AUGUSTO HENRIQUE HENTZ

COMPRESSÃO DE BANCOS DE FALA PARA SISTEMAS DE SÍNTESE CONCATENATIVA DE ALTA QUALIDADE

FLORIANÓPOLIS

2009

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

COMPRESSÃO DE BANCOS DE FALA PARA SISTEMAS DE SÍNTESE CONCATENATIVA DE ALTA QUALIDADE

Dissertação submetida à
Universidade Federal de Santa Catarina
como parte dos requisitos para a
obtenção do grau de Mestre em Engenharia Elétrica

AUGUSTO HENRIQUE HENTZ

Florianópolis, Setembro de 2009

COMPRESSÃO DE BANCOS DE FALA PARA SISTEMAS DE SÍNTESE CONCATENATIVA DE ALTA QUALIDADE

Augusto Henrique Hentz

'Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, Área de Concentração em *Comunicações e Processamento de Sinais*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.'

Prof. Rui Seara, Dr.
Orientador

Prof. Roberto de Souza Salgado, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

Prof. Rui Seara, Dr.
Presidente

Prof. Sidnei Noceti Filho, Dr.

Prof. Fernando Santana Pacheco, Dr.

Agradecimentos

A meus pais, Paulo e Maria, e minha irmã Isabel, pelo apoio e incentivo sempre presentes.

À Fabiana, pelo carinho.

Ao Prof. Rui, pela orientação e contribuições para o sucesso deste trabalho.

Aos membros da banca, Prof. Fernando e Prof. Sidnei, por suas valiosas contribuições.

Aos amigos do LINSE, pelas horas de convivência agradável.

À CAPES, pelo apoio financeiro.

A todos os que de alguma forma contribuíram para esse trabalho.

Resumo da Dissertação apresentada à UFSC como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica

COMPRESSÃO DE BANCOS DE FALA PARA SISTEMAS DE SÍNTESE CONCATENATIVA DE ALTA QUALIDADE

Augusto Henrique Hentz

Setembro/2009

Orientador: Rui Seara, Dr.
Área de Concentração: Comunicações e Processamento de Sinais.
Palavras-chave: *Codec* iLBC, compressão de sinais de fala, conversão texto-fala, LSFs.
Número de Páginas: 52.

RESUMO: Sistemas de conversão texto-fala baseados na técnica de síntese concatenativa com seleção de unidades são capazes de produzir fala sintética de muito boa qualidade, com inteligibilidade e naturalidade próximas às da fala humana. Para conseguir tal feito, é necessário o uso de bancos de fala contendo exemplos de diversos contextos fonéticos e prosódicos. Freqüentemente, os bancos utilizados em sintetizadores de muito boa qualidade têm duração de dezenas de horas, tornando sua ocupação de memória muito elevada. Além dos bancos de gravações, sistemas de síntese concatenativa utilizam um conjunto de informações para o cálculo de custos no processo de seleção de unidades, contribuindo para a ocupação de memória. O presente trabalho apresenta técnicas para reduzir a ocupação de memória de sistemas de síntese concatenativa de fala, considerando o sintetizador desenvolvido no LINSE (Laboratório de Circuitos e Processamento de Sinais do Departamento de Engenharia Elétrica da UFSC). O banco de gravações do sistema considerado é compactado utilizando o *codec* iLBC, que proporciona a capacidade de acesso aleatório aos dados codificados, fundamental para a aplicações em síntese de fala concatenativa. O banco de parâmetros, por sua vez, é compactado usando quantização vetorial dos coeficientes espectrais no processo de seleção de unidades. As técnicas propostas permitem reduzir a ocupação de memória do sistema considerado em até 79 %, sem grandes perdas na qualidade da fala sintética.

Abstract of Dissertation presented to UFSC as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

COMPRESSION OF SPEECH CORPORA FOR HIGH-QUALITY CONCATENATIVE SPEECH SYNTHESIS

Augusto Henrique Hentz

September/2009

Advisor: Rui Seara, Dr.
Area of Concentration: Communications and Signal Processing.
Keywords: iLBC codec, speech compression, text-to-speech (TTS) conversion, LSFs.
Number of Pages: 52.

ABSTRACT: Unit-selection concatenative text-to-speech (TTS) systems can produce very-high-quality synthetic speech with intelligibility and naturalness close to that of human speech. To achieve such a quality, these systems have to use speech recording databases with many different phonetic and prosodic contexts. Recording databases used in very-high-quality TTS systems often have tens of hours of duration, causing the memory usage of such systems to be very high. Besides the speech recordings, concatenative TTS systems use a parameter database containing data to evaluate costs in the unit selection process, which adds to the system memory usage. This work presents techniques to reduce the memory footprint of concatenative speech synthesizers, focusing on the TTS system developed at LINSE (Circuits and Signal Processing Laboratory of the Federal University of Santa Catarina). System speech recording database is compressed with the iLBC codec, which provides random access capability to the compressed data that is a fundamental feature in concatenative speech synthesis applications. In turn, the feature database is compressed by vector quantizing the spectral coefficients used in the unit selection process. The proposed techniques provide up to a 79 % reduction in memory usage, with small loss in synthetic speech quality.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

— Meu Deus, ele fala!

Disse, espantado, Dom Pedro II, quando apresentado ao telefone. O imperador logo veio a saber que o telefone, na verdade, não falava, mas esse episódio ilustra bem o espanto e o fascínio causados por máquinas capazes de falar. O desejo de criar tais dispositivos é antigo. Já no século XVIII, aparelhos mecânicos capazes de criar sons semelhantes à fala humana foram desenvolvidos. Na primeira metade do século XX, foram concebidos sintetizadores de fala eletro-eletrônicos. Esses primeiros sistemas eram operados por usuários experientes, ajustando um conjunto de controles, de modo análogo ao uso de instrumentos musicais. Com a evolução dos computadores digitais, o foco passou a ser o desenvolvimento de algoritmos computacionais para a síntese de fala. Atualmente, existem sistemas de conversão texto-fala (*text-to-speech* – TTS) capazes de converter texto irrestrito em fala sintética de qualidade semelhante à fala natural.

A fala é um dos métodos de comunicação mais naturais entre os seres humanos. Por isso, equipamentos capazes de se comunicar com seus usuários através da fala despertam grande interesse. Além do mais, há situações em que a comunicação verbal de mensagens é mais conveniente do que a escrita, ou mesmo a única forma viável. Dentre as inúmeras aplicações de sistemas de conversão texto-fala, pode-se destacar:

1. Auxílio para deficientes visuais.
2. Auxílio a pessoas com problemas de fala.
3. Automação de acesso a informações por telefone (saldos bancários, horários de filmes, etc.).
4. Passar informações auxiliares em situações em que não se pode desviar a atenção de uma tarefa principal. Por exemplo, sistemas de posicionamento global (*global positioning system* – GPS) em automóveis.
5. Tornar a interação com máquinas mais amigável.

Para que sistemas de conversão texto-fala possam ser empregados com sucesso, alguns requisitos básicos devem ser observados. Primeiramente, as mensagens produzidas pelo sistema devem ser prontamente entendidas pelos usuários. Além do mais, a fala sintética deve ter características semelhantes à fala humana natural. Essas características são denominadas, respectivamente, inteligibilidade e naturalidade, sendo os principais aspectos da qualidade de sistemas de conversão texto-fala.

1.1 Justificativa e Objetivos

Sistemas TTS atuais são capazes de produzir fala sintética de muito boa qualidade, com inteligibilidade e naturalidade muito próximas às da fala humana. A grande maioria desses sistemas se baseia na técnica de síntese concatenativa, produzindo o sinal de fala sintético através da concatenação de trechos de gravações extraídos de um banco de dados segmentado e transcrito foneticamente. Em geral, bancos de maior duração contêm maior diversidade de contextos fonéticos, levando à produção de fala sintética de melhor qualidade. Isso faz com que sistemas com muito boa qualidade possuam bancos de grande duração, tornando sua ocupação de memória muito elevada. Por exemplo, o Orador, sistema TTS desenvolvido no LINSE e objeto de estudo deste trabalho, possui um banco de gravações de 28 horas que, amostrado a 8 kHz e quantizado não-linearmente através da curva da lei A [1], ocupa 800 MB. Associado ao banco de gravações, há um conjunto de parâmetros adicionais, fazendo com que a ocupação de memória total do sistema seja de aproximadamente 1,6 GB em disco e 750 MB de memória principal.

Embora esse nível de ocupação de memória não inviabilize o uso de sistemas TTS de muito boa qualidade em computadores atuais, ainda há interesse em desenvolver técnicas para reduzi-lo. Atualmente, o uso de sistemas TTS de estado-da-arte é comum apenas em equipamentos dedicados, como em dispositivos de resposta vocal para telefonia. Reduzindo o uso de memória desses sistemas, seria possível utilizá-los em computadores pessoais de baixo custo ou até mesmo em sistemas embarcados.

Uma alternativa para reduzir a ocupação de memória de sistemas TTS (mantendo a diversidade de contextos fonéticos e prosódicos) é através da compressão do banco de gravações, utilizando técnicas de codificação de fala. No entanto, a compressão de bancos de gravações visando síntese de fala apresenta algumas características particulares que a distinguem das aplicações mais usuais da codificação de fala, tais como sistemas de comunicação em tempo real. A característica mais importante da aplicação considerada é que, no processo de síntese, o banco de gravações é acessado de maneira aleatória. Desse modo, o algoritmo utilizado para compressão deve permitir a decodificação parcial de

qualquer trecho do banco.

Algoritmos atuais de codificação de fala, baseados em predição linear e análise por síntese, permitem a compressão eficiente do sinal de fala, reduzindo muito a taxa de bits necessária para representá-lo sem prejudicar significativamente sua qualidade. Para tal, grande parte desses algoritmos exploram correlações de longo termo do sinal de fala. Uma abordagem bastante comum é o uso de *codebook* adaptativo, que permite considerar excitações passadas na composição da excitação do quadro corrente. No entanto, é necessário manter os estados internos do codificador e do decodificador sincronizados, para que o sinal possa ser decodificado sem distorções. Isso significa que, caso o sinal comece a ser decodificado a partir de um dado quadro (como ocorre no acesso aleatório), haverá grandes distorções até que os estados do decodificador e do codificador sejam sincronizados. Desse modo, *codecs* que tiram proveito de correlações entre quadros consecutivos não são adequados quando for necessário acessar os dados codificados de maneira aleatória, tornando-se inviáveis para compressão de um banco de gravações para síntese de fala.

Para comprimir um banco de gravações visando síntese de fala, Vrecken et. al. [2] propõem um codificador que utiliza um *codebook* adaptativo e diversos *codebooks* estocásticos. No início de cada segmento, para reduzir as distorções causadas pela falta de sincronismo entre codificador e decodificador, utiliza-se maior contribuição dos *codebooks* estocásticos. No entanto, ainda há distorções consideráveis no início de cada segmento. Por sua vez, Lee et. al. [3] propõem substituir o *codebook* adaptativo do *codec* G.729 por um *codebook* fixo de sinais de excitação, treinado com gravações do próprio banco do sistema de síntese. O codificador proposto apresenta qualidade semelhante à do *codec* G.729 para situações em que as gravações são acessadas de maneira contínua. Embora esse codificador permita acessar os dados codificados de maneira aleatória, não há resultados perceptuais avaliando a degradação causada nessa situação.

Neste trabalho, propõe-se codificar o banco de gravações de sistemas TTS utilizando o *codec* iLBC (*internet low bit rate codec*) [4], [5], desenvolvido especificamente para ser robusto à perda de quadros comum em aplicações de transporte de voz sobre protocolo internet (*voice over internet protocol – VoIP*). Uma característica importante deste *codec* é a independência entre quadros, permitindo minimizar as distorções decorrentes da perda de pacotes, podendo ser usado com vantagem em sistemas TTS.

