, o pavilhão auricular tem também uma contribuição importante aumentando consideravelmente a largura de banda do filtro.

Embora a característica passa-banda dependa da localização da fonte acústica, produz em geral um reforço da gama de frequências 2-5 KHz, importante na percepção de alguns sons.

O ouvido médio actua como adaptador de impedâncias entre o ouvido externo e a cóclea. Como a propagação do som no ouvido externo, é feita através do ar e como a cóclea é preenchida por fluidos com impedância acústica muito superior, só uma pequena parcela da energia do sinal de entrada seria transferida para a cóclea se a cóclea fosse excitada directamente pela onda acústica do canal auditivo. A ligação entre o ouvido externo e interno é realizada, no ouvido médio, pela membrana timpânica e por três ossículos (martelo, bigorna e estribo), que comunicam o movimento de vibração da membrana timpânica a uma membrana da cóclea, designada por janela oval. Este processo é acompanhado por um aumento de pressão (30 dB) que é justificado principalmente pela diferença entre as áreas da membrana timpânica e da extremidade do estribo, mas também por um "efeito de alavanca" na ligação dos ossículos e pela configuração cónica da membrana timpânica. O ouvido médio tem também a função de proteger o ouvido interno contra sons demasiado intensos, através da contracção dos músculos ligados aos ossículos e da transformação do movimento transversal dos ossículos num movimento de rotação, reduzindo em ambos os casos a eficiência da ligação entre o estímulo acústico e a cóclea.

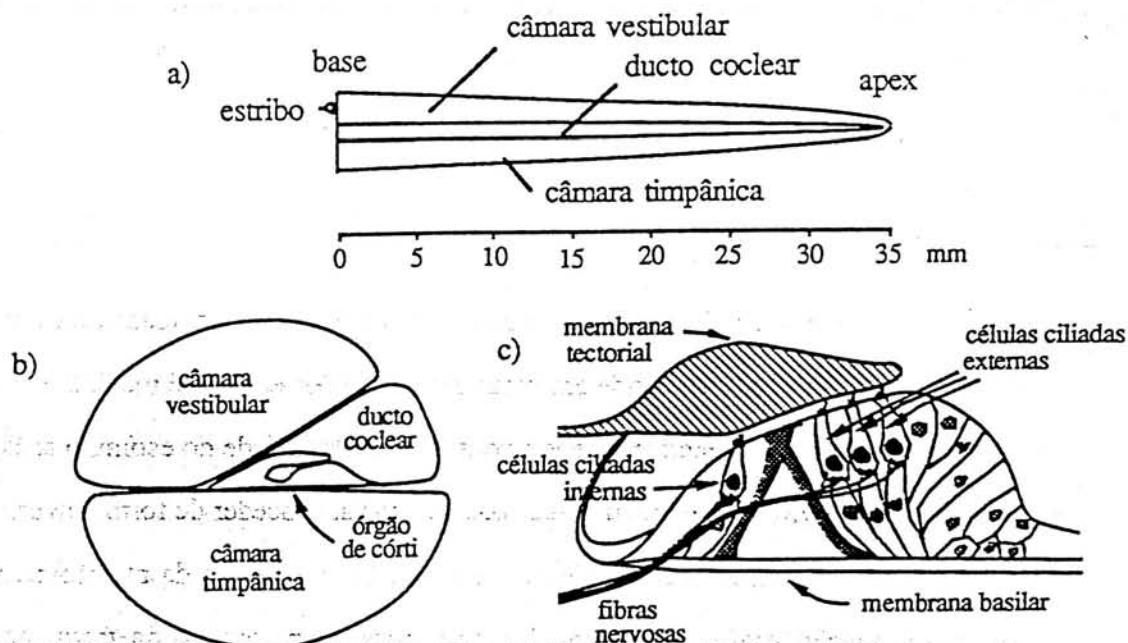
A maioria das transformações de interesse para a compreensão do mecanismo auditivo ocorrem ao nível do ouvido interno. É aí que as vibrações mecânicas transmitidas pelo ouvido médio sofrem uma espécie de transformação em frequência e são convertidas em impulsos, enviados para a cérebro através do nervo auditivo.

A cóclea é um órgão em forma de caracol dividido em três câmaras (ver figura seguinte); duas maiores: câmara vestibular e câmara timpânica, e uma menor: ducto coclear. A separar estas três câmaras estão duas membranas: a membrana basilar que separa a câmara timpânica do ducto coclear e a membrana de Reissner que separa o

ducto coclear da câmara vestibular. Ligado à membrana basilar encontra-se o órgão de CORTI, estrutura complexa composta por células ciliadas, por células de suporte e por terminações do nervo auditivo, sendo coberto por uma membrana gelatinosa designada membrana tectorial (ver figura seguinte).

O órgão de Corti, é composto por dois tipos de células ciliadas com funções bem diferenciadas: as células ciliadas internas distribuídas em linha ao longo da cóclea e as células ciliadas externas, em número maior, organizadas em três linhas.

A importância do órgão de Corti reside no facto de protagonizar a transdução de movimentos de vibração da membrana basilar em impulsos eléctricos das fibras do nervo auditivo. transdução que é realizada, tanto quanto se sabe pelas células ciliadas internas.



Estrutura anatómica da cóclea: a) cóclea desenrolada;

b) secção transversal; c) órgão de Corti (adaptado de [1])

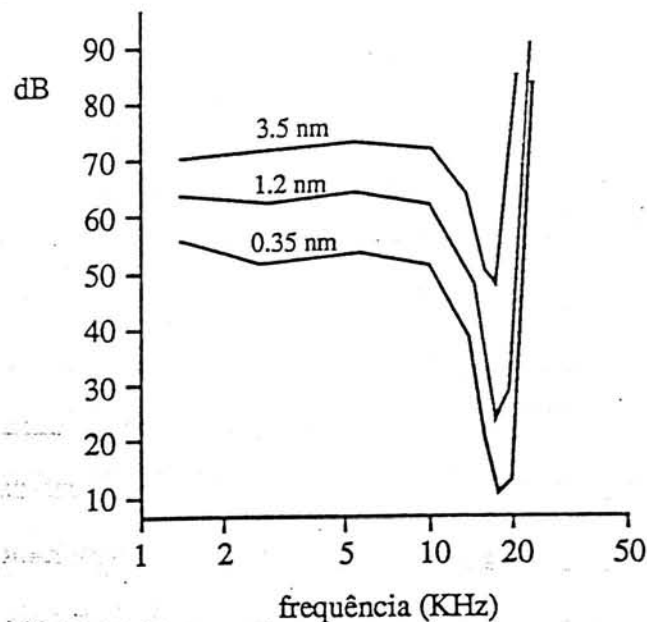
A cóclea está ligada através da janela oval à extremidade do estribo. Por isso, a vibração dos ossículos do ouvido médio produz um movimento de vibração nas membranas cocleares e, em particular, um deslocamento da membrana basilar que se propaga ao longo da cóclea desde a base até ao apex (ver figura anterior). Este comportamento foi observado experimentalmente pela primeira vez por Von Békésy nos anos 20, tendo sido posteriormente estudado com técnicas mais sofisticadas. Verificou-se, usando estímulos sinusoidais, que para cada frequência do sinal de entrada existe um ponto da membrana basilar em que a amplitude da vibração é máxima, sendo a distância desse ponto ao apex aproximadamente proporcional ao logaritmo da frequência. Com estímulos mais complexos, diferentes componentes de frequência são distribuídas por diferentes localizações na membrana basilar o que sugere a interpretação do movimento da membrana basilar como uma transformação em frequência do sinal acústico. Noutras experiências, mediu-se a amplitude de vibração da membrana basilar em diferentes pontos da membrana em função da frequência do sinal de entrada. Obtiveram-se características passa-banda em que a frequência do máximo depende do ponto da membrana basilar considerado.

Pode-se assim considerar a resposta de cada ponto da membrana basilar como a saída de um filtro passa-banda sintonizado para uma frequência que depende da localização do ponto.

Como o movimento da membrana basilar não depende linearmente da intensidade do estímulo acústico, a função de transferência depende em geral das condições em que é medida. Em vez de se calcular a resposta em frequência mantendo constante a intensidade do estímulo acústico e medindo a amplitude da vibração da membrana basilar, é usual proceder de forma inversa, ou seja, medir a intensidade da entrada capaz de produzir um deslocamento da membrana pré-fixado. Na figura seguinte, mostra-se a representação gráfica da intensidade da entrada em função da frequência, o que é designado de curva de sintonia.

As curvas podem apresentar elevada selectividade que varia contudo, em função da amplitude do deslocamento da membrana basilar (não linearidade). É também muito característico o carácter assimétrico das curvas de sintonia que apresentam declives mais abruptos acima da frequência do mínimo (frequência de sintonia). A comparação entre curvas de sintonia obtidas em diferentes pontos da membrana basilar permite concluir que a banda varia de acordo com o ponto considerado obtendo-se curvas com factor de qualidade aproximadamente constante para frequências de sintonia acima de 1Khz. Isto faz com que se utilizem, por vezes,

escalas não lineares de frequência onde as larguras de banda das curvas de sintonia permanecem aproximadamente constantes.



Curvas de sintonia da membrana basilar no ponto de 18 KHz para deslocamentos de 0.35 nm, 12 nm e 3.5 nm (adaptado de [80])

A selectividade observada em cada ponto da membrana basilar, reflecte-se em níveis superiores do sistema auditivo e até mesmo em termos perceptuais.

O movimento em cada ponto da membrana basilar é convertido pelo órgão de Córti em disparos das fibras nervosas pertencentes ao nervo auditivo. Pensa-se que esta conversão é feita pelas células ciliadas internas que se encontram ligadas por

sinapses à grande maioria (90-95%) das células nervosas do nervo auditivo e cujo potencial eléctrico é controlado pelos movimentos da membrana basilar.

Em resposta a um estímulo acústico sinusoidal, o potencial no interior das células ciliadas é composto por um nível constante a que está adicionada uma componente alternada, cuja amplitude depende da frequência de excitação. A determinação da intensidade do estímulo acústico capaz de provocar uma componente alternada de amplitude pré especificada, para diferentes valores da frequência da entrada, permite traçar curvas de sintonia que são muito semelhante às obtidas para a membrana basilar. Esta semelhança traduz a ligação estreita que existe entre o movimento da membrana basilar e o potencial medido no interior das células ciliadas internas.

A ligação estreita com o funcionamento da membrana basilar observa-se ainda ao nível do nervo auditivo. O mecanismo de disparo das fibras nervosas no nervo auditivo é complexo e não linear. Na ausência de estímulo, muitas fibras apresentam actividade espontânea, disparando de forma aleatória. Quando se apresenta um estímulo acústico, o ritmo de disparo das fibras nervosas é aumentado. Além disso, experiências com estímulos sinusoidais, mostram que os disparos tendem a ocorrer de forma síncrona com os movimentos da membrana basilar na vizinhança da fibra, mas ocorrem apenas durante metade do ciclo de excitação e cada fibra dispara só em alguns dos ciclos. O aumento da taxa de disparo com a aplicação de um estímulo sinusoidal depende da frequência do estímulo e da fibra considerada. Este comportamento é habitualmente caracterizado através de curvas de sintonia, curvas que são calculadas medindo para cada frequência do estímulo sinusoidal, a intensidade da excitação capaz de produzir um aumento pré estabelecido da taxa de disparos. As curvas de sintonia das fibras do nevo auditivo, são ainda semelhantes às curvas de sintonia da membrana basilar na vizinhança da fibra.

3 - MODELIZAÇÃO ESPECTRAL

Um sinal de fala é fortemente não estacionário, com grandes variações de energia e composto por uma grande diversidade de sons com características muito diferenciadas, quer do ponto de vista espectral, quer do ponto de vista da descrição temporal.

De acordo com isto, é habitual considerar a fala como quase estacionária apenas em intervalos de tempo de valor entre 10-30ms. Logo em vez de se utilizar um único modelo para descrever segmentos muito longos, ou mesmo frases inteiras, procura-se desenvolver modelos simples válidos em intervalos curtos onde o sinal possa ser considerado como quase-estacionário. Estes modelos traduzem o comportamento local do sinal de fala e são por isso designados modelos locais.

O uso de modelos locais é também justificado pela necessidade de restringir o atraso que, em muitas aplicações, não pode exceder algumas dezenas de milissegundos e pela analogia existente entre o processamento localizado do sinal de fala do sistema auditivo periférico que realiza continuamente uma espécie de análise em frequência com base nos últimos sons da excitação acústica.

A transformação realizada pelo ouvido interno sugere também a formulação da modelização local do sinal de fala no domínio espectral, pois aí, os erros de modelização podem ser interpretados como alterações no comportamento dinâmico da membrana basilar e mais facilmente correlacionados com resultados fisiológicos e psico-acústicos de mascaramento auditivo.

A representação em frequência de um determinado sinal, pode ser obtida através da transformada de Fourier localizada, que traduz as características espectrais do sinal analisado, na vizinhança da cada instante de tempo.

A modelização do sinal de fala, pode então ser feita no domínio da transformada através da escolha de um modelo espectral que aproxime o espectro original. A modelização é feita escolhendo uma estrutura para a modelização espectral ajustando o modelo à transformada do sinal de fala por minimização de uma medida de distorção. Finalmente, constrói-se o sinal de fala sintético a partir dos modelos espectrais utilizando, por exemplo, a técnica do overlap-add. O conceito de modelo local, a transformada de Fourier localizada, algumas medidas de distorção e a síntese são seguidamente apresentados.

3.1 - MODELIZAÇÃO LOCAL

Por modelização local num instante t_K entende-se não a aproximação directa do sinal de fala numa vizinhança de t_K mas do sinal de fala multiplicado por uma janela que selecciona o intervalo que se pretende analisar.

3.2 - MEDIDAS DE DISTORÇÃO

Uma análise de frequência localizada, permite a identificação de muitas características perceptualmente relevantes, tais como:

Estrutura harmónica, pitch e formantes. O facto de esta transformação por em evidência características importantes está relacionado com a analogia que existe entre a transformada de Fourier localizada e o processamento realizado ao nível do ouvido interno pela cóclea. Embora hajam diferenças acentuadas entre os dois tipos de transformações operadas no sinal, pois a análise em frequência realizada pela cóclea tem uma resolução maior nas baixas frequências que nas altas é seguida por um conjunto de operações não lineares complexas. Pode-se considerar que a TFL fornece uma representação do sinal de fala mais "próximo" do que é observável ao nível da membrana basilar do que a descrição original no domínio do tempo. Parece por isso ser um bom domínio para a definição de medidas de distorção entre o sinal original e o sinal sintético resultante da modelização.

3.3 - MODELO HARMÓNICO GENERALIZADO

Este modelo foi desenvolvido no início da década de 80 por Almeida Tribolet, o modelo harmónico generalizado (MHG) propõe uma solução para a modelização espectral da fala vocalizada constituída por um conjunto de aproximações do espectro localizado de crescente complexidade. Estas aproximações temam em consideração o carácter não estacionário da fala vocalizada, descrevendo-a como a soma de exponenciais complexas de frequências constantes e amplitudes variantes no tempo.

O tipo de variação admitida para as amplitudes, depende da complexidade pretendida; no caso mais simples, designado modelo de ordem zero, as amplitudes são constantes e a solução obtida é equivalente ao modelo harmónico clássico apresentado anteriormente.

O outro extremo, é o caso em que não se impõe nenhuma restrição ao número de parâmetros do modelo obtendo-se uma representação perfeita do sinal de fala.

Uma característica importante deste modelo, ao contrário de outros modelos é que ele pode ser derivado com rigor a partir de um modelo de produção de fala; consegue assim fornecer um enquadramento elegante para o problema da modelização sinusoidal da fala vocalizada, estando por um lado enraizado num modelo de realidade física e permitindo, por

outro lado, diferentes graus de sofisticação na aproximação do sinal de fala.

3.4 - CONCLUSÕES PARCIAIS RELATIVAS A ESTES ASSUNTOS

O modelo harmónico generalizado, permite aproximar o sinal de fala com precisão arbitrariamente grande em qualquer intervalo finito.

Esta propriedade revela-se de grande importância pois estabelece que não há nenhum limite à precisão do modelo harmónico desde que não se restrinja o número de parâmetros.

De qualquer forma, esta conclusão não é assaz surpreendente, uma vez que as amplitudes dos harmónicos são moduladas por séries de potências, logo bastaria no limite uma única onda harmónica para aproximar o sinal com a precisão pretendida em qualquer intervalo de finito.

O uso de uma única onda harmónica, é uma solução sem interesse porque apesar de ter potencialidades para apresentar com precisão o sinal, exige um número muito elevado de parâmetros.

A questão fundamental no desenvolvimento de um modelo, é a de atingir a "precisão" desejada com o menor número possível de parâmetros. O modelo tem por isso que estar adaptado à classe de sinais a que se destina e ao critério usado para avaliar o seu desempenho. O modelo harmónico, é uma solução adequada destes dois pontos de vista. A característica mais importante da classe de sinais a que se aplica é serem quase-periódicos, ou no domínio espectral, terem espectros compostos por riscas localizadas em múltiplos da frequência fundamental.

Acresce a isto que o modelo harmónico é especialmente eficiente na aproximação deste tipo de espectros desde que a frequência fundamental seja determinada com precisão. No que diz respeito à avaliação dos resultados, o critério de avaliação é, em última análise, o sistema auditivo humano.

De acrescentar ainda que o modelo harmónico permite uma boa representação de todo o espectro localizado, com vantagem para as frequências mais baixas onde a estrutura de riscas é habitualmente mais marcada.

3.5 - ESTIMAÇÃO ESPECTRAL

O problema de modelização não estacionário do sinal de fala pela sobreposição de funções de base na vizinhança da origem, pode ser formulado no domínio da frequência, tentando-se aproximar o espectro localizado do sinal de fala $S(t_0, \omega)$ pelo modelo espectral $S(\omega)$ de DSLMF (decomposição de sinais linearmente modulados em frequência). Este problema, pode ser resolvido por estimação directa dos parâmetros do modelo por minimização de um critério de erro quadrático, pela equação:

$$D = \frac{1}{2\pi} \int_{-\infty}^{\infty} |S(t_0, \omega) - \hat{S}(\omega)|^2 d\omega$$

Esta atitude, pode conduzir ao problema de optimização não linear envolvendo a estimação simultânea dos parâmetros de todas as sinusóides. Daí que, por vezes se adopte uma solução diferente, que pode ser a decomposição e a estimação de múltiplas sinusóides numa série de questões, mais simples, em que apenas se estima uma única função de base e de uma forma linear.

3.6 - CONCLUSÕES PARCELAES DA RELAÇÃO COM O MÉTODO DE GAUSS-SEIDEL

A estimação de múltiplas sinusóides pelo método de eliminação espectral está relacionado intimamente com o método de Gauss-Seidel para a resolução de sistemas de equações não lineares.

O algoritmo de eliminação espectral envolve não só a estimação das amplitudes complexas das exponenciais, mas também, dos parâmetros que controlam a forma das riscas espectrais usadas para aproximar o espectro localizado. A extracção óptima destes parâmetros de acordo com um critério de erro quadrático é um problema não linear e por isso, o método de eliminação espectral pode ser interpretado como a decomposição subóptima de um problema não linear em vários problemas de

estimação lineares. No entanto, se a forma das riscas espectrais for mantida constante, a solução óptima pode ser obtida pela resolução de um sistema de equações não lineares e o algoritmo de eliminação espectral, é nesse caso, equivalente ao algoritmo de Gauss-Seidel para a resolução de sistemas lineares.

3.7 - INTRODUÇÃO DA ESCALA DE TEMPO

Em segmentos vocalizados, o espectro localizado é idealmente composto por riscas espectrais localizadas em frequências de valor aproximado a múltiplos da frequência fundamental.