Além do banco de gravações, o sistema TTS considerado possui um conjunto de informações utilizadas no processo de seleção de unidades para a síntese. Esse conjunto, denominado banco de parâmetros, também ocupa grande quantidade de memória, principalmente devido a parâmetros espectrais armazenados para cada unidade (fonema) do

banco de gravações. As informações contidas nesse banco devem ser mantidas em memória principal (ao invés de simplesmente armazenadas em disco) para um melhor desempenho do sistema. Desse modo, a compressão do banco de parâmetros traz grandes benefícios para o desempenho do sistema TTS, já que o custo da memória principal é muito maior do que o custo de armazenamento em disco. Sendo assim, outro objetivo de grande importância desse trabalho é reduzir a ocupação de memória do banco de parâmetros.

1.2 Organização do Trabalho

Este trabalho é organizado como segue. No Capítulo 2, se discute brevemente o processo de produção da fala humana. São abordados aspectos fisiológicos do aparelho fonador humano, bem como modelos matemáticos desenvolvidos para compreender e simular a produção da fala. Aspectos de conversão texto-fala são abordados no Capítulo 3. Funções de processamento lingüístico são brevemente discutidas, assim como técnicas de síntese de fala baseadas em simulação do aparelho fonador humano. A abordagem de síntese concatenativa, largamente utilizada em sistemas de conversão texto-fala atuais, é discutida em mais detalhes. O Capítulo 4 discute a codificação de fala. O conceito de predição linear é apresentado e discutido no contexto de codificação de fala. A representação LSF dos coeficientes de predição linear e suas propriedades são descritas. Aspectos fundamentais do *codec* iLBC são também apresentados. As técnicas propostas para reduzir a ocupação de memória de sistemas TTS baseados em síntese concatenativa são discutidas no Capítulo 5. São apresentadas técnicas para reduzir a ocupação de memória de bancos de parâmetros e de concatenação. O resultado das técnicas propostas é apresentado e comentado no Capítulo 6. São analisadas a capacidade de redução de memória proporcionada pelas técnicas propostas e seu efeito na qualidade da fala sintética. As considerações finais e propostas para futuros trabalhos são apresentadas no Capítulo 7.

2 Produção da Fala Humana

A fala humana consiste de um sinal acústico não-estacionário, produzido pela ação coordenada dos órgãos do aparelho fonador. Aspectos relacionados ao processo de produção da fala humana e de modelos matemáticos desse processo são apresentados neste capítulo, baseando-se nas referências [6], [7] e [8].

O aparelho fonador, ilustrado na Fig. 2.1, é composto por órgãos dos sistemas respiratório, fonador e articulatório. Os órgãos do sistema respiratório são responsáveis por criar o fluxo de ar que gera o sinal de fala. Esse fluxo é produzido nos pulmões e levado pela traquéia até a laringe.

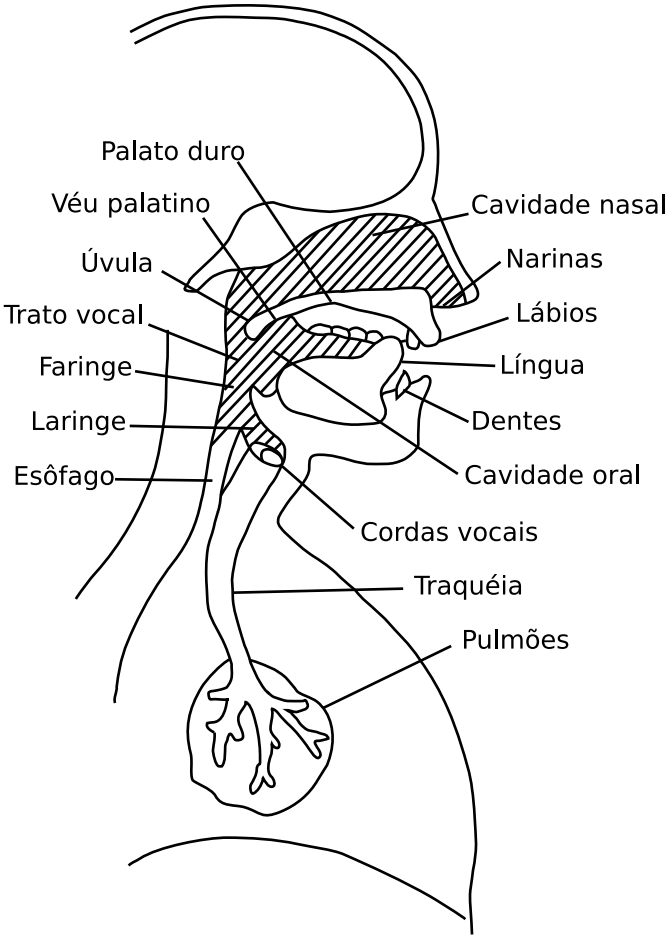

Fig. 2.1: Órgãos do aparelho fonador.

Na laringe, órgão do sistema fonador, estão localizadas as cordas vocais, um par de músculos estriados que alteram as características do fluxo aerodinâmico. Quando as cordas vocais formam uma abertura estreita, o fluxo de ar proveniente dos pulmões as faz vibrar, gerando pulsos aerodinâmicos periódicos. Esses pulsos glotais são responsáveis pela criação dos chamados sons vozeados. Por outro lado, quando a glote (abertura formada pelas cordas vocais) mantém-se levemente aberta, o fluxo de ar proveniente dos pulmões não é mais periódico, adquirindo características ruidosas. Os sons gerados nessa situação são chamados não-vozeados.

O fluxo de ar glotal passa pelos órgãos do sistema articulatório, composto pelo trato vocal e pela cavidade nasal. O trato vocal, que inicia na glote e termina nos lábios, pode ser interpretado como um tubo de seção não-uniforme e variável ao longo do tempo. O efeito do trato vocal sobre o fluxo de ar é de filtragem acústica, enfatizando algumas freqüências do sinal de excitação e atenuando outras. A amplificação de uma freqüência é chamada ressonância e as ressonâncias do trato vocal são chamadas formantes. Através da movimentação da língua e dos lábios, a resposta em freqüência do trato vocal pode ser alterada de diversas maneiras, permitindo a produção de diferentes tipos de sons.

Finalmente, o sinal de fala é irradiado para o ambiente. As ondas acústicas podem ser irradiadas pelos lábios (sons orais), pelas narinas (sons nasais) ou por ambos (sons nasalizados). A cavidade nasal é acoplada ao trato vocal através do véu palatino. Quando o véu palatino é erguido, a cavidade nasal é bloqueada e o som é radiado somente pelos lábios. O acoplamento da cavidade nasal também altera a resposta em freqüência do trato vocal.

A produção da fala pode ser interpretada como um processo de criação de sinais de excitação e filtragem acústica. Sendo assim, é possível desenvolver modelos matemáticos desse processo. Uma visão geral do chamado modelo fonte-filtro de produção de fala é apresentada na seqüência.

2.1 Modelo Fonte-Filtro de Produção de Fala

Como mencionado anteriormente, a fala humana é produzida pela filtragem acústica de uma fonte sonora e uma característica de radiação para o ambiente. Como o sinal de fala não é estacionário, a modelagem deve ser variante no tempo. O sinal de fala pode ser considerado estacionário quando analisado em quadros de curta duração (até 30 ms). Dessa maneira, é possível modelar a produção de fala como um processo invariante por partes. Na seqüência, são apresentados aspectos da modelagem da produção de fala, considerando quadros de curta duração. São discutidos aspectos da produção do sinal de

excitação, da filtragem pelo trato vocal e da característica de radiação para o ambiente.

Fonte sonora. Para a criação de sons não-vozeados, a fonte sonora pode ser simulada por um gerador de ruído branco. Já para criar os sons vozeados, é necessário produzir sinais simulando os pulsos glotais. O sinal de excitação vozeado pode ser obtido por modelos paramétricos dos pulsos glotais, como os modelos de Liljencrants-Fant e Rosemberg-Klatt [9]. Também é possível simular a excitação vozeada aplicando um trem de impulsos periódico a um filtro $G(z)$ que simule o efeito dos pulsos glotais.

Trato vocal. O trato vocal pode ser modelado pela associação em cascata de um conjunto de tubos uniformes, com diferentes seções transversais. Cada tubo, correspondente a um formante, é caracterizado como um ressoador acústico de segunda ordem. O sistema completo é descrito por

$$V(z) = \frac{G}{\prod_{i=1}^{N}(1 - p_i z^{-i})} \tag{2.1}$$

onde p_i são os N pólos do sistema e G é um fator de ganho. O modelo composto apenas por pólos representa adequadamente a resposta do trato para a criação de sons vozeados. No entanto, na produção de sons nasais e nasalizados, a resposta do trato vocal apresenta zeros. Para modelar adequadamente essas situações, é necessário que o número de pólos do sistema seja maior do que o número de formantes. Considerando uma taxa de amostragem de 8 kHz, a resposta do trato vocal pode ser adequadamente representada por 4 a 6 ressonâncias (8 a 12 pólos).

Característica de radiação. A característica de radiação para o ambiente pode ser modelada como um sistema derivador $R(z)$. Assim,

$$R(z) = 1 - z^{-1}. \tag{2.2}$$

Modelo completo. O modelo completo de produção de fala, ilustrado na Fig. 2.2 é obtido associando os modelos de fonte sonora, trato vocal e radiação para o ambiente. Variando os parâmetros desse modelo, é possível simular as características não-estacionárias do sinal de fala real.

Modelo simplificado. É possível associar os modelos de simulação do pulso glotal, do trato vocal e da radiação em um único filtro discreto $H(z)$. Assim,

$$H(z) = G(z) \cdot V(z) \cdot R(z) \tag{2.3}$$

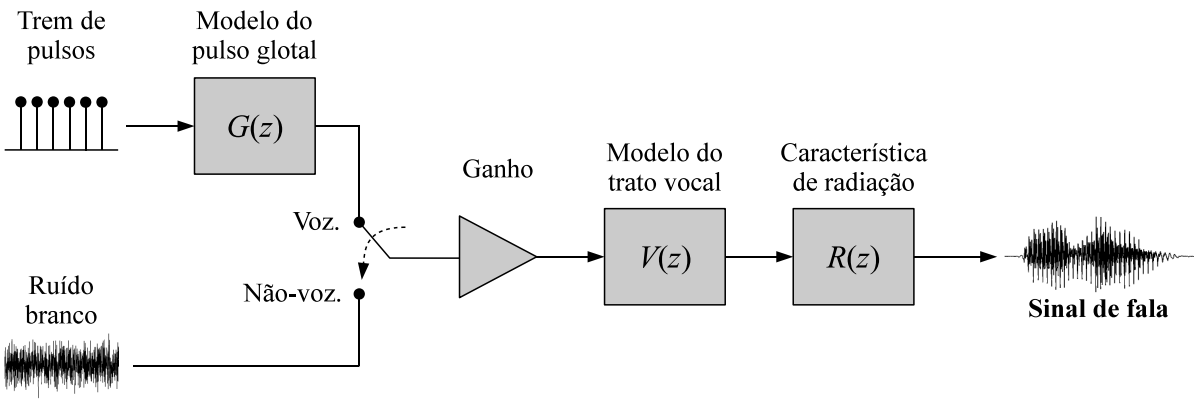

Fig. 2.2: Modelo completo de produção de fala.

onde $G(z)$ é o filtro simulando o efeito dos pulsos glotais, $V(z)$ é a resposta do trato vocal e $R(z)$ é a característica de radiação para o ambiente. Com tal simplificação, a fonte de excitação se torna apenas um trem de impulsos para a geração de sons vozeados e um gerador de ruído branco para a produção de sons não-vozeados. O modelo simplificado é mostrado na Fig. 2.3.

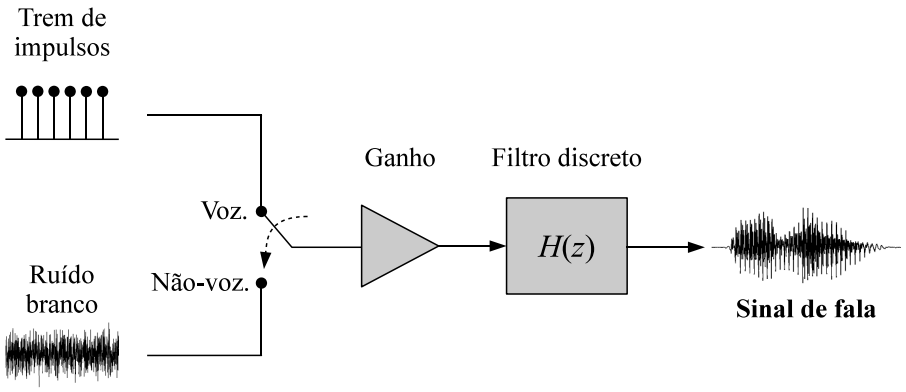

Fig. 2.3: Modelo simplificado de produção de fala.

Modelo auto-regressivo. Essa abordagem introduz uma simplificação adicional: a fonte de excitação é unicamente um gerador de ruído branco, representando o fluxo de ar proveniente dos pulmões e toda a informação espectral do sinal de fala é gerada por um processo auto-regressivo [10]. Caso a ordem desse processo seja suficientemente elevada, é possível gerar a excitação periódica para os sons vozeados a partir do ruído branco. O modelo auto-regressivo de produção da fala é ilustrado na Fig. 2.4.

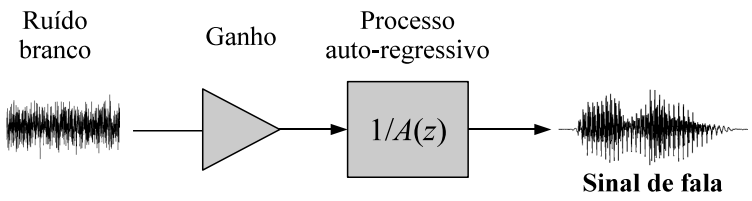

Fig. 2.4: Modelo auto-regressivo de produção de fala.

2.2 Sumário e Comentários

A fala humana é produzida pela ação dos órgãos do aparelho fonador, em um processo de filtragem acústica de um sinal de excitação. A fonte de sinal é o fluxo de ar proveniente dos pulmões, que adquire características de ruído branco ou trem de pulsos periódicos ao passar pela glote. Esse sinal acústico tem sua resposta em freqüência alterada pelo trato vocal, sendo então radiado para o ambiente pelos lábios e/ou pelas narinas.

O processo de produção de fala pode ser modelado pelo chamado modelo fonte-filtro. Esse modelo considera que a fala é produzida pela filtragem de um sinal de excitação. Para a produção de sons não-vozeados, a excitação é gerada por um ruído branco. Já para a produção de sons vozeados, a excitação é criada por um trem de impulsos periódico. Um filtro discreto modela a resposta em freqüência do trato vocal e a característica de radiação para o ambiente. O processo de produção de fala é considerado estacionário por partes, ou seja: para quadros de curta duração, o sinal de excitação é estacionário e a resposta do filtro é invariante. A característica não-estacionária do sinal de fala real é obtida variando os parâmetros do modelo para cada quadro.

O modelo fonte-filtro de produção de fala é bastante utilizado em diversas aplicações de processamento de fala. Esse modelo é a base de algumas abordagens de síntese de fala, discutidas na Seção 3.2.1, e de grande parte dos algoritmos de codificação de fala, apresentados no Capítulo 4.

3 Conversão Texto-Fala

O objetivo final de um sistema de conversão texto-fala (*text-to-speech* – TTS) é transformar qualquer texto em fala sintética que tenha inteligibilidade e naturalidade muito próximas às da fala humana. Em um sistema TTS ideal, os usuários devem ser levados a crer que um locutor está lendo o texto apresentado ao sistema. Para que esse processo complexo possa ser realizado de maneira satisfatória, é necessário uma divisão em etapas, como ilustrado na Fig. 3.1 [6]. A primeira etapa, descrita na Seção 3.1, é o processamento lingüístico, responsável por converter o texto em uma representação estruturada, descrevendo detalhadamente os sons que devem ser produzidos para codificar a mensagem textual. Em seguida, as informações obtidas na etapa de processamento lingüístico são utilizadas por um módulo de síntese de fala para produzir o sinal de fala artificial. Na Seção 3.2, são apresentadas diferentes abordagens para o problema de síntese de fala, com destaque para a síntese concatenativa, que é a abordagem utilizada nos sistemas de estado-da-arte atuais.

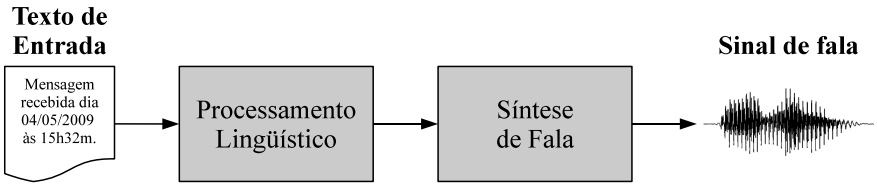

Fig. 3.1: Componentes de um sistema de conversão texto-fala.

3.1 Processamento Lingüístico

Conforme citado, o objetivo da etapa de processamento lingüístico é converter texto irrestrito em uma representação adequada para o sistema de síntese de fala. Executar essa etapa com desempenho adequado é um fator muito importante para a qualidade do sistema TTS, já que falhas nesse procedimento podem causar a impressão de que o sistema é de má qualidade por produzir palavras com pronúncia ou entonação erradas. É possível dizer que as funções de processamento lingüístico proporcionam ao sistema de conversão texto-fala a capacidade de "ler" corretamente.

O processamento lingüístico é uma tarefa complexa e fortemente ligada ao idioma utilizado pelo sistema. Desse modo, para implementar adequadamente funções que realizam essa tarefa, é necessário conhecer detalhadamente a estrutura da língua. Usualmente, o processamento do texto é dividido em etapas, como mostrado na Fig. 3.2. As etapas desse processo são detalhadas na seqüência.

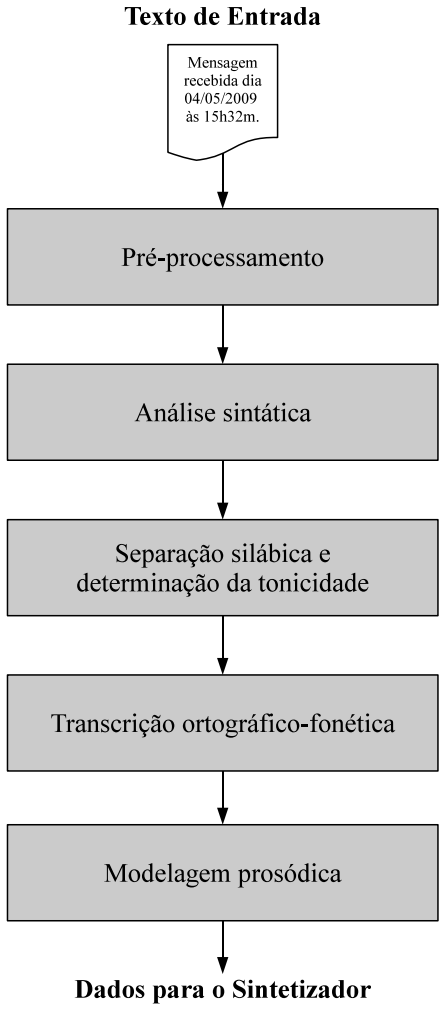

Fig. 3.2: Etapas do processamento lingüístico.

Pré-processamento. A primeira etapa do processamento lingüístico de um dado texto é sua divisão em grupos de palavras para simplificar a análise. Usualmente, as palavras do texto são agrupadas em frases, processadas de maneira independente. Em cada frase, são executadas as seguintes operações:

- Expansão de abreviaturas (por exemplo, a abreviatura "Dr." é substituída pela palavra "doutor").
- Decodificação de siglas, de acordo com seu modo usual de pronúncia (por exemplo, a

sigla “FGTS” é pronunciada como uma seqüência de letras, enquanto a sigla “UFSC” é pronunciada como uma palavra).

- Decodificação de elementos não-textuais, como, por exemplo, numerais, datas, valores monetários, endereços de *e-mail* e de sítios de internet.

- Identificação dos sinais de pontuação, para inserção de pausas nos momentos adequados.

Como resultado das operações descritas, um dado texto é convertido em um conjunto de frases, compostas unicamente por palavras e sinais de pontuação.

Análise sintática. A análise sintática é uma etapa com grande impacto na qualidade da fala sintética, pois permite distinguir palavras com mesma grafia e pronúncias diferentes, como por exemplo o verbo “almoço” (“Eu almoço em casa todos os dias.”) e o substantivo “almoço” (“O almoço estava delicioso.”). Além disso, a estrutura sintática de uma frase é bastante importante para a determinação de sua prosódia.

Separação silábica e determinação da tonicidade. Após a análise sintática, as palavras do texto são divididas em sílabas para determinar a sílaba tônica de cada palavra. Palavras contendo diacríticos (acentos gráficos) já têm sua sílaba tônica determinada. Para as demais, é necessário utilizar um conjunto de regras [11]. Além de determinar a sílaba tônica de cada palavra, é necessário classificar as palavras da frase quanto à tonicidade, enfatizando palavras de conteúdo, tais como verbos, substantivos e adjetivos.

Transcrição ortográfico-fonética. A etapa seguinte do processamento lingüístico é converter as palavras em seqüências de fonemas. Esse procedimento não é trivial, já que não há um mapeamento único entre as letras do alfabeto e os fonemas da língua. Por exemplo, uma determinada letra pode representar mais de um fonema, como nas palavras “centena” e “coelho”, nas quais a letra “c” inicial representa os fonemas /s/ e /k/, respectivamente. Além disso, um mesmo fonema pode ser representado por letras diferentes, como, por exemplo, nas palavras “senhora” e “centena”, onde o fonema /s/ inicial é representado pelas letras “s” e “c”, respectivamente. Uma abordagem bastante utilizada na língua portuguesa é a transcrição por regras, baseada no contexto das letras analisadas [11]. No entanto, há palavras que não podem ser transcritas corretamente de maneira automática. Para esses casos, é necessário usar um dicionário de exceções contendo a transcrição fonética de tais palavras.

Modelagem prosódica. A última etapa do processamento lingüístico, modelagem prosódica, tem grande impacto na naturalidade da fala sintética. A prosódia, caracterizada pela entonação e ritmo empregados na leitura de um dado texto, é determinada considerando o resultado das etapas previamente discutidas, com destaque para a identificação da pontuação, a análise sintática e a determinação da tonicidade. A prosódia desejada pode ser obtida de maneira explícita, em que a freqüência fundamental, duração e intensidade dos fonemas são manipulados segundo um modelo de geração de prosódia, ou de maneira implícita, em sistemas de síntese concatenativa com bancos de longa duração. Tais sistemas, discutidos na Seção 3.2.2, possibilitam reproduzir as características prosódicas da fala do locutor presente no banco de gravações sem manipulações adicionais no sinal de fala.