Esta estrutura de risca é habitualmente mais aparente nas baixas frequências (0-2kHz) que nas altas, o que levanta dificuldades na estimação das sinusóides pois não se sabe se existe alguma estrutura que deva ser preservada ou se as altas frequências devem simplesmente, ser tratadas como ruído. Esta dificuldade pode também aparecer em exemplos com frequências relacionadas harmonicamente em que não existem dúvidas acerca da existência de uma estrutura harmónica subjacente.

O que acontece nestes casos, é que quando as riscas espectrais estão muito próximas ou quando a variação da frequência instantânea é grande, há forte interacção espectral entre riscas vizinhas, a qual distorce o espectro de amplitude, criando picos deslocados das frequências centrais das sinusóides.

Isto dá origem a efeitos de distorção linear de escala de tempo na representação espectral do sinal de fala vocalizado.

4 - MODELIZAÇÃO SINUSOIDAL NO DOMÍNIO DO TEMPO

O sinal de fala contém habitualmente zonas vocalizadas extensas, com estrutura quase-periódica, produzidas por vibração das cordas vocais. Quando analisadas através de espectrogramas, as zonas vocalizadas revelam ser compostas por harmónicos de frequência fundamental que evoluem de forma lenta ao longo do tempo. Esta estrutura sugere a modelização da fala vocalizada através da soma de sinusóides com frequências variantes no tempo, acompanhando a evolução da frequência fundamental. Nas zonas não-vocalizadas, já não existe motivação tão forte para a utilização de funções de base sinusoidais pois o espectro localizaado já não é formado

por riscas espectrais bem separadas em frequência. No entanto, por motivos de homogeneidade e robustez, prefere-se muitas vezes estender a modelização sinusoidal a zonas não vocalizadas a ter que utilizar modelos diferentes. Como o espectro dos sons não-vocalizadas já não tem estrutura harmónica, é necessário utilizar um número maior de sinusóides na modelização da fala não-vocalizada.

Em termos históricos realço o seguinte:

O uso de sinusóides para a geração de sons musicais estacionários, é antigo. Foi por exemplo determinante nos estudos realizados por Ohm (1843) e por Helmholtz (1875) sobre a percepção de sons periódicos. Os resultados destes estudos tiveram forte influência sobre gerações de investigadores. Mais recentemente, modelos sinusoidais no domínio do tempo têm sido usados em diferentes áreas do processamento da fala nomeadamente em codificação, em síntese, em mudanças de escala e melhoramento da fala distorcida. É difícil encontrar uma lista extensa de contribuições na área de modelização sinusoidal da fala, e mesmo muitos dos trabalhos pioneiros não são do conhecimento geral.

4.1 - MODELOS SINUSÓIDAI DA FALA

São dois os modelos que usaremos para sintetizar o sinal de fala e que serão designados (modelo sinusoidal básico (MSB) e modelo sinusoidal em quadratura (MSQ)).

O modelo sinusoidal básico é um modelo relativamente conhecido e que tem grande aplicação em termos de codificação e fala; quanto ao segundo é de desenvolvimento Português e relativamente recente.

Em qualquer um destes modelos, a escala de tempo é dividida em segmentos justapostos (não necessariamente de comprimento constante) e o sinal de fala é aproximado, em cada segmento, por uma soma de sinusóides moduladas em amplitude e frequência. Os modelos ficam especificados definindo o tipo de variação adoptada para as amplitudes e fases das sinusóides e as relações que devem existir entre as sinusóides pertencentes a segmentos consecutivos.

4.2 - ESTIMAÇÃO DE AMPLITUDES E FASES

Na modelização sinusoidal no domínio do tempo, usam-se intervalos de síntese não sobrepostos, e aproxima-se o sinal de fala em cada intervalo pela sobreposição de sinusóides.

Os modelos por vezes especificam a evolução da fase e da amplitude de cada sinusóide parametrizando-a através de um conjunto de vários parâmetros: os valores instantâneos da amplitude, frequência e fase nos extremos do segmento.

A estimação conjunta destes parâmetros é um problema difícil, tendo-se por isso preferido estimar primeiro as frequências e só depois calcular as amplitudes e as fases. As soluções habitualmente utilizadas para estimar as amplitudes e fases das sinusóides, não são métodos de análise por síntese, pois não se baseiam na minimização de um funcional de custo entre o sinal sintético e o sinal de fala original. Em vez disso procuram aproximar localmente o sinal de fala, na vizinhança dos instantes de estimação, usando um modelo sinusoidal estacionário. As estimativas assim obtidas nada têm a ver com a forma como o modelo de síntese interpola os valores das amplitudes e das fases ao longo do segmento de síntese obtendo-se por isso estimativas não óptimas das amplitudes e fases das sinusóides.

Na continuação deste descrição, fala-se agora de dois métodos não óptimos de estimação da amplitudes e fases, que têm tido grande aplicação em modelização sinusoidal: o método de amostragem espectral e o método de mínimos quadrados estacionário. Poderíamos também falar de métodos quase-óptimos de estimação de amplitudes.

Apesar da sua divulgação, estes dois métodos têm sido apresentados como soluções desligadas, não tendo havido interesse na compreensão da relação que existe entre eles, nem a preocupação em se descreverem propriedades que auxiliem a sua implementação. De acordo com o objectivo de criar um quadro comum que permita estabelecer relações entre diferentes modelos sinusóideais, a compreensão das relações entre os métodos de análise é de grande importância. De grande importância é também a representação de propriedades que possam reduzir significativamente a complexidade dos algoritmos e torná-los mais competitivos em termos de implementação.

4.3 - ESTIMAÇÃO POR AMOSTRAGEM ESPECTRAL

Constitui a técnica mais simples para a estimação de amplitudes e fases. Este método admite estacionariedades das amplitudes e fases das sinusóides que compõem o sinal de fala na vizinhança do instante de estimação e despreza a interacção espectral entre riscas diferentes.

4.4 - MÉTODOS DE MÍNIMOS QUADRADOS ESTACIONÁRIO

O método dos mínimos quadrados estacionário (MQE) tenta aproximar o sinal de fala por uma soma de sinusóides com amplitudes e frequências constantes na vizinhança do instante em que se pretende realizar a estimação. Esta vizinhança é frequentemente designada por intervalo de análise e está habitualmente centrada num extremo do segmento de síntese.

4.5 - ESTIMAÇÃO QUASE-ÓTIMA DE AMPLITUDES

Os métodos de análise descritos anteriormente estão baseados em hipóteses simplificativas. Admitem Por exemplo que as sinusóides têm frequências e amplitudes constantes no intervalo em que é realizada a análise, admitem independência da estimação de parâmetros em instantes de tempo distintos, e não tomam em consideração a forma como as amplitudes e fases das sinusóides são interpoladas ao longo do segmento de síntese. Em contrapartida, conduzem a algoritmos e a fala de alta qualidade em zonas vocalizadas.

4.6 - MÉTODO QUASE-ÓTIMO MODIFICADO

Enquanto que na análise quase-ótima se estima em cada iteração as amplitudes de todas as funções de base auxiliares que não sejam identicamente nulas no horizonte temporal. Outra possibilidade consiste em utilizar apenas as funções auxiliares totalmente contidas no horizonte temporal, eliminando portanto as que nascem no

último segmento, isto leva a uma interpretação matricial muito mais prática e interessante.

4.7 - ESTIMAÇÃO DE SINAIS CONSTANTES

É instrutivo estudar a estimação de sinais constantes pelas técnicas anteriores. A análise do comportamento dos algoritmos resulta simplificada pelo facto de o modelo utilizar, em cada segmento e tal como é lógico, uma única sinusóide de frequência nula. Isto permite responder a questões relacionadas com os modelos anteriores, nomeadamente a de saber se as estimativas do modelo quase-óptimo são centradas ao evocar o efeito do comprimento do horizonte temporal do comportamento do método.

5 - REPRESENTAÇÃO DA FALA POR FUNÇÕES DE BANDA ESTREITA

De uma forma geral, os modelos sinusoidais são técnicas eficientes de modelização de zonas vocalizadas quase-periódicas, pois permitem obter fala vocalizada com muito alta qualidade, mesmo quando se utilizam segmentos de síntese relativamente longos (20-30 ms). No entanto, os modelos sinusoidais têm maior dificuldade em representar zonas não vocalizadas do sinal de fala, assim como algumas zonas vocalizadas em que o sinal deixa de poder ser considerado como quase periódico. O modelização destas zonas exige habitualmente um número maior de sinusóides e, mesmo assim, a qualidade degrada-se mais rapidamente quando se aumenta o comprimento do segmento de síntese. Quando se reduz o número de sinusóides ou quando se aumenta o comprimento do segmento de síntese mantendo constante o número de sinusóides, obtêm-se fala com um carácter tonal intenso e muitos artefactos que não são aceitáveis em qualidade telefónica, ou mesmo de comunicações.

Para se obter fala não-vocalizada de alta qualidade, é necessário utilizar um número de sinusóides da ordem de metade do número de amostras de fala de cada segmento. Como cada sinusóide tem dois parâmetros, os modelos sinusoidais não permitem redução significativa do número de parâmetros a codificar nos segmentos

não-vocalizadas, o que dificulta a sua utilização em codificação a ritmos médios e baixos.

Este capítulo estuda um conjunto alternativo de funções de base, mais adaptado a zonas não-vocalizadas. Tentou-se incorporar na escolha desse conjunto algum conhecimento sobre o funcionamento do sistema auditivo periférico e do mecanismo de percepção auditiva. As funções de base escolhidas são sinusóides moduladas em amplitude por sinais aleatórios passa-baixo, e serão designadas funções de base de banda estreita, uma vez que a sua energia está concentrada em torno da frequência central da sinusóide. O uso destas funções na modelização de fricativas não-vocalizadas permite a obtenção de qualidade alta ou muito alta, sem artefactos tonais, com um número muito menor de parâmetros, o que torna esta representação promissora em termos de codificação.