3.2 Síntese do Sinal de Fala

O sinal de fala artificial é produzido por um sintetizador de fala, com base em uma representação detalhada dos sons que codificam um dado texto, obtida na etapa de processamento lingüístico. É possível destacar duas abordagens distintas para a síntese de fala: síntese baseada em simulação do trato vocal e síntese concatenativa. As duas abordagens são discutidas nas seções subseqüentes.

3.2.1 Síntese Baseada em Simulação do Trato Vocal

Nesta abordagem, a fala sintética é produzida simulando o processo de produção da fala humana. Comum a todas as técnicas baseadas nessa abordagem é o uso do modelo fonte-filtro de produção de fala, apresentado na Seção 2.1. As principais técnicas baseadas nessa abordagem são a síntese por formantes e a síntese articulatória, discutidas a seguir.

3.2.1.1 Síntese por Formantes

A abordagem de síntese por formantes modela a resposta em freqüência do trato vocal pela associação de um conjunto de ressoadores de segunda ordem (filtros caracterizados por dois pólos), representando as ressonâncias do trato vocal (formantes). Alterando a freqüência e a largura de banda de cada formante ao longo do tempo, é possível simular a característica não estacionária do trato vocal. A fala sintética é então obtida aplicando um sinal de excitação conveniente a esse sistema variante no tempo.

Os ressoadores podem ser associados em cascata, em paralelo ou em uma estrutura híbrida, combinando as duas associações. Uma implementação bastante conhecida da síntese por formantes é o sintetizador de Klatt [12]. Esse sintetizador possui 5 ressoadores

responsáveis pela simulação do trato vocal, que podem ser associados em cascata ou em paralelo, de acordo com um parâmetro de configuração. Há ainda um ressoador e um anti-ressoador (sistema inverso ao ressoador, contendo zeros ao invés de pólos) para a síntese de sons nasais ou nasalizados. Tanto a excitação vozeada quanto a excitação não vozeada podem ser especificadas detalhadamente através de um conjunto de parâmetros. Finalmente, um diferenciador de primeira ordem simula a característica de radiação para o ambiente. O sintetizador de Klatt possui, no total, 39 parâmetros de controle, atualizados a cada 5 ms.

A síntese por formantes, com o ajuste apropriado dos parâmetros, permite a produção de fala sintética de muito boa qualidade. Experimentos chamados de síntese por cópia (*copy synthesis*) eram comumente realizados para avaliar a qualidade de um dado sintetizador. Nesses experimentos, um operador experiente ajusta cuidadosamente os parâmetros do sistema para reproduzir uma dada gravação. Freqüentemente, a fala sintética obtida pelo processo de síntese por cópia é indistinguível da gravação original. A flexibilidade dessa técnica também é um ponto importante. Como se tem o controle completo do processo de produção da fala sintética, é possível fazer com que o sistema produza diferentes vozes ou, ainda, uma dada voz com diferentes características (por exemplo, com ênfase ou suspirando). Sistemas de conversão texto-fala baseados na técnica de síntese por formantes, em geral, produzem fala sintética altamente inteligível, mas com naturalidade pobre. Isso ocorre pela dificuldade de realizar o mapeamento da especificação dos sons a serem produzidos (determinada pelas funções de processamento lingüístico) para os parâmetros de controle do sintetizador. O método clássico de realizar esse mapeamento, que consiste em um conjunto de regras desenvolvidas por especialistas, foi praticamente abandonado. No entanto, há pesquisas recentes visando determinar automaticamente os parâmetros de controle do sintetizador através da especificação obtida pelas funções de processamento lingüístico [13], [14]. Ainda assim, a síntese por formantes foi praticamente abandonada no desenvolvimento de sistemas de conversão texto-fala comerciais.

3.2.1.2 Síntese Articulatória

A síntese articulatória é baseada na simulação direta do aparelho fonador humano. É potencialmente a melhor técnica para geração de fala sintética, por permitir o controle total sobre todos os aspectos de sua produção. No entanto, é uma técnica muito complexa, exigindo conhecimento detalhado de todo o processo de produção da fala humana.

O sinal de fala é criado simulando a passagem do fluxo de ar oriundo da glote por um tubo de seção não-uniforme, representando o trato vocal. Alterando a excitação e a forma do tubo, é possível sintetizar todos os sons produzidos pelo aparelho fonador humano. De

maneira semelhante à síntese por formantes, a maior dificuldade da síntese articulatória consiste em determinar os parâmetros de controle do sintetizador a partir da especificação derivada do texto.

Atualmente, a abordagem de síntese articulatória não é ainda uma alternativa viável para produzir fala sintética de boa qualidade. Contudo, é uma ferramenta de pesquisa poderosa em estudos de produção de fala, por permitir a simulação detalhada do processo de produção da fala humana [8].

3.2.2 Síntese Concatenativa

Em contraste com as abordagens discutidas na seção anterior, a técnica de síntese concatenativa tem por objetivo produzir fala sintética concatenando trechos de gravações. Para tal, é necessário um conjunto de gravações transcritas e segmentadas em unidades. Denomina-se **unidade** o menor segmento de fala utilizado pelo sistema de síntese, em geral fonema ou difone. É possível sintetizar qualquer seqüência de palavras utilizando um banco de dados contendo apenas um exemplo de cada unidade da língua. Para obter resultados satisfatórios, é necessário que as unidades utilizadas sejam extraídas de contextos fonéticos neutros, de modo a reduzir efeitos de coarticulação indesejáveis no processo de síntese. Além do mais, as unidades não podem ser pronunciadas isoladamente, sendo necessário o uso de frases veículo para construção do banco. As frases veículo contém palavras em que as unidades desejadas para a criação do banco são pronunciadas de maneira neutra e estável, com reduzido efeito de coarticulação.

Embora um exemplo de cada unidade seja suficiente para sintetizar um dado conjunto de palavras, a fala sintética sem a prosódia esperada não é natural. Sendo assim, é necessário utilizar alguma técnica para modificar os parâmetros prosódicos (*pitch* e duração) das unidades utilizadas na síntese. Uma técnica bastante utilizada é a PSOLA (*pitch-synchronous overlap-and-add*) [15], que permite alterar adequadamente o contorno de *pitch* e duração da fala sintética. Essa técnica consiste em duas etapas: os períodos do sinal de fala são identificados e isolados, sendo então recombinados de maneira conveniente para obter a duração e o *pitch* desejados.

O uso de apenas um exemplo de cada unidade para a síntese está associado a uma premissa: toda a variabilidade prosódica da fala pode ser simulada alterando o *pitch* e a duração de um pequeno conjunto de unidades básicas, com prosódia neutra. Essa premissa é muito forte e é um fator que limita a qualidade da fala sintética. Por exemplo, na fala natural, um dado fonema soa diferente quando ocorre no início ou no final da frase, mesmo que a duração e a freqüência fundamental sejam idênticas nas duas situações. É possível

contornar essa limitação construindo bancos de dados contendo diferentes exemplos de cada unidade, extraídos em diferentes situações (início/fim de frase, sílaba tônica ou não, etc.). No entanto, o processo de construção de um banco de dados de unidades isoladas em um grande número de situações é complexo e ineficiente. Para construir um banco com essas características, é necessário elaborar um grande número de frases veículo que permitam extrair as unidades em diferentes contextos. No entanto, apenas um pequeno número de unidades é extraído de cada frase veículo, descartando grande parte do conteúdo das gravações. Uma abordagem mais interessante para o problema de síntese concatenativa é extrair as unidades diretamente de gravações contínuas. Desse modo, é possível elaborar um conjunto de frases em que as unidades apareçam em todas as situações desejadas, e as gravações poderiam ser aproveitadas em sua totalidade. As técnicas utilizadas para extrair unidades para síntese concatenativa de um conjunto de gravações contínuas, denominadas algoritmos de seleção de unidades, são descritas a seguir.

3.2.2.1 Seleção de Unidades

Como mencionado, é possível melhorar a qualidade da técnica de síntese concatenativa utilizando diversas unidades extraídas diretamente de gravações de fala natural, ao invés de um pequeno número de unidades isoladas. No entanto, com o aumento do número de unidades, surge o problema de escolher quais são mais apropriadas para sintetizar uma dada frase. A grande maioria dos sistemas atuais de síntese de fala aborda esse problema utilizando alguma variação do algoritmo de seleção de unidades proposto por Andrew Hunt e Alan Black [16].

A essência do algoritmo é buscar, em um banco de gravações contínuas, a seqüência de unidades mais apropriada para sintetizar um dado texto. O banco de gravações é um fator de grande importância na qualidade da fala sintética. As gravações devem ser realizadas em ambiente controlado e sem ruído por um locutor experiente, de modo a manter as características da voz consistentes em todo o banco. Além disso, é importante que as gravações contenham exemplos de todos os contextos fonéticos e variações prosódicas presentes na língua. Essas características podem ser obtidas definindo adequadamente os roteiros de gravação.

Para que unidades possam ser buscadas no banco, é necessário identificá-las e caracterizá-las com um conjunto de parâmetros. Alguns dos parâmetros usualmente empregados são de natureza acústica, como a freqüência fundamental, a duração ou alguma representação compacta do conteúdo espectral (por exemplo, coeficientes de predição linear ou parâmetros mel-cepstrais). As unidades também são caracterizadas por parâmetros contextuais, dentre os quais se destacam:

- Suas unidades vizinhas.
- O tipo de sílaba em que a unidade está inserida (por exemplo, Vogal-Consoante ou Consoante-Vogal-Consoante).
- A posição dessa sílaba na palavra.
- A classe gramatical da palavra em que a unidade está inserida.
- A posição dessa palavra na frase.

Usando as informações obtidas pelas funções de processamento lingüístico, é determinada uma seqüência alvo $T = \{t_1, \ldots, t_n\}$ descrevendo as n unidades necessárias para a síntese. Então, a seqüência de unidades mais apropriada, $\hat{U} = \{\hat{u}_1, \ldots, \hat{u}_n\}$, é buscada no banco. A seqüência $\hat{U}$ é aquela que minimiza uma dada função custo, ou seja

$$\hat{U} = \arg\min_{U} C(T,U) \tag{3.1}$$

onde

$$C(T,U) = \sum_{i=1}^{n} C_{\mathrm{t}}(t_i,u_i) + \sum_{i=2}^{n} C_{\mathrm{c}}(u_{i-1},u_i) \tag{3.2}$$

e C_{t} e C_{c} são, respectivamente, as funções custo alvo e custo de concatenação. O custo alvo é uma estimativa da semelhança entre um alvo t_i e uma unidade candidata u_i, e o custo de concatenação avalia a qualidade da junção de duas unidades candidatas consecutivas. Essas funções são discutidas em detalhes a seguir.

O custo alvo é composto pela soma ponderada de um conjunto de custos parciais, levando em consideração diferenças acústicas e contextuais entre as unidades alvo e candidata. Como uma unidade alvo consiste de um conjunto de informações extraídas do texto, os únicos parâmetros acústicos válidos para caracterizá-la são aqueles que podem ser obtidos por funções de modelagem prosódica: *pitch* e duração. Sendo assim, observa-se que o custo alvo é fortemente influenciado por parâmetros contextuais. De fato, é possível obter bons resultados sem empregar quaisquer parâmetros acústicos no cálculo do custo alvo. A quantidade de parâmetros utilizados para caracterizar as unidades do banco é um fator importante. Embora um número elevado de parâmetros possibilite especificar as unidades alvo com elevada precisão, podem ocorrer situações em que o banco não contenha unidades cujos parâmetros sejam iguais aos de um dado alvo. A função custo alvo deve ser desenvolvida considerando essa particularidade.

Ao contrário do custo alvo, o custo de concatenação é uma medida utilizada para comparar duas unidades do banco. Sendo assim, é possível utilizar informações acústicas detalhadas, tais como coeficientes que caracterizem o conteúdo espectral das unidades.

Embora também seja possível usar parâmetros contextuais, o custo de concatenação é usualmente uma medida acústica. As funções custo alvo e de concatenação utilizadas no sistema de conversão texto-fala considerado nesse trabalho são descritas na Seção 5.1.

De maneira geral, é inviável calcular a função custo para todas as seqüências fonéticas contidas no banco. Para reduzir o espaço de busca, é necessário adotar algum procedimento sub-ótimo. A escolha desse método influi de maneira importante na qualidade da fala sintética.

A técnica de seleção de unidades associada a bancos de longa duração adequadamente construídos possibilita criar fala sintética com boa naturalidade sem a necessidade de alterações de parâmetros prosódicos no sinal de fala. O algoritmo de seleção de unidades permite encontrar segmentos em contextos muito próximos aos necessários para a síntese, reproduzindo assim o estilo de fala do locutor presente nas gravações. Nessas situações, alterações prosódicas podem até degradar a qualidade da fala sintética.

3.3 Sumário e Comentários

O processo de converter texto em fala pode ser dividido em duas etapas principais: o processamento lingüístico, responsável por obter uma representação lingüística detalhada do texto e a síntese de fala, que cria um sinal de fala artificial a partir dessa representação.

A síntese do sinal de fala pode ser realizada usando modelos de produção da fala humana ou através da concatenação de segmentos de gravações. A primeira abordagem, na forma de síntese por formantes, foi muito utilizada em sistemas comerciais nas décadas de 1970 e 1980, mas caiu em desuso devido à dificuldade de criação de regras para determinar os parâmetros de síntese a partir do texto. Atualmente, os sistemas de conversão texto-fala de estado-da-arte empregam a técnica de síntese concatenativa de unidades extraídas de bancos de gravações de longa duração, utilizando algoritmos de seleção de unidades.

Um aspecto importante da técnica de seleção de unidades é que a qualidade da fala sintética tende a melhorar com o aumento do banco de unidades, já que assim há maior diversidade de contextos fonéticos e situações prosódicas. Isso faz com que sistemas de conversão texto-fala de muito boa qualidade possuam, em geral, bancos de grande duração, o que torna a ocupação de memória de tais sistemas elevada. Técnicas para reduzir a ocupação de memória de sistemas de síntese concatenativa sem impactar significativamente a qualidade são propostas no Capítulo 5.

4 Codificação de Fala

A representação usual de sinais de fala digitalizados é a modulação por código de pulso (*pulse code modulation – PCM*). Codificação de fala é o conjunto de técnicas utilizadas para representar um sinal de fala de formas alternativas à codificação PCM, que sejam mais adequadas para a transmissão ou armazenamento. A Fig. 4.1 mostra o diagrama de blocos de um sistema de codificação de fala. O sinal de fala digitalizado é processado pelo codificador de fonte, removendo informações redundantes e reduzindo sua taxa de bits. Em seguida, é processado pelo codificador de canal, para proporcionar robustez a erros introduzidos pelo canal e, opcionalmente, criptografia. O sinal codificado é então transmitido pelo canal, processado pelos decodificadores de canal e fonte e convertido novamente para o domínio contínuo.

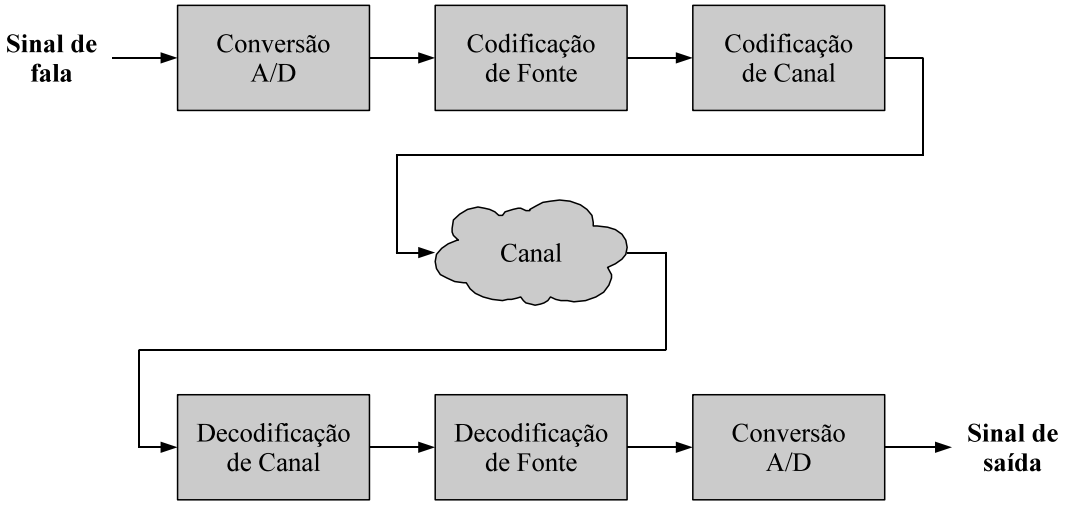

Fig. 4.1: Diagrama de blocos de um sistema de codificação de fala.

De maneira geral, o foco dos algoritmos de codificação de fala é redução da taxa de bits do sinal, ou seja, codificação de fonte. A maior parte dos codificadores/decodificadores de sinais de fala (*speech coders/decoders – codecs*) opera como mostrado no diagrama de blocos da Fig. 4.2. O sinal de fala digitalizado é primeiramente segmentado em quadros de análise de curta duração (até 30 ms), nos quais o sinal de fala pode ser considerado estacionário. Os quadros são então processados pelo codificador, extraindo um conjunto de

parâmetros que são posteriormente empacotados e transmitidos. O decodificador recebe esses dados, os desempacota e constrói uma estimativa do sinal original. Em geral, o sinal reconstruído não é igual ao sinal original, caracterizando um processo de codificação com perdas.

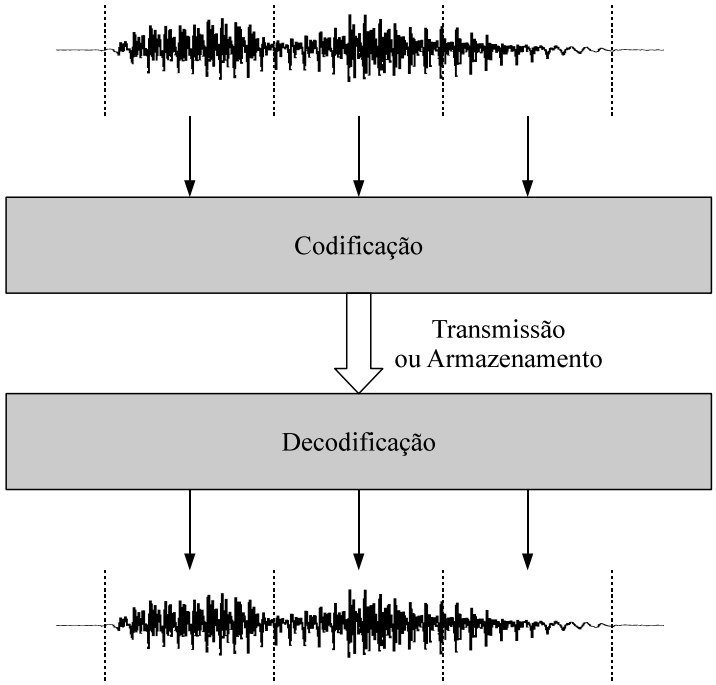

Fig. 4.2: Diagrama de blocos de um *codec*.

Os *codecs* de fala podem ser classificados, quanto à abordagem utilizada na codificação, como codificadores de forma de onda, paramétricos ou híbridos. Codificadores de forma de onda buscam preservar a forma de onda do sinal original, proporcionando qualidade satisfatória com taxas de bit relativamente altas (de 2 a 4 vezes menores que a taxa do sinal representado pela codificação PCM com quantização linear). Por sua vez, os codificadores paramétricos extraem parâmetros do sinal de fala para reconstruí-lo utilizando algum modelo, sem compromisso com a forma de onda do sinal original. Em geral, codificadores paramétricos possibilitam representar o sinal de fala com uma taxa de bits muito menor do que os codificadores de forma de onda (redução maior do que 20 vezes em relação à taxa da codificação PCM linear). Finalmente, codificadores híbridos unem características das abordagens paramétrica e de forma de onda: o sinal é reconstruído a partir de um conjunto de parâmetros obtidos de modo a aproximar a forma de onda do sinal original. As informações extraídas pelos codificadores híbridos são relativamente detalhadas, permitindo uma adequada aproximação do sinal original, ao contrário dos codificadores paramétricos, que descartam muitos detalhes. A abordagem híbrida, utilizada pela maior parte dos *codecs* atuais, proporciona qualidade muito boa com taxas de bits

relativamente baixas (de 8 a 15 vezes menor que a taxa da codificação PCM linear). A Tabela 4.1 apresenta uma comparação entre taxa de bits e qualidade para as três abordagens de codificação citadas. Nessa tabela, três *codecs* padronizados, representativos de cada uma das abordagens de codificação, são comparados com a codificação PCM linear a 16 bits, tomada como referência.

Tabela 4.1: Comparação entre taxa de bits e qualidade para as diferentes abordagens de codificação.

Codec	**Abordagem**	**Taxa de bits (kb/s)**	**Qualidade**
PCM linear 16 bits	Referência	128	Muito boa
FS1015 LPC [10]	Paramétrico	2,4	Pobre
G.726 ADPCM [17]	Forma de onda	16–40	Boa
G.729 CS-ACELP[18]	Híbrido	8	Boa

O campo de codificação de fala é vasto, incluindo conceitos de processamento de sinais, produção e percepção da fala, além de otimização de algoritmos [10], [19]. Os conceitos mais relevantes para o trabalho são discutidos a seguir.

4.1 Predição Linear

A predição linear é uma ferramenta de análise bem conhecida, com foco em aplicações de processamento de fala. O conceito de predição linear pode ser utilizado para estimação do envelope espectral, redução de redundâncias, determinação de parâmetros de modelos de produção de fala, dentre outras aplicações [20]. O princípio fundamental da predição linear é determinar uma estimativa $\hat{s}(n)$ de um sinal de fala $s(n)$ pela combinação linear de suas N amostras passadas

$$\hat{s}(n) = \sum_{i=1}^{N} a_i s(n-i) \tag{4.1}$$

onde N é a ordem do preditor linear, caracterizado pelos coeficientes a_i, denominados coeficientes de predição linear (*linear prediction coefficients* – LPCs). O erro de predição é definido como a diferença entre o sinal $s(n)$ e sua estimativa $\hat{s}(n)$

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{N} a_i \hat{s}(n-i). \tag{4.2}$$

Aplicando a transformada z em (4.2), obtém-se

$$E(z) = S(z)\left[1 - \sum_{i=1}^{N} a_i z^{-i}\right] = S(z)A(z) \tag{4.3}$$

onde $S(z)$ é a transformada z do sinal $s(n)$ e $E(z)$, a transformada z do erro de predição $e(n)$. Dessa forma, o erro de predição (também chamado resíduo) pode ser obtido filtrando o sinal $s(n)$ por um sistema com função de transferência $A(z) = 1 - a_1 z^{-1} - \cdots - a_N z^{-N}$, denominado filtro de análise. O sinal original $s(n)$ pode então ser recuperado filtrando o erro de predição pelo filtro de síntese $1/A(z)$. Assim,

$$S(z) = \frac{E(z)}{A(z)}. \tag{4.4}$$

É possível determinar um conjunto de coeficientes $\mathbf{a} = \{a_1\ a_2\ \cdots\ a_N\}$ que minimize o erro de predição. Diferentes critérios de minimização dão origem aos métodos da correlação, da covariância e *lattice* para determinação dos coeficientes [20]. A abordagem mais utilizada em codificação de fala é o método da correlação, obtido minimizando a média quadrática do erro de predição. De acordo com esse método, os coeficientes são determinados por

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r}, \tag{4.5}$$

onde $\mathbf{R}$ é a matriz de autocorrelação de ordem N de $s(n)$ e $\mathbf{r}$ é o vetor de autocorrelação de ordem N de $s(n)$. Essa equação pode ser resolvida usando diferentes procedimentos. Um método eficiente, bastante utilizado, é o algoritmo de Levinson-Durbin [20]. As raízes do filtro caracterizado pelos coeficientes determinados através de (4.5) se localizam no círculo unitário [10]. Dessa maneira, o filtro de análise $1/A(z)$ obtido pelo método da correlação é garantidamente estável.