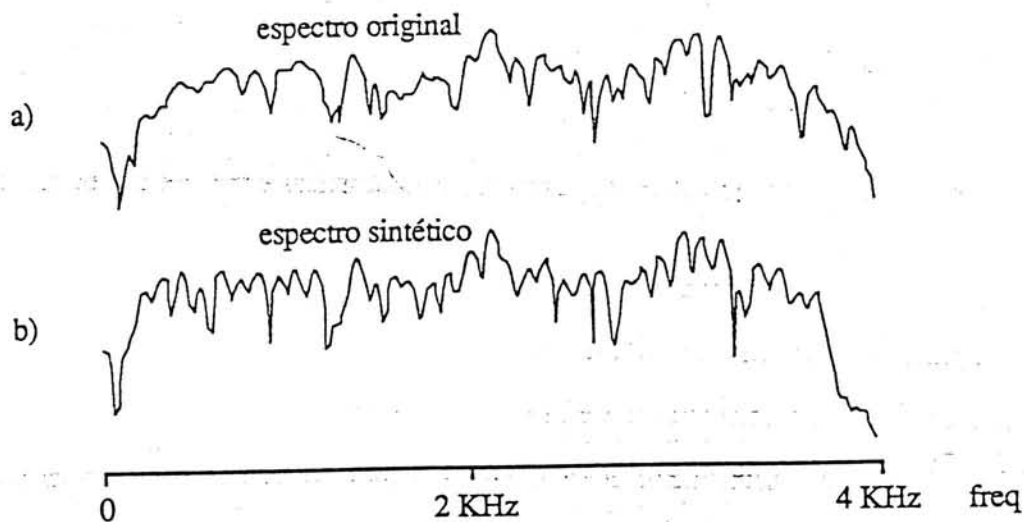
5.1 - MODELIZAÇÃO DE FALA NÃO-VOCALIZADA COM FUNÇÕES DE BANDA ESTREITA

Tal como na modelização sinusoidal, o sinal de fala é dividido em segmentos de comprimento T_j e em cada segmento é aproximado por uma soma de funções de base multiplicadas por amplitudes linearmente variantes no tempo. Para permitir uma estimação mais eficaz dos parâmetros utilizar-se-á uma estrutura semelhante à do modelo sinusoidal em quadratura em que o sinal de fala no segmento j é aproximado por:

$$\hat{s}_j^t(t) = \sum_{k=1}^{n_j} a_{ck}^j(t) \phi_{ck}^j(t) + a_{sk}^j(t) \phi_{sk}^j(t)$$

e em que $\phi_{ck}^j(t)$, $\phi_{sk}^j(t)$ são as funções de base da decomposição;
 $a_{ck}^j(t)$, $a_{sk}^j(t)$ são as respectivas amplitudes variantes no tempo e n_j é o número de pares de funções de base usadas na decomposição.

Para a escolha das funções de base tentar-se-á compreender a razão do insucesso do modelo sinusoidal quando se reduz o número de sinusóides. Este insucesso não deve ser interpretado apenas em termos de conceitos de teoria dos sinais, ou usando critérios de qualidade como a relação sinal-ruído segmental, mas deve sobretudo ser relacionado com aspectos da percepção dos sons. O carácter tonal observado quando o número de sinusóides é reduzido está provavelmente relacionado com a capacidade do sistema auditivo para detectar periodicidades no estímulo acústico, o que parece ser uma característica importante na audição humana. Esta interpretação sugere que se tente destruir, de alguma forma, a periodicidade das funções de base sinusoidais. Outra explicação para o carácter tonal provocado pela redução do número de sinusóides é a sensibilidade do ouvido humano a "vazios" no espectro de potência ou seja, regiões com pouca energia (ver figura seguinte).

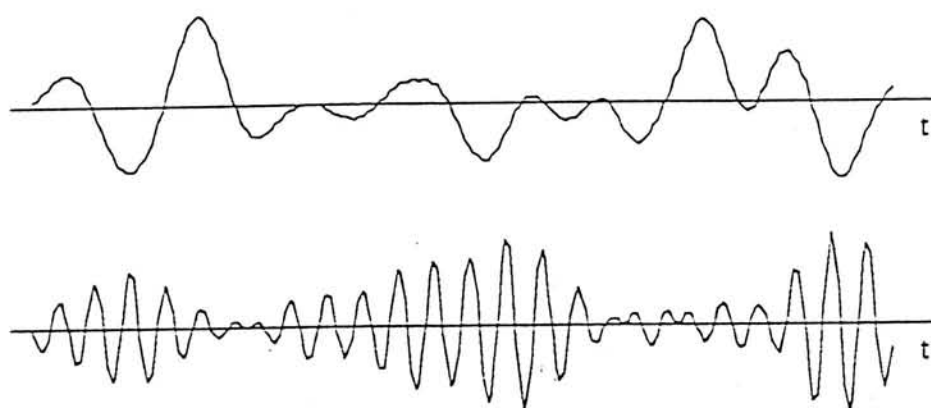


Espectro localizado de segmento não-vozeado (dB):

a) espectro original; b) espectro sintético usando 40 sinusóides e segmentos de 20 ms.

Estes "vazios" ocorrem sempre que a diferença entre as frequências centrais das sinusóides é menor que a sua largura de banda. Como os espectros dos sons não-vocalizados não apresentam estrutura harmónica e têm geralmente uma envolvente espectral suave, é de esperar que possam ser bem modelizados por funções de base com largura de banda maior. Assim, as sinusóides vão ser substituídas nas zonas não-vocalizadas por funções de base de banda estreita, obtidas modulando a amplitude de sinusóides por processos estocásticos passa-baixo.

A figura seguinte mostra um exemplo de duas funções de base de banda estreita.



Funções de base de banda estreita

5.2 - CODIFICAÇÃO DE FALA NÃO VOCALIZADA COM FUNÇÕES DE BANDA ESTREITA

As experiências anteriores de análise/síntese mostraram ser possível obter fala não-vocalizada de alta qualidade com um número reduzido de funções de base desde que se utilizem funções de banda estreita em vez de sinusóides. Tendo uma largura de banda maior que a das sinusóides, as funções de banda estreita permitem cobrir todo o espectro com um número menor de funções e simultaneamente aproximar com detalhe suficiente a envolvente espectral de amplitude nas zonas não-vocalizadas. Estas características sugerem a aplicação das funções de banda estreita na codificação de fala a ritmos médios e baixos.

6 - CODIFICAÇÃO SINUSOIDAL

O progresso tecnológico e a crescente importância das redes de telecomunicações na organização social e comercial das sociedades actuais têm aberto perspectivas novas e aliciantes à comunicação digital da fala. A resposta a este desafio tem tido duas componentes:

- i) O desenvolvimento de melhores algoritmos de codificação de fala.
- ii) O desenvolvimento de um grande esforço para o estabelecimento de standards de codificação.

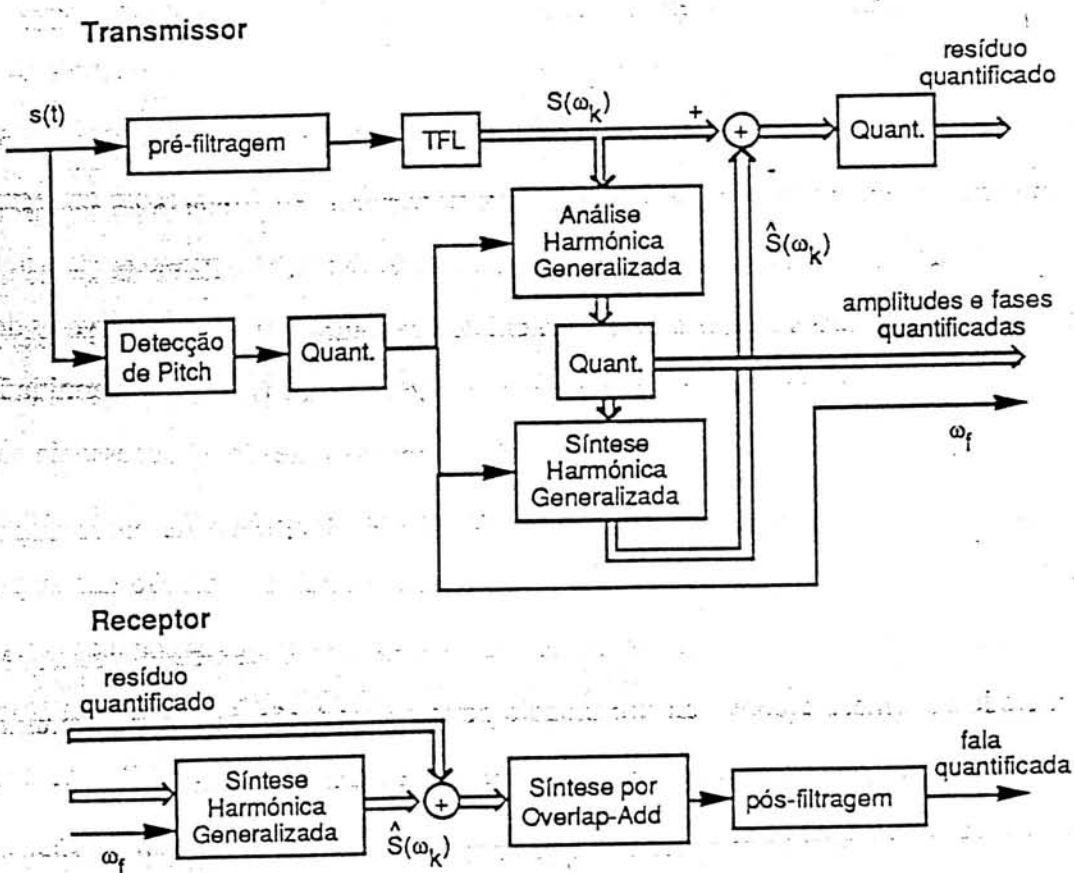
Uma dificuldade comum a muitas das actuais áreas da comunicação digital da fala (p. ex. comunicação por satélite, telemóvel e aplicações militares) é a necessidade de se utilizarem canais de comunicação com larguras de banda reduzidas e consequentemente ritmos de transmissão médios ou baixos.