O desempenho do preditor linear é medido através do ganho de predição, dado por

$$G_{\mathrm{p}} = 10\log\left(\frac{\sigma_s^2}{\sigma_e^2}\right), \tag{4.6}$$

onde σ_s^2 e é a variância do sinal de entrada e σ_e^2 é a variância do erro de predição. Erros de predição com pequena variância caracterizam um adequado desempenho do preditor linear, produzindo um ganho de predição alto.

Caso a ordem N do preditor linear seja suficientemente elevada, o erro de predição $e(n)$ se assemelha a um ruído branco. Para sons vozeados, essa ordem é elevada, próxima ao período de *pitch* (até 150, para sinais amostrados a 8 kHz). Embora seja possível utilizar preditores de elevada ordem para aumentar o ganho de predição, o custo computacional dessa abordagem é muito alto. Uma solução alternativa, que apresenta bons resultados, é

utilizar um preditor de longo termo

$$H(z) = 1 + bz^{-T} \tag{4.7}$$

em cascata com um preditor de ordem reduzida (em torno de 10), denominado preditor de curto termo. Os parâmetros b e T do preditor de longo termo são determinados de maneira análoga aos coeficientes de $A(z)$, minimizando a média quadrática do erro de predição [10]. O preditor de longo termo, também denominado preditor de *pitch*, remove o efeito da periodicidade em sinais vozeados, proporcionando aumento no ganho de predição.

Os conceitos previamente discutidos são válidos para sinais estacionários. No entanto, o sinal de fala só pode ser considerado estacionário quando analisado em trechos de curta duração. Sendo assim, para aplicar a predição linear em sinais de fala, é necessário uma segmentação em quadros de curta duração, como mencionado anteriormente. É importante notar que os parâmetros do preditor de longo termo devem ser atualizados mais freqüentemente do que os do preditor de curto termo, para compensar pequenas variações no período de *pitch* que ocorrem em um quadro do sinal de fala [10]. Dessa maneira, muitos *codecs* utilizam a estrutura de quadros e subquadros, ilustrada na Fig. 4.3. O sinal de fala é dividido em quadros de análise, nos quais são determinados os parâmetros do preditor de curto termo. Em cada quadro, o erro de predição de curto termo é então calculado e segmentado em subquadros de menor duração, nos quais os parâmetros do preditor de longo termo são estimados. O erro de predição final é então obtido filtrando o erro de predição de curto termo em cada subquadro pelo preditor de longo termo correspondente. Muitos *codecs* atuais realizam a codificação de longo termo usando o conceito de *codebook* adaptativo [10]. Usualmente os subquadros têm duração de 5 ms e os quadros são formados por 2 a 6 subquadros (duração de 10 a 30 ms).

O erro de predição total e os coeficientes de predição linear de curto e longo termo são parâmetros suficientes para reconstruir um dado quadro do sinal de fala. Considerando que os parâmetros de um dado quadro podem ser armazenados em b_q bits, a taxa de bits do sinal codificado é dada por

$$r_\mathrm{c} = \frac{b_\mathrm{q}}{T_\mathrm{q}} \tag{4.8}$$

onde T_q é a duração do quadro em segundos. O processo de predição linear remove redundâncias do sinal de fala, de modo que os parâmetros podem ser representados de maneira compacta, tornando r_c menor do que a taxa de bits do sinal representado pela codificação PCM, dada por

$$r_\mathrm{o} = f_\mathrm{s} b \tag{4.9}$$

onde f_s é a taxa de amostragem do sinal, e b é o número de bits utilizados para representar

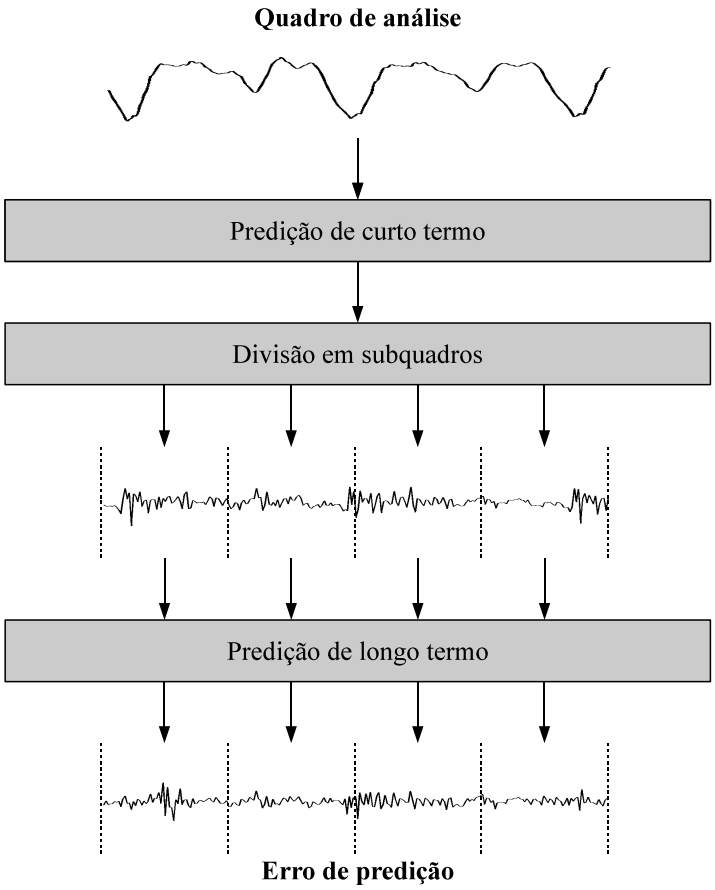

Fig. 4.3: Estrutura de quadros e subquadros.

cada amostra. Como exemplo de codificação de fala utilizando o conceito de predição linear, a Seção 4.2 descreve o *codec* iLBC.

A codificação de fala por predição linear também pode ser interpretada como uma implementação do modelo fonte/filtro de produção de fala discutido na Seção 2.1. A excitação é criada filtrando o resíduo com filtro de síntese de longo termo, e a resposta do trato vocal é caracterizada pelo filtro de síntese de curto termo, como ilustrado na Fig. 4.4. Uma diferença marcante entre o esquema mostrado aqui e aquele discutido anteriormente é a presença de apenas uma fonte de excitação, já que o filtro de síntese de longo termo é capaz de produzir excitações periódicas a partir de um sinal com característica ruidosa. Sendo assim, a classificação dos quadros do sinal como vozeados ou não-vozeados se torna desnecessária, simplificando o sistema de codificação e melhorando a qualidade da fala codificada.

4.1.1 Representação LSF dos Coeficientes da Predição Linear

Como mencionado, os coeficientes de predição linear devem ser representados de maneira compacta para reduzir a taxa de bits do sinal codificado. Uma possível solução é

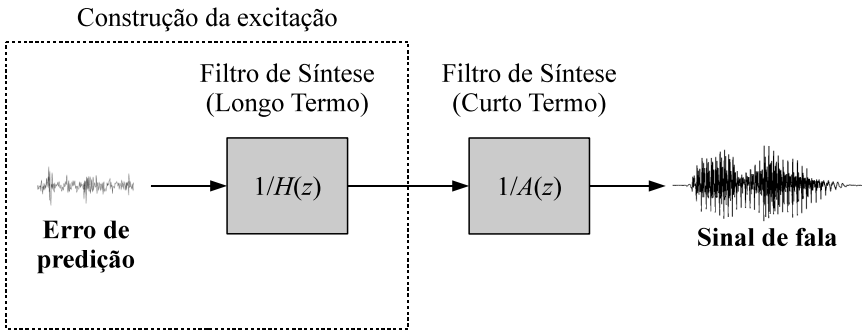

Fig. 4.4: Modelo fonte-filtro de produção de fala utilizando predição linear.

quantizar os coeficientes utilizando reduzido número de bits. No entanto, os coeficientes de $A(z)$ não são apropriados para quantização, pelos seguintes motivos:

1. A faixa dinâmica dos coeficientes a_i não é limitada, dificultando o projeto de um quantizador adequado.

2. Mudanças nos coeficientes causadas pela quantização alteram a posição das raízes de $A(z)$, que podem ser deslocadas para fora do círculo unitário, tornando o filtro de síntese $1/A(z)$ instável.

Dessa maneira, é importante utilizar alguma representação alternativa para os coeficientes, mais robusta à quantização. Dentre as possíveis representações, se destacam os coeficientes de reflexão, os parâmetros LAR (*log area ratio*) e os parâmetros LSF (*line spectral frequency*) [21]. A representação LSF, discutida nessa seção, é utilizada em diversos codificadores atuais.

Os coeficientes LSF são obtidos a partir das raízes dos polinômios $P(z)$ e $Q(z)$, derivados do preditor linear $A(z)$ através das relações

$$P(z) = A(z) + z^{-M-1}A(z) \tag{4.10a}$$

$$Q(z) = A(z) - z^{-M-1}A(z) \tag{4.10b}$$

onde M é a ordem do preditor linear.

Uma propriedade importante das relações (4.10) é que, caso os zeros de $A(z)$ se localizem no círculo unitário, as raízes de $P(z)$ e $Q(z)$ são intercaladas sobre a circunferência de raio unitário, ou seja, têm magnitude unitária e seus ângulos obedecem à seguinte ordem:

$$\theta_1^{(P)} < \theta_1^{(Q)} < \theta_2^{(P)} < \theta_2^{(Q)} < \dots$$

onde $\theta_i^{(P)} = \arg z_i^{(P)}$ denota o ângulo da i-ésima raiz de $P(z)$ e $\theta_i^{(Q)}$, o ângulo da i-ésima raiz de $Q(z)$. Essa propriedade é ilustrada na Fig. 4.5. Sua recíproca também é verdadeira,

ou seja, caso as raízes de $P(z)$ e $Q(z)$ sejam intercaladas sobre a circunferência de raio unitário, as raízes de $A(z) = \frac{1}{2}[P(z) + Q(z)]$ se localizam no círculo unitário [22].

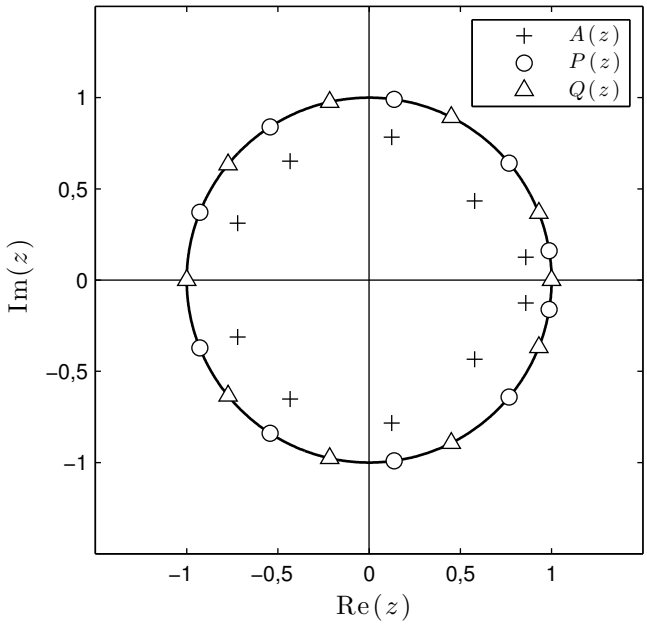

Fig. 4.5: Propriedade de intercalamento dos coeficientes LSF.

Como as raízes de $P(z)$ e $Q(z)$ estão sobre a circunferência de raio unitário, é suficiente representá-las por seus ângulos. Ainda, como as raízes se apresentam em pares complexos conjugados, basta armazenar aquelas com ângulos positivos. Desse modo, os coeficientes LSF são definidos através do ângulo das N raízes de $P(z)$ e $Q(z)$ localizadas no semiplano superior do plano z. A representação LSF dos coeficientes de predição linear tem alguns aspectos vantajosos para aplicações em codificação de fala, a saber:

1. O número de parâmetros LSF é igual à ordem de $A(z)$, de modo que essa representação não introduz redundâncias.

2. Ao contrário dos coeficientes de $A(z)$, os parâmetros LSF têm faixa dinâmica limitada, o que os torna mais adequados para a quantização.

3. Desde que a propriedade de intercalamento seja mantida, erros de quantização não tornam o filtro de síntese $1/A(z)$ instável.

4. É possível interpolar os parâmetros LSF. Essa característica é aproveitada em *codecs* que utilizam a estratégia de divisão de quadros em subquadros. Assim, é possível determinar um preditor para cada subquadro interpolando os coeficientes calculados utilizando o quadro completo.

O preditor linear pode ser representado de maneira muito compacta quantizando vetorialmente o conjunto de coeficientes LSF. O conceito fundamental da quantização vetorial é particionar um espaço N-dimensional em L regiões, representando cada região por um vetor (*codeword*) [23]. O conjunto de *codewords* é denominado *codebook*. Um dado vetor de coeficientes $\mathbf{v}$ é quantizado vetorialmente através de

$$\hat{\mathbf{a}} = \arg\min_i \mathrm{d}(\mathbf{a},\mathbf{c}_i) \quad \forall\, \mathbf{c}_i \in \mathcal{C} \tag{4.11}$$

onde $\mathbf{c}_i$ são as *codewords* de um *codebook* $\mathcal{C}$ e d é uma função distância (por exemplo, euclidiana ou de Mahalanobis). Assim, os vetores de coeficientes são representados pelos índices das *codewords* mais próximas.

A quantização provoca distorções no envelope espectral representado pelos coeficientes, definida como [19]

$$d_\mathrm{e} = \frac{1}{\pi}\int_0^\pi \left[20\log\left|A'(\mathrm{e}^{\mathrm{j}\omega})\right| - 20\log\left|A(\mathrm{e}^{\mathrm{j}\omega})\right|\right]^2 \mathrm{d}\omega \tag{4.12}$$

onde $A(\mathrm{e}^{\mathrm{j}\omega})$ é a resposta em freqüência do filtro de análise derivado dos coeficientes não quantizados e $A'(\mathrm{e}^{\mathrm{j}\omega})$, a resposta em freqüência do filtro derivado dos coeficientes quantizados. É possível avaliar (4.12) numericamente utilizando a transformada discreta de Fourier (DFT). Considerando uma DFT de N pontos, obtém-se

$$D_\mathrm{e} = \frac{1}{N/2}\sum_{k=0}^{N/2-1} \left[20\log|A'(k)| - 20\log|A(k)|\right]^2 \tag{4.13}$$

onde $A'(k)$ e $A(k)$ são as amostras das respostas em freqüência dos filtros de análise, calculadas pela DFT. Usualmente, a distorção espectral é calculada utilizando uma porção limitada do espectro, tipicamente de 125 a 3100 Hz, pois as componentes fora dessa faixa não são perceptualmente significativas mas podem prejudicar a avaliação da distorção. É possível avaliar o desempenho de um dado quantizador calculando a distorção espectral de um conjunto de vetores de parâmetros em um banco de teste. Para que a quantização seja transparente, ou seja, não cause distorção audível na fala codificada, os seguintes critérios devem ser observados [21]:

1. A distorção espectral média do banco de teste deve ser menor que 1 dB.
2. Não mais que 2 % dos vetores do banco devem ter distorção maior que 2 dB.
3. Nenhum vetor do banco deve ter distorção maior que 4 dB.

Utilizando quantização vetorial, vetores de parâmetros LSF usualmente empregados em codificação de fala (ordem 10 a 12) podem ser codificados de maneira transparente utili-

zando apenas 20 bits (*codebooks* com 2^{20} elementos) [19].

4.2 O *internet low bitrate codec* – iLBC

A maioria dos codificadores atuais explora correlações de longo termo no erro de predição, removendo redundâncias que ultrapassam os limites dos quadros. Dessa maneira, o sinal de fala pode ser representado com muito boa qualidade utilizando reduzidas taxas de bit. No entanto, a perda de um quadro causa distorções que se propagam por longos períodos [4]. Essa característica impede o acesso aleatório a um conjunto de dados codificados, que é fundamental para aplicações em síntese de fala concatenativa. Para tais aplicações, é necessário o uso de um *codec* que seja robusto a perda de quadros codificados. Um *codec* padronizado que atende a esse requisito é o iLBC, concebido para ter boa qualidade em aplicações de comunicações de voz através da Internet (*voice over IP* – VoIP). Para obter robustez à perda de quadros, o iLBC não explora as correlações de longo termo entre quadros distintos. Essa característica, diferente da maior parte dos *codecs* atuais, permite decodificar, sem distorções, qualquer trecho de uma gravação codificada com o iLBC. Sendo assim, esse *codec* é vantajoso para aplicações em síntese concatenativa de fala, proporcionando acesso aleatório ao banco de gravações. Esta seção apresenta algumas características do iLBC, descrito detalhadamente em [5].

O *codec* trabalha com sinais de fala de banda estreita (amostrados a 8 kHz), quantizados em 16 bits. De acordo com o modo de operação, o iLBC pode codificar sinais de fala em uma das seguintes taxas:

- 15,2 kbps, utilizando quadros de 20 ms.
- 13,33 kbps, utilizando quadros de 30 ms.

Em cada quadro, 10 coeficientes LPC são calculados, convertidos para a representação LSF e quantizados vetorialmente em 20 bits. Os coeficientes LSF calculados em um dado quadro são interpolados, de modo a obter um preditor para cada subquadro de 5 ms. O erro de predição do quadro é então obtido filtrando o sinal de cada subquadro pelo preditor correspondente.

A codificação do erro de predição é a etapa que distingue o iLBC dos demais *codecs* utilizados atualmente. Para não considerar informações de quadros passados, o algoritmo de codificação analisa o erro de predição do quadro corrente, determinando o trecho de maior energia (com duração igual a 57 amostras quando se utilizam quadros de 20 ms e 58 amostras para quadros de 30 ms). Esse trecho é então quantizado, servindo como

estado inicial de um *codebook* adaptativo para codificar o restante do quadro. Antes da quantização, o estado inicial é filtrado por um sistema *all-pass* derivado do preditor linear, de modo a obter uma distribuição mais uniforme das amplitudes de suas amostras. Então, a amostra de maior magnitude é encontrada e seu valor é quantizado em 6 bits. As amostras do estado inicial são então normalizadas pelo valor da amostra de maior magnitude, e quantizadas em 3 bits usando um esquema de quantização diferencial.

Dessa forma, um dado quadro contém todas as informações necessárias à sua reconstrução. Devido a essa característica, a taxa de bits do iLBC é maior do que a de outros *codecs*, como, por exemplo, o G.729 [18]. Essa característica de independência entre quadros é vantajosa para a aplicação discutida nesse trabalho, possibilitando a compressão do banco de gravações de sistemas de síntese concatenativa sem prejuízo à capacidade de acesso aleatório.

5 Redução da ocupação de memória do TTS

Este capítulo apresenta a arquitetura do sistema TTS considerado, bem como as alterações propostas a essa arquitetura, visando reduzir a ocupação de memória do sistema.

5.1 Arquitetura do TTS do LINSE

A arquitetura do sistema TTS considerado é ilustrada pelo diagrama de blocos da Fig. 5.1. Primeiramente, funções de processamento lingüístico são aplicadas ao texto de entrada, obtendo uma seqüência alvo. A seqüência alvo é processada pelo módulo de seleção de unidades, determinando as unidades mais apropriadas para a síntese, de acordo com informações contextuais e acústicas contidas no banco de parâmetros. Finalmente, as unidades selecionadas são extraídas do banco de gravações e concatenadas, construindo a fala sintética.

Conforme mencionado no Capítulo 3, o algoritmo de seleção de unidades determina a seqüência de unidades $\hat{U}$ mais apropriada para a síntese minimizando uma função custo C, composta por duas parcelas: o custo alvo C_t e o custo de concatenação C_c. Dessa forma,

$$\hat{U} = \arg\min_{U} \left[\sum_{i=1}^{n} C_\mathrm{t}(t_i,u_i) + \sum_{i=2}^{n} C_\mathrm{c}(u_{i-1},u_i) \right] \tag{5.1}$$

onde t_i são as n unidades da seqüência alvo T e u_i são as unidades de uma dada seqüência candidata U.

O custo alvo é calculado como a soma de n_t custos parciais, referentes aos diferentes parâmetros contextuais utilizados. Assim,

$$C_\mathrm{t}(t_i,u_i) = \sum_{j=1}^{n_\mathrm{t}} c_\mathrm{t}^{(j)}(t_i^{(j)},u_i^{(j)}) \tag{5.2}$$

onde $c_\mathrm{t}^{(j)}$ é a função custo parcial referente ao j-ésimo parâmetro contextual, $t_i^{(j)}$ é o j-ésimo parâmetro contextual da unidade alvo t_i e $u_i^{(j)}$ é o j-ésimo parâmetro contextual da unidade candidata u_i. O custo parcial é definido como zero caso o parâmetro da

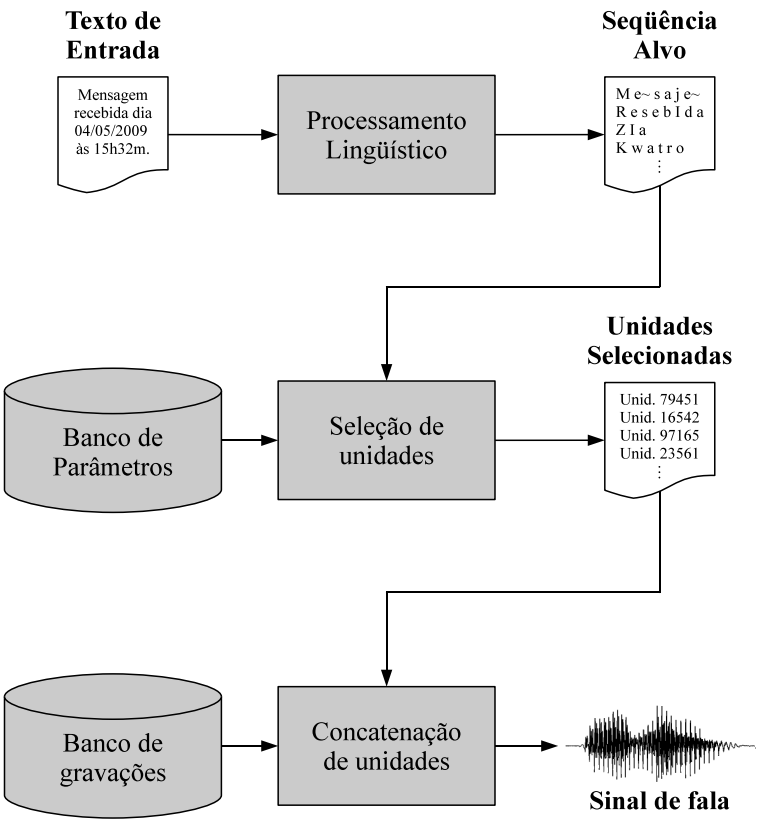

Fig. 5.1: Diagrama de blocos do sistema de conversão texto-fala.

unidade candidata seja igual ao parâmetro da unidade alvo e uma penalidade $p^{(j)}$, caso os parâmetros sejam diferentes. Assim,

$$c_{\mathrm{t}}^{(j)}(t_i^{(j)},u_i^{(j)}) = \begin{cases} 0 & \text{se } t_i^{(j)} = u_i^{(j)} \\ p^{(j)} & \text{se } t_i^{(j)} \neq u_i^{(j)}. \end{cases} \tag{5.3}$$

O sistema considerado utiliza 18 parâmetros contextuais, dentre os quais se destacam:

- O contexto fonético das unidades.
- A posição das unidades na palavra.
- A posição das unidades na frase.
- O tipo de sílaba em que as unidades estão inseridas.
- A classe gramatical da palavra que contém uma dada unidade.