No final dos anos 70 havia tipicamente duas classes de codificadores: os codificadores de forma de onda e os vocoders, que funcionam a ritmos muito mais baixos com qualidade sintética. Actualmente a tendência é para o preenchimento da lacuna entre estas duas classes de codificadores, procurando soluções que funcionando a ritmos médios e baixos, mantêm muitas características dos codificadores de forma de onda, como: fala de alta qualidade, alta inteligibilidade e desempenho independente do ruído ambiente. Estas soluções têm sido desenvolvidas de forma a adaptarem-se aos diversos ritmos de transmissão em estudo. Temos ainda, uma alternativa menos explorada, os codificadores sinusoidais e harmónicos em que o sinal de fala é

sintetizado por meio de sinusóides. O codificador harmónico foi originalmente desenvolvido para melhorar a eficiência do ATC (adaptive transform coding) nas zonas vocalizadas, substituindo a transmissão de todos os valores da transformada pela transformação das amplitudes e fase dos harmónicos, que são em menor número e permitem uma codificação mais eficiente, por intermédio do uso de técnicas de predição.

Posteriormente foram seguidas duas linhas distintas: continuação do uso de técnicas de modelização e codificação do espectro localizado com síntese por overlap-add e técnicas de modelização e síntese no domínio do tempo.

A nível de curiosidade apresenta-se de seguida a figura representativa da estrutura de um codificador harmónico.



Estrutura do codificador harmónico

6.1 - CODIFICAÇÃO DE FASE

A codificação de fase é uma das principais fontes de degradação de qualidade da fala sintética em codificadores sinusoidais, merecendo por isso uma atenção especial. Quando o ritmo de transmissão é baixo há um número reduzido de bits para codificar as fases das sinusóides. Por exemplo, nos segmentos não vocalizados, em que se usa um número muito elevado de sinusóides para representar o sinal de fala, existe habitualmente menos um bit por sinusóide para codificar as fases.

A codificação das fases, deve por esse motivo, obedecer a uma estratégia ou modelo bem definidos que permitam manter as características perceptualmente mais importantes.

6.2 - ESTRUTURA DA FASE EM ZONAS NÃO-VOCALIZADAS

Nas zonas vocalizadas, a importância da fase é grande. A ideia habitual de que o "ouvido é surdo à fase", só é válida para uma classe muito restrita de distorções de fase em que não se inclui o tipo de distorções, variantes no tempo, que resultam da quantificação da fase em codificadores sinusoidais.

Nas zonas não vocalizadas, a importância perceptual da fase não é bem conhecida. Aparentemente, tem um papel muito menor que nas zonas vocalizadas; não significando isto que todas as modificações de fase sejam desprezáveis ou toleráveis.

Posso agora referir certas experiências realizadas e que chegaram até mim em termos de bibliografia, para estudar o efeito da transformação de fase em zonas não-vocalizadas, experiências estas realizadas com sons fricativos não-vocalizadas e com ruído branco. Nestas experiências, utilizou-se o modelo sinusoidal básico com amplitudes e fases estimadas por mínimos quadrados estacionários e frequências equiespaçadas na gama de frequências de interesse, usando segmentos de síntese de 20 ms e um máximo de 80 frequências. Num primeiro teste, a informação de fase foi eliminada anulando a fase de todas as sinusóides nos extremos dos segmentos de síntese. A fala sintetizada ficou fortemente degradada apresentando um carácter tonal intenso. Numa segunda experiência, fez-se uma filtragem passa-baixo das fases em função da frequência, substituindo cada fase pela média da três valores vizinhos, apesar de tudo obteve-se ainda um resultado com carácter fortemente tonal.

É lícito concluir daqui, que há transformações na fase que produzem distorções fortes, apesar do ouvido não ser provavelmente muito sensível a distorções de fase em zonas não vocalizadas.

É nossa convicção que o aspecto mais importante na percepção da fase em zonas não-vocalizadas não é o valor individual de cada uma das fases, mas sim, a sua estrutura aleatória, tanto no domínio da frequência como ao longo do tempo. Ou, em termos mais concretos, se não quisermos gastar bits a codificar as fases das sinusóides num troço não-vocalizado, mais vale utilizar fases aleatórias, distribuídas uniformemente entre $-\Pi$ e Π , que anular as fases.

6.3 - DESENHO E ESTUDO DE UM CODIFICADOR DE 4.8 kb/s

6.3.1 - ESTRUTURA BÁSICA

Este ponto é também extraído da bibliografia, a razão pela qual lhe dou relevância é porque me parece importante, mais que não seja, pelo detalhe da descrição.

Uma das primeiras opções a tomar no desenho de um codificador sinusoidal, diz respeito à sua estrutura. É sabido que existem várias técnicas sinusoidais de análise e de síntese de fala e várias estruturas de codificação. Temos assim várias opções para o desenvolvimento de um codificador sinusoidal, conforme se escolhem modelos no domínio do tempo ou da frequência, ou consoante se usa uma decisão rígida entre métodos diferentes dedicados a zonas particulares do sinal de fala ou se tente integrar numa única técnica a modelização e codificação de todo o sinal.

As dificuldades existentes em alguns codificadores híbridos, baseados numa decisão rígida entre técnicas muito diferenciadas, levam a uma maior cautela nesse ponto, de forma a que se possa adoptar uma solução tão homogénea quanto possível. Talvez por isso, o autor tenha decidido usar um modelo sinusoidal válido tanto em zonas vocalizadas, como em zonas não-vocalizadas, escolhendo o modelo sinusoidal básico com amplitudes e fases estimadas pelo método de mínimos quadrados estacionário.

O principal obstáculo à utilização de modelos sinusoidais em codificação a ritmos médios e baixos tem sido o número excessivamente grande de parâmetros que é

preciso codificar. Esta dificuldade não se faz sentir tão pesadamente em zonas vocalizadas, pois o carácter quase-periódico do sinal de fala, permite reduzir o número de parâmetros a transmitir por unidade de tempo, e simultâneamente desenvolver técnicas de codificação eficientes. O problema põe-se, sobretudo, nas zonas não-vocalizadas e fortemente não estacionárias que necessitam de um maior número de sinusóides por unidade de tempo, não permitindo o uso eficiente de técnicas de predição na codificação dos seus parâmetros.

Outra questão delicada é a escolha do comprimento de síntese. Nas zonas vocalizadas, os parâmetros das sinusóides evoluem habitualmente de uma forma suave, o que tem a vantagem de permitir utilizar segmentos de síntese relativamente longos (20-30 ms), deixando portanto um maior número de bits disponíveis para codificar os parâmetros das sinusóides. A codificação cuidada dos parâmetros nas zonas vocalizadas, (fase por exemplo) é muito importante, uma vez que o ouvido é especialmente sensível a distorções da forma de onda em zonas estacionárias quase-periódicas, como é o caso da fala vocalizada. No entanto o uso de segmentos longos, provoca dificuldades de modelização.

Temos desta forma que a escolha do segmento de síntese, representa um compromisso entre erros de modelização e erros de quantificação.

6.3.2 - CODIFICAÇÃO DO MODELO SINUSOIDAL A RITMOS MÉDIOS E BAIXOS

De seguida, vou falar da codificação dos parâmetros do modelo sinusoidal básico, ou seja, a codificação das amplitudes, frequências e fases das sinusóides. O primeiro aspecto que vou tratar é o da determinação das frequências das sinusóides.

Como os modelos sinusoidais, foram usados inicialmente para representar sinais de fala quase-periódicos, escolheram-se nos estudos iniciais de modelização e codificação, frequências múltiplas da frequência fundamental (modelização harmónica). Quando o uso de modelos sinusoidais foi estendido a zonas não-vocalizadas, que não possuem estrutura harmónica, pensou-se que as frequências não poderiam continuar a ser harmónicamente relacionadas e deveriam ser independentes.

Esta opção implica contudo, a transmissão de todas as frequências, o que é impraticável não só a ritmos baixos, como mesmo a ritmos médios. Felizmente, resultados desenvolvidos mostraram que era possível manter a restrição harmónica das frequências sem deterioração da qualidade perceptual da fala produzida.

Resulta desta forma um esquema mais dependente de estimativas de pitch, ficando todas as frequências especificadas por um único parâmetro: a frequência fundamental, que é por conveniência mantida constante e pré-definida nos segmentos não-vocalizados. Assim para além da decisão de vocalização, só a frequência fundamental tem que ser codificada, somente nos segmentos vocalizados.

Uma vez que a codificação das frequências envolve apenas um único parâmetro, a grande maioria dos bits disponíveis, são gastos na codificação das amplitudes e fases. De qualquer forma, a ritmos médios e baixos o número de bits disponíveis em cada segmento é manifestamente insuficiente para codificar com precisão as amplitudes e fases das sinusóides. Existe por isso um compromisso grande entre o número de bits atribuídos às amplitudes e o número de bits usados para codificar as fases tornando-se assim necessário encontrar formas de codificação dos parâmetros em bloco. Para codificar as amplitudes das sinusóides utilizou-se a envolvente espectral de LPC (abstenho-me de explicar esta envolvente).

As amplitudes são calculadas por amostragem do espectro de LPC nas frequências das sinusóides. Experiências realizadas com esta técnica, mantendo os restantes parâmetros não quantificados, permitiram concluir que a fala produzida tem ainda qualidade alta, não só nas zonas vocalizadas quase-estacionárias, como já tinha sido notado, mas também em todo o sinal de fala em geral. A codificação da amplitude das sinusóides fica portanto reduzida ao cálculo e codificação do filtro de LPC o que pode ser feito usando técnicas bem conhecidas.

A figura seguinte mostra a estrutura simplificada do transmissor e receptor do codificador sinusoidal. Ao contrário do que acontece em outros codificadores, existe uma separação clara, ao nível do transmissor, entre a estimação de parâmetros e a sua quantificação. A estimação de parâmetros é realizada em três blocos:

- i) Extração de pitch e decisão de vocalização,
- ii) Cálculo das frequências das sinusóides como múltiplos da frequência fundamental e
- iii) Estimação de fases pelo método de mínimos quadrados estacionário (MQE).

A quantificação dos parâmetros do modelo harmónico envolve, para além dos blocos de quantificação propriamente ditos, o cálculo e quantificação da envolvente espectral que é usada na atribuição de bits para a codificação de fase e na geração das amplitudes das sinusóides.

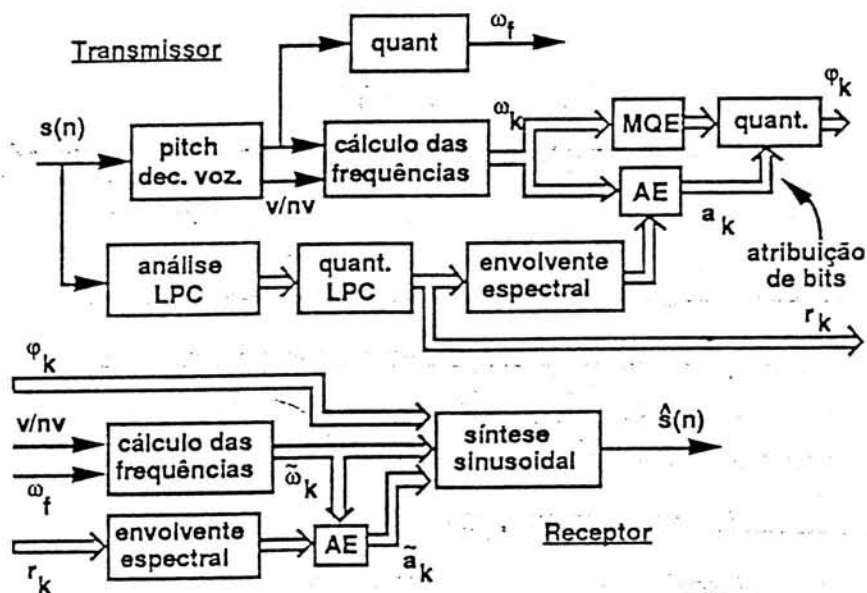


Diagrama de blocos do codificador harmónico

O funcionamento do codificador pode ainda ser compreendido através do diagrama temporal do transmissor e do receptor. Para se sintetizar um segmento de fala no receptor, é necessário conhecer as amplitudes, frequências e fases das sinusóides nos extremos do intervalo de síntese. Os parâmetros das sinusóides no início do intervalo são já conhecidos do receptor, pois foram os parâmetros finais do segmento anterior. Basta, por isso, estimar e transmitir os parâmetros das sinusóides no final do segmento actual de síntese.

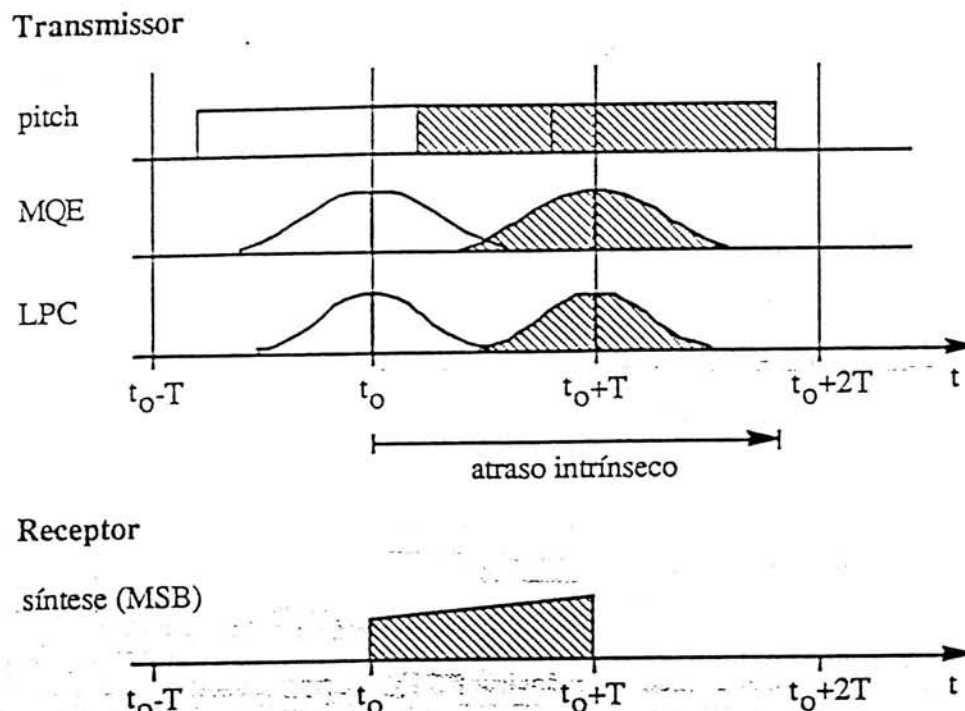


Diagrama temporal do transmissor e receptor

a tracejado: posição actual das janelas; a branco: posição anterior

No codificador sinusoidal a 4.8 Kb/s usam-se intervalos de síntese de 20 ms. A frequência fundamental é calculada a partir de um bloco de 32 ms indicado a tracejado na figura anterior. O método de mínimos quadrados estacionário usa uma janela de comprimento variável entre 20 e 32 ms e o filtro LPC é calculado usando uma janela de 20 ms. O atraso intrínseco associado a este codificador é portanto de um intervalo de síntese mais metade do maior bloco de análise (20+16 ms). Este atraso não inclui o tempo de processamento no transmissor que pode atingir a duração de um intervalo de síntese (20 ms) nem o tempo gasto na transmissão que é também de 20 ms se o ritmo de transmissão for constante. É desta forma realista pensar num atraso total da ordem dos 70 ms.

6.3.3 - DESCRIÇÃO DO CODIFICADOR DE 4.8 Kb/s

Seguidamente, vou descrever as principais características do codificador, bem como algumas opções tomadas ao nível da bibliografia.

6.3.4 - COMPRIMENTO DO INTERVALO DE SÍNTESE

Na escolha do intervalo de síntese temos em contraposição dois factores principais: a distorção da modelização, que é tanto maior, quanto maior for o intervalo de síntese e a distorção de quantificação que diminui à medida que o intervalo de síntese aumenta, dado que aumenta também o número de bits disponíveis para codificar os parâmetros do modelo sinusoidal. O melhor compromisso que se conseguiu estabelecer entre estes dois factores de distorção, resultou segundo as experiências realizadas na bibliografia no uso de intervalos de 20 ms. Como o sinal de fala é amostrado a 8 KHz um intervalo de 20 ms corresponde a 160 amostras.

6.3.5 - CÁLCULO E CODIFICAÇÃO DA FREQUÊNCIA FUNDAMENTAL

A frequência fundamental, é obtida usando algoritmos apropriados. As estimativas da frequência fundamental, podem ainda ser corrigidas a partir de coeficientes do modelo harmónico generalizado.

6.3.6 - CÁLCULO DAS FREQUÊNCIAS DAS SINUSÓIDES

As frequências das sinusóides nos extremos dos segmentos de síntese são calculados a partir da frequência fundamental, obtida pelo algoritmo apropriado, nas zonas vocalizadas e pré-especificada de forma conveniente, nas zonas não-vocalizadas.

As frequências das sinusóides, são calculadas tanto no transmissor como no receptor, mas por motivos diferentes. No transmissor são necessárias para a estimação das fases (e amplitudes) pelo método de MQE. Utiliza-se por isso o valor não-quantificado da frequência fundamental para o cálculo das frequências das sinusóides, os desvios no valor da frequência fundamental, degradam fortemente a estimação de amplitudes e fases, especialmente nos harmónicos de ordem mais elevada. No receptor, utilizam-se naturalmente valores quantificados da frequência fundamental na síntese do sinal de fala pelo MSB.

6.3.7 - CÁLCULO E CODIFICAÇÃO DE UM FILTRO DE LPC

Para o cálculo do filtro de LPC, usou-se o método da covariância pesado, não se tendo contudo realizado qualquer correcção de alta frequência como é usual em codificadores no domínio do tempo. A escolha do método da covariância deve-se ao facto de conduzir a menor erro de predição e melhor ajuste espectral que o método da correlação embora a diferença seja pequena para intervalos de 20 ms (mais ou menos 160 amostras), utilizados no cálculo dos coeficientes.

O facto do método da covariância, não garantir estabilidade do filtro de LPC não constitui desvantagem nesta aplicação pois nunca é utilizado para filtragem no domínio do tempo mas apenas para o cálculo de amplitudes das sinusóides. Foi também pelo mesmo motivo que não se utilizou correcção de alta frequência no cálculo do filtro de LPC pois modifica artificialmente a envolvente espectral para frequências altas. Para além de melhorar o condicionamento numérico na estimação do LPC, a correcção de alta frequência costuma ser justificada pelo facto do filtro inverso de LPC ser utilizado em codificadores no domínio do tempo para a predição do sinal de fala a partir das amostras anteriores quantificadas em vez das amostras originais, que foram usadas na estimação do preditor. Também esta argumentação, não se aplica no contexto em que o filtro de LPC é aqui utilizado. Finalmente a utilização da janela (com um funcionamento equiparado ao de um peso) no critério dos mínimos quadrados usado na estimação do filtro de LPC tem a finalidade habitual: reduzir a dependência dos parâmetros estimados em relação à localização do intervalo de optimização.

A codificação dos parâmetros de LPC é um assunto intensamente estudado uma vez que o número de codificadores, a diversos ritmos de transmissão, utiliza o filtro de LPC em predição ou no cálculo da envolvente espectral do sinal de fala. Não quer isto

dizer que a codificação desses parâmetros seja um assunto encerrado. Antes pelo contrário, têm-se verificado progressos importantes nos últimos anos.

6.3.8 - CÁLCULO DAS AMPLITUDES A PARTIR DA ENVOLVENTE ESPECTRAL

As amplitudes das sinusóides são calculadas por amostragem da envolvente espectral de LPC quantificada, ou seja, por amostragem do módulo da função de transferência,

$$a_k^q = |S_{LPC}^q(\omega_k)|$$

A função de transferência do filtro de LPC é dada por

$$S_{LPC}^q(\omega) = \frac{G}{A^q(e^{j\omega})}$$

em que

$$A^q(z) = 1 + a_1^q z^{-1} + \dots + a_L^q z^{-L}$$

é o filtro preditor de LPC e G um ganho. O ganho G é calculado no transmissor, por forma a que as amplitudes das sinusóides obtidas por amostragem do espectro de LPC, tenham o mesmo valor quadrático médio que as amplitudes estimadas por MQE. O valor do ganho é depois quantificado logaritmicamente, codificado e transmitido para o receptor. no receptor, as amplitudes das sinusóides são calculadas de forma idêntica, usando os parâmetros de LPC e o ganho quantificados.

6.3.9 - ESTIMAÇÃO DE FASES

As fases das sinusóides são estimadas pelo método de mínimos quadrados estacionário que, de entre as várias técnicas, representa o melhor compromisso entre qualidade e esforço computacional.

6.3.10 - SÍNTESE SINUSOIDAL

O único bloco do receptor que não existe no transmissor é o da síntese sinusoidal. A síntese sinusoidal é o bloco de maior complexidade computacional de todo o codificador, com excepção para a correcção da frequência fundamental por pós-processamento harmónico, bloco que pode ser, no entanto, substituído pelo algoritmo de alinhamento de picos, de baixa complexidade computacional. A síntese sinusoidal realiza a interpolação de amplitude e fase do modelo sinusoidal básico a partir dos parâmetros das sinusóides nos extremos do segmento de síntese, e a construção do sinal de fala sintético por sobreposição dessas sinusóides.

6.3.11 - ATRIBUIÇÃO DE BITS DO CODIFICADOR SINUSOIDAL

Como se escolheu um intervalo de síntese de 20 ms, o codificador sinusoidal a 4.8 Kb/s dispõe de 96 bits para transmitir a informação correspondente a cada intervalo de síntese. A distribuição dos bits pelos diferentes parâmetros a codificar foi feita de acordo com a tabela seguinte. O critério usado foi o de atribuir o maior número possível de bits para a codificação das fases, sem comprometer a codificação dos restantes parâmetros. Para determinar o número de bits necessários para codificar as amplitudes sem degradação acentuada de qualidade foram realizados alguns testes em que apenas as amplitudes foram codificadas. Conclui-se serem necessários cerca de 40 bits para codificar os parâmetros de LPC com o algoritmo de codificação utilizado.

	V	NV
pitch	7	0
v/nv	1	1
ganho	5	5
LPC	40	40
fase	43	50
total	96	96

Atribuição de bits do codificador sinusoidal a 4.8 Kb/s

7 - CONCLUSÕES E RESUMO GERAL DO TRABALHO

A ideia de modelizar a fala através de sinusóides é sugerida pela existência de uma estrutura quase periódica em zonas extensas do sinal de fala (zonas vocalizadas). Sabendo que um sinal periódico pode ser modelizado de uma forma exacta por uma série de Fourier, ou seja, pela soma de sinusóides com amplitudes e frequências constantes, o problema que se põe é o de estender este conceito de forma a tomar em consideração a natureza não estacionária do sinal de fala. Estas não estacionariedades traduzem-se por um lado, em variações da frequência fundamental, resultantes de alterações do ritmo de vibração das cordas vocais e, por outro lado, em modificações progressivas da forma de onda correspondendo a diferentes configurações do tracto vocal. Um segundo problema no estabelecimento de modelos sinusoidais da fala consiste na escolha da estratégia a adoptar em zonas não vocalizadas e de transição em que o sinal de fala é aperiódico. Nestas zonas, as funções de base sinusóides, já não estão também adaptadas à estrutura do sinal, sendo necessário um número maior de sinusóides para se conseguir uma modelização adequada das zonas não vocalizadas.

Foram estudados modelos sinusoidais no domínio da frequência e modelos sinusoidais no domínio do tempo. No domínio da frequência estudou-se a modelização do espectro localizado do sinal de fala pela soma de "riscas espectrais"

correspondentes à transformada de Fourier localizada das funções de base sinusoidais. Foram considerados quatro modelos:

- i) O modelo harmónico clássico,
- ii) O modelo harmónico generalizado,
- iii) A decomposição em sinais linearmente modulados em frequência e
- iiii) O modelo Gaussiano.

O modelo mais simples é o modelo harmónico clássico em que as funções de base são sinusóides com amplitudes constantes e frequências múltiplas da frequência fundamental. Devido à estacionariedade das funções de base, o desempenho do modelo harmónico clássico é fortemente afectado quando, no interior da janela de análise, o sinal de fala apresenta desvios de periodicidade significativos. Esta incapacidade na representação de não estacionariedades limita a aplicação do modelo harmónico clássico a intervalos de síntese curtos obrigando por isso, à utilização de um número elevado de modelos por unidade de tempo. O modelo harmónico generalizado é uma extensão do modelo harmónico clássico em que o sinal de fala é aproximado por exponenciais complexas de frequências constantes e amplitudes complexas variantes no tempo. Embora o modelo harmónico generalizado possa aproximar o sinal de fala vocalizada com precisão arbitrariamente grande, é necessário, na prática, limitar a evolução das amplitudes complexas das exponenciais, por forma a poderem ser parametrizadas com um número limitado de parâmetros. Assim, a evolução das amplitudes é descrita através de polinómios complexos de ordem pré-especificada, dando origem a diferentes modelos harmónicos consoante a ordem escolhida. Em particular, o modelo harmónico de ordem 0, coincide com o modelo harmónico clássico, uma vez que tanto as amplitudes como as frequências das sinusóides são mantidas constantes durante o intervalo de análise.

Embora o modelo harmónico generalizado permita representar o sinal de fala com precisão arbitrariamente grande, a utilização de modelos harmónicos de ordem elevada é limitada pelo número de parâmetros, que cresce linearmente com a ordem do modelo. Por isso apenas modelos de ordem baixa têm sido utilizados na prática, o que limita fortemente o tipo de não estacionariedades que podem ser bem aproximadas. Em particular os modelos de ordem reduzida, não permitem uma boa modelização de zonas com acentuada variação da frequência fundamental. Para ultrapassar esta dificuldade desenvolveu-se um modelo novo, utilizando exponenciais cuja frequência tinha a particularidade de ter uma frequência variante no tempo. A esta representação do sinal de fala, foi chamada de decomposição em sinais linearmente modulados em

frequência (DSLME) pois as funções de base escolhidas foram exponenciais complexas com evolução de frequência linear.

Para tornar o modelo menos dependente da decisão de vocalização e das estimativas de Pitch, as frequências centrais das sinusóides são estimadas directamente a partir do espectro localizado do sinal de fala, em vez de estarem restringidas a múltiplos da frequência fundamental. Isto permite utilizar DSLME não só em zonas vocalizadas como em zonas não-vocalizadas do sinal de fala sem qualquer modificação da estrutura do modelo do algoritmo de estimação dos parâmetros.

Estudos experimentais realizados relativamente à DSLME revelou clara superioridade em relação à modelização sinusoidal estacionária em zonas com forte variação da frequência fundamental confirmando a importância da modelização não estacionária das frequências das sinusóides. Existe no entanto um problema, pois não é fácil estender a modelização não estacionária das frequências a outros tipos de evoluções temporais, como por exemplo, evoluções polinomiais de ordem superior, uma vez que se torna difícil a obtenção de modelos espectrais manuseáveis analiticamente e de métodos de estimação de parâmetros adequados. No entanto, é possível conciliar de forma muito natural e prática a modelização não estacionária de amplitudes e frequências.

A obtenção deste objectivo, passa pela modulação das amplitudes das funções de base anteriores por sinais Gaussianos. As funções de base assim obtidas são sinais Gaussianos complexos e os modelos designados por modelos Gaussianos.

Os resultados experimentais obtidos com o modelo Gaussiano confirmam a importância da modelização não estacionária do conjunto de amplitudes e frequências, comparando-se favoravelmente com os resultados do modelo harmónico generalizado para um número de parâmetros semelhante.

Em termos muito gerais, podemos dizer que estão a desenvolver técnicas para a modelização espectral do sinal de fala com funções de base sinusoidais. Existe presentemente grande flexibilidade de escolha dos modelos espectrais: sejam estes, modelos com amplitudes e frequências constantes (MHC), modelos com amplitudes variantes no tempo e frequências constantes (MHG), modelos com amplitudes constantes e frequências variantes no tempo (DSLME) e finalmente modelos em que tanto as amplitudes como as frequências são variantes no tempo (MG).

O quadro da modelização espectral da fala está portanto bastante completo, tendo sido desenvolvidas em vários trabalhos inclusive Portugueses técnicas que envolvem modelização não estacionária de frequência.

De realçar ainda que muitos destes modelos são baseados num elevado número de parâmetros que dificulta a sua utilização em codificação a ritmos médios e baixos.

Nomeadamente, só o modelo harmónico clássico, com todas as suas limitações, tem servido de base a codificadores harmónicos. A dificuldade em codificar um número muito elevado de parâmetros impede a utilização de modelos espectrais não estacionários e motivou o estudo de outra classe de modelos mais promissora em termos de codificação, os modelos sinusoidais no domínio do tempo.

No que diz respeito à modelização sinusoidal no domínio do tempo, podem-se distinguir dois traços importantes que o separam da modelização espectral:

Em primeiro lugar, o aspecto óbvio de se pretender aproximar o sinal de fala directamente no domínio do tempo.

Em segundo lugar, a modelização baseia-se na compreensão de certas grandezas, nomeadamente as frequências instantâneas e as amplitudes das sinusóides, devem evoluir de forma suave ao longo do tempo e por isso, o modelo utilizado para aproximar o sinal de fala num dado segmento não é independente dos modelos usados nos segmentos adjacentes. Assim, assegura-se habitualmente continuidade das amplitudes, frequências e fases das sinusóides nas fronteiras entre segmentos. Com estas hipóteses o problema da modelização sinusoidal no domínio do tempo fica quase completamente estruturado. Restando apenas especificar as leis de evolução das amplitudes e fases das sinusóides em cada segmento, de forma a que se verifiquem as condições de continuidade na amplitude, a frequência e fase das sinusóides nos extremos dos segmentos de síntese.

Duas soluções foram estudadas extensamente, o modelo sinusoidal básico com amplitudes lineares e fases descritas por polinómios de terceiro grau, e o modelo sinusoidal em quadratura em que cada sinusóide é decomposta na soma de duas componentes em quadratura com amplitudes lineares e independentes, e com fases quadráticas em qualquer dos modelos, as evoluções das amplitudes e fases das sinusóides são especificadas pelos valores instantâneos das amplitudes, frequências e fases nos extremos do segmento, motivo pelo qual as leis que determinam essa evolução são também designadas leis de interpolação.

Como a evolução das amplitudes, frequências e fases das sinusóides é calculada por interpolação dos valores destes parâmetros nos extremos do segmento de síntese, a estimação do modelo sinusoidal resume-se à determinação dos parâmetros das sinusóides nos extremos dos segmentos. É no entanto muito difícil estimar conjuntamente estes parâmetros para um número elevado de sinusóides, tomando em

consideração a forma como são interpolados ao longo do segmento de síntese. Por este motivo, separa-se habitualmente a estimação de frequências da estimação de amplitudes e fases.

Como a modelização sinusoidal da fala começou por ser aplicada apenas em zonas vocalizadas, as frequências foram inicialmente escolhidas como múltiplos da frequência fundamental, hipótese que foi designada por restrição harmónica. Ao estender a modelização sinusoidal a zonas não-vocalizadas e de transição pensou-se inicialmente ser necessário eliminar a restrição harmónica na determinação das frequências, uma vez que não existe estrutura

quase-periódica no sinal que justifique o emprego de frequências múltiplas.

As frequências passaram então a ser obtidas independentemente, por uma técnica de detecção de máximos do espectro localizado (método de detecção de picos). Esta atitude, tem uma desvantagem óbvia em termos de codificação pois exige a transmissão de todas as frequências das sinusóides em vez de um único parâmetro, a frequência fundamental. Em trabalhos deste tipo, retoma-se a hipótese harmónica completando-se com um critério de determinação de frequências em zonas não vocalizadas. Nestas zonas, utilizou-se um número elevado de sinusóides, com frequências equiespaçadas entre zero e a frequência máxima do sinal. Um estudo detalhado das duas técnicas em termos de relação sinal-ruído segmental mostrou que a estimação harmónica de frequências tem melhor desempenho nas zonas vocalizadas e que o método de detecção de picos é superior nas zonas não-vocalizadas. Contudo a estimação harmónica revelou melhor desempenho em termos perceptivos tendo permitido a eliminação de algumas das distorções tonais produzidas em transições e zonas não-vocalizadas quando as frequências são estimadas a partir dos picos do espectro localizado.

Na estimação das amplitudes e fases das sinusóides estudaram-se as propriedades e interligações de dois métodos que têm sido muito utilizados em trabalhos recentes: o método da amostragem espectral e o método de mínimos quadrados estacionário. Para além da análise destas duas técnicas de estimação descreveu-se um algoritmo novo que é mais potente e complexo em termos computacionais: a estimação quase-ótima de amplitudes. Resumidamente, os dois primeiros métodos calculam as amplitudes e fases das sinusóides utilizando um modelo sinusoidal estacionário que é adaptado ao sinal de fala na vizinhança do instante em que se pretende realizar a estimação. Assim, o modelo estacionário utilizado na estimação das amplitudes e fases nada tem a ver com o modelo realmente utilizado na síntese do sinal de fala, modelo em que as amplitudes e frequências das

sinusóides variam ao longo do tempo. Além disso, no método de amostragem espectral despreza-se a correlação entre as funções de base sinusóides, estimando-se os parâmetros de cada sinusóide como se as outras sinusóides não existissem. Com estas hipóteses simplificativas a estimação de amplitudes e fases tem baixa complexidade computacional. O método de mínimos quadrados estacionário resume-se à resolução de um sistema de equações lineares, que pode ser realizado de formas muito eficientes do ponto de vista computacional. No caso do método da amostragem espectral, o problema tem ainda solução mais simples.

Apesar dos métodos de amostragem espectral e de mínimos quadrados estacionário estarem baseados em hipóteses simplificativas permitem a obtenção de fala sintética de muito alta qualidade em zonas quase-estacionárias. No entanto, observam-se frequentemente distorções tonais associadas a variações rápidas da estrutura do sinal, distorções que são muito audíveis quando as frequências das sinusóides são estimadas pelo método de detecção de picos. Para resolver esta dificuldade desenvolveu-se o método quase-óptimo que toma em consideração a evolução temporal das amplitudes e frequências das sinusóides e a influência exercida pelos parâmetros estimados na modelização dos segmentos seguintes.

A estimação é feita tentando adaptar a saída do modelo de síntese ao sinal de fala ao longo de um horizonte temporal que inclui não só o segmento em que se pretende estimar os parâmetros, mas também alguns segmentos futuros. A designação de "quase-óptimo" justifica-se, uma vez que a utilização do horizonte temporal limita a influência dos segmentos futuros na estimação actual dos parâmetros. Nos testes realizados, o algoritmo quase-óptimo, mostrou clara superioridade em relação aos métodos anteriores, permitindo a obtenção de fala sintética de muito alta qualidade indistinguível do original na maioria das fases testadas. A complexidade computacional do algoritmo quase-óptimo dificulta no entanto a sua implementação em tempo real.

As técnicas de modelização sinusoidal no domínio do tempo permitem a obtenção de fala sintética de muito alta qualidade tanto em zonas vocalizadas como em zonas não vocalizadas. No entanto, a modelização nas zonas não vocalizadas exige em geral um número maior de parâmetros por unidade de tempo, o aumento do intervalo de síntese não permite aliviar esta dificuldade pois tem que ser compensado com o aumento correspondente do número de sinusóides nas zonas não vocalizadas, para evitar que a fala produzida adquira um carácter tonal intenso. A necessidade de aumento do número de sinusóides nas zonas não vocalizadas traduz o facto da modelização sinusoidal ter sido prioritariamente concebido para as zonas quase-

periódicas do sinal de fala, onde o espectro localizado é composto por riscas espectrais podendo, por isso, ser modelizado com um número menor de sinusóides.

Para reduzir o número de parâmetros em zonas não vocalizadas, considerou-se um conjunto diferente de funções de base, mais adaptado a este tipo de sons. As funções escolhidas foram designadas funções de banda estreita por terem ainda a sua energia concentrada em torno de uma frequência central, apesar de terem largura de banda maior que no caso sinusoidal. As funções de banda estreita foram definidas como sinusóides moduladas em amplitude por envolventes estocásticas passa-baixo; a sua largura de banda, fica portanto especificada a partir da largura de banda da envolvente estocástica. Modelando a fala não-vocalizada pela sobreposição de funções de banda estreita consegue-se uma redução considerável do número de funções de base e portanto, do número de parâmetros do modelo, mantendo uma qualidade alta na fala sintetizada. Foram realizadas algumas experiências de codificação de fala não-vocalizada utilizando funções de banda estreita. Embora o algoritmo testado não tenha sido completamente codificado, os resultados obtidos confirmam a utilidade das funções de banda estreita na codificação da fala não-vocalizada com alta qualidade, a ritmos baixos.

Como aplicação deste estudo realizado acerca da modelização sinusoidal da fala, analisou-se o problema de codificação a ritmos médios e baixos e desenvolveu-se um codificador sinusoidal a 4.8 Kb/s. O modelo utilizado foi o modelo sinusoidal básico com frequências múltiplas e segmentos de síntese de 20 ms. O aspecto mais delicado na codificação dos parâmetros do modelo sinusoidal é a codificação das fases.

Ao contrário do que habitualmente se pensa acerca da importância perceptual da fase, traduzido na chamada lei de Ohm: "o ouvido é surdo à fase", é grande a distorção introduzida no sinal sintético pela codificação da fase a ritmos baixos. A codificação da fase foi, por isso, objecto de estudo. Nas zonas vocalizadas, exploram-se técnicas de predição de fase, tendo-se codificado os erros de predição de fase com um quantificador adaptativo directo. O esquema resultante corresponde à utilização de um codificador APC aplicado a cada uma das fases. Nas zonas não vocalizadas, considerou-se de grande importância manter a estrutura aleatória da fase mesmo quando as fases não podem ser codificadas com precisão.

Assim, mesmo quando o número de bits disponíveis é insuficiente para representar com precisão cada uma das fases das sinusóides, a fase quantificada não perde a sua estrutura aleatória. Os resultados de codificação a 4.8 Kb/s revelam um bom desempenho do codificador em algumas das fases do conjunto de teste, em que se

obteve qualidade alta, mas mostram também distorções fortes em algumas das frases processadas. Estas distorções parecem estar localizadas em zonas não estacionárias do sinal de fala onde o modelo sinusoidal com segmentos de síntese de 20 ms, representa algumas dificuldades de modelização que são depois agravadas pela codificação dos parâmetros das sinusóides. Um estudo detalhado das técnicas de modelização e codificação em zonas não estacionárias revela-se indispensável para tornar a codificação sinusoidal competitiva com os melhores codificadores actualmente existentes para ritmos médios e baixos.

8 - PARA FUTURO

O âmbito deste trabalho resumiu-se praticamente a uma pesquisa e estudo bibliográficos, daí que muitas das experiências referidas e estudadas, tiveram como berço o trabalho de alguns autores. Relativamente ao meu trabalho fica a certeza do esforço desenvolvido e a mensagem escassa de um assunto, de per si, bastante extenso, complexo e interessante.



FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

BIBLIOTECA



0000101582