O custo de concatenação é definido como a soma ponderada de n_{c} distâncias acústicas. Assim,

$$C_{\mathrm{c}}(u_{i-1},u_i) = \sum_{k=1}^{n_{\mathrm{c}}} q^{(k)} d_{\mathrm{c}}^{(k)}(u_{i-1}^{(k)},u_i^{(k)}) \tag{5.4}$$

onde $q^{(k)}$ é o fator de ponderação do k-ésimo parâmetro acústico e $d_{\mathrm{c}}^{(k)}(u_{i-1}^{(k)},u_i^{(k)})$ é a função distância utilizada para comparar o k-ésimo parâmetro acústico de duas unidades candidatas. Os parâmetros acústicos utilizados no sistema considerado são:

- *Pitch* médio da unidade.
- Energia.
- Envelope espectral.

As funções distância de *pitch* e energia são definidas, respectivamente, como o módulo da diferença do *pitch* e da energia nas unidades consideradas. Já no caso do envelope espectral, a função distância é definida como a distância euclidiana entre os vetores de coeficientes que caracterizam o envelope espectral das unidades. No sistema considerado, o envelope espectral é caracterizado por 12 coeficientes mel-cepstrais (*mel-frequency cepstral coefficients* – MFCCs).

A estrutura do banco de parâmetros é ilustrada na Fig. 5.2. Há um conjunto de parâmetros contextuais e acústicos armazenados para cada unidade do banco. Cada unidade é dividida em quadros, correspondendo aos períodos de *pitch* do sinal de fala em trechos vozeados, enquanto nos demais trechos, têm duração fixa de 10 ms. Embora a figura mostre um conjunto de parâmetros acústicos em cada unidade, há um conjunto distinto de parâmetros acústicos em cada quadro. Utilizando os parâmetros acústicos de cada quadro, é possível calcular o custo de concatenação entre segmentos menores do que as unidades. Sendo assim, há maior liberdade para a escolha do ponto de concatenação das unidades, melhorando a qualidade da fala sintética.

O banco de parâmetros ocupa grande espaço de memória (em torno de 700 MB). Os dados contidos nesse banco são acessados com muita freqüência, por serem utilizados no cálculo dos custos do processo de seleção de unidades. Desse modo, para se obter um melhor desempenho desse algoritmo, é necessário manter o banco de parâmetros na memória principal do sistema, que proporciona acesso muito mais rápido do que o armazenamento em disco. Sendo assim, a compactação do banco de parâmetros traz grandes benefícios ao TTS, por reduzir a ocupação de memória principal do sistema. As técnicas propostas para compactar o banco de parâmetros são apresentadas na Seção 5.2.

O banco de gravações contém sentenças gravadas por um locutor, cujos trechos são concatenados para construir a fala sintética. As gravações são amostradas a 8 kHz e quantizadas não-linearmente em 8 bits, utilizando a curva da lei A. O sistema considerado utiliza 28 horas de gravações, ocupando 800 MB de memória. Os dados do banco de gravações são acessados em menor freqüência do que os dados do banco de parâmetros.

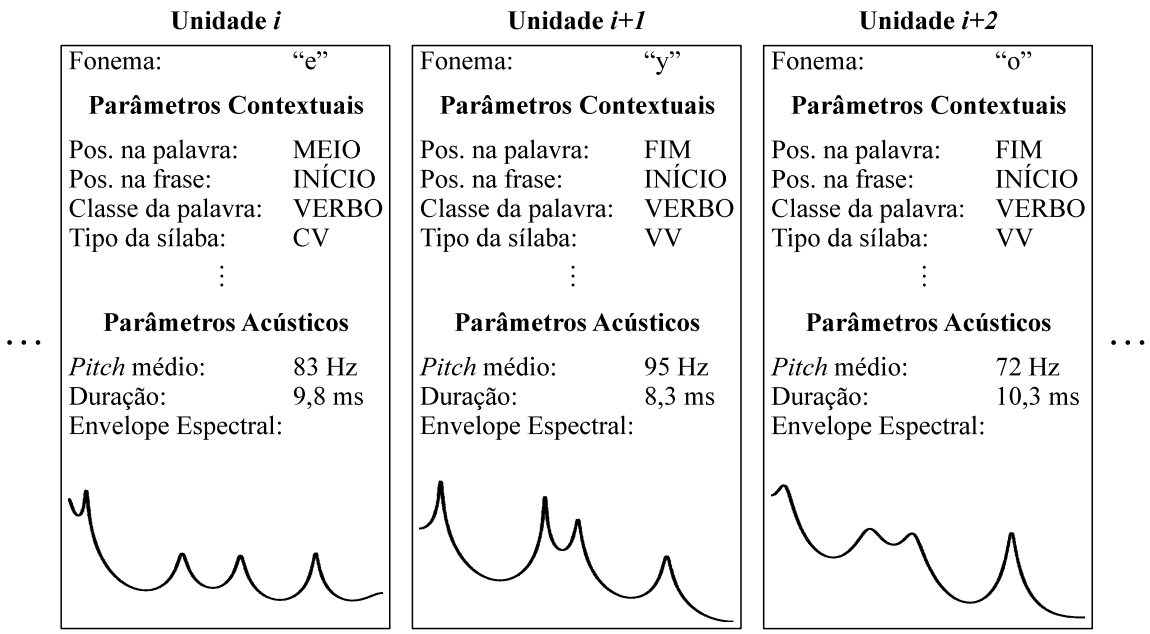

Fig. 5.2: Banco de parâmetros.

Desse modo, não há necessidade de manter todas as gravações na memória principal do sistema. Além do mais, é desejável reduzir a ocupação de memória desse banco, para possibilitar o uso do TTS em equipamentos com restrição de espaço em memória secundária (por exemplo, disco rígido ou memória *flash*) ou utilizar diferentes vozes em um único equipamento. A compressão do banco de gravações é discutida na Seção 5.3.

5.2 Compactação do banco de parâmetros

Grande parte da ocupação de memória no banco de parâmetros se deve aos 12 coeficientes MFCC utilizados na caracterização do envelope espectral das unidades, armazenados como vetores representados em ponto flutuante com precisão simples (32 bits). Propõe-se substituir os coeficientes mel-cepstrais por 10 coeficientes LSF, que além de serem uma representação adequada do envelope espectral, podem ser quantizados vetorialmente. Outro aspecto vantajoso do uso de coeficientes LSF é a possibilidade de reaproveitá-los na codificação do banco de gravações, como discutido em detalhes na Seção 5.3. Além do mais, a quantização vetorial permite pré-calcular as distâncias acústicas de todos os vetores de parâmetros, reduzindo o esforço computacional no cálculo do custo de concatenação.

Os coeficientes LSF são quantizados vetorialmente, utilizando o esquema de quantização subvetorial [19]: considera-se que o vetor de 10 elementos é composto por 3 subvetores com dimensões 3, 3 e 4. Cada subvetor é então quantizado independentemente, utilizando um quantizador vetorial próprio.

As *codewords* do quantizador vetorial utilizado na aplicação considerada devem ser de dimensão múltipla de 8 bits, por questões de alinhamento de memória. Sendo assim, três quantizadores vetoriais, com *codewords* de 16, 24 e 32 bits, foram treinados pelo algoritmo LBG [24] utilizando um banco contendo 2,3 milhões de vetores de parâmetros, correspondendo a 5 horas de gravações. O desempenho desses quantizadores em relação à distorção espectral D_{e} foi então avaliado, como discutido na Seção 4.1.1. Os resultados da avaliação, realizada em um banco de testes contendo 465 mil vetores de parâmetros, são mostrados na Tabela 5.1. Os critérios para quantização transparente são mostrados entre parênteses. O quantizador de 16 bits viola o critério de distorção média por uma larga margem. Por sua vez, o quantizador de 24 bits viola os critérios de distorção média e distorção menor do que 4 dB por pequena margem. Apenas o quantizador de 32 bits atende aos três critérios apresentados. Optou-se por utilizar o quantizador de 24 bits, que, apesar de violar alguns dos critérios por uma margem bastante pequena, proporciona maior compactação do banco de parâmetros do que o quantizador de 32 bits. A alocação de bits utilizada por esse quantizador é de 8 bits por subvetor.

Tabela 5.1: Desempenho dos quantizadores vetoriais

Número de bits	16	24	32
D_{e} médio (≤ 1 dB)	2,83 dB	1,09 dB	0,45 dB
$D_{\mathrm{e}} > 2$ dB ($\leq 2\,\%$)	0,59 %	0,10 %	0,01 %
$D_{\mathrm{e}} > 4$ dB (0 %)	0,19 %	0,01 %	0,00 %

5.3 Compactação do banco de gravações

A redução de memória do banco de gravações é obtida com a compressão das gravações por algoritmos de codificação de fala. Conforme mencionado no Capítulo 4, grande parte dos codificadores de fala considera informações de quadros passados na codificação, não sendo apropriados para a aplicação em questão, que exige acesso aleatório aos dados do banco. Sendo assim, propõe-se codificar o banco de gravações utilizando o *codec* iLBC, que proporciona qualidade adequada, sem prejuízo à capacidade de acesso aleatório.

Além de serem utilizados para o cálculo do custo de concatenação no algoritmo de seleção de unidades, os vetores de coeficientes LSF armazenados no banco de parâmetros podem ser reaproveitados na codificação do banco de gravações. A dimensão dos vetores de coeficientes foi escolhida para ser igual à ordem dos filtros de análise e síntese utilizados pelo *codec* iLBC (ordem 10). Os coeficientes armazenados no banco de parâmetros são

calculados de acordo com os períodos de *pitch* do sinal de fala. Para que possam ser aproveitados na codificação, esses coeficientes devem se apresentar na taxa de um a cada subquadro de 5 ms. Sendo assim, a conversão entre bases de tempo se faz necessária. Para realizar essa conversão, os coeficientes do banco de parâmetros são interpolados linearmente. Então, a média dos valores interpolados é calculada para cada trecho de 5 ms, como ilustrado na Fig. 5.3. Esse algoritmo é utilizado tanto na codificação, realizada na construção do banco, quanto na decodificação, realizada no processo de síntese. Essa técnica proporciona uma redução da taxa de bits do *codec* iLBC de 15,2 para 14,4 kbps (ou 12 kbps para a taxa de 13,3 kbps).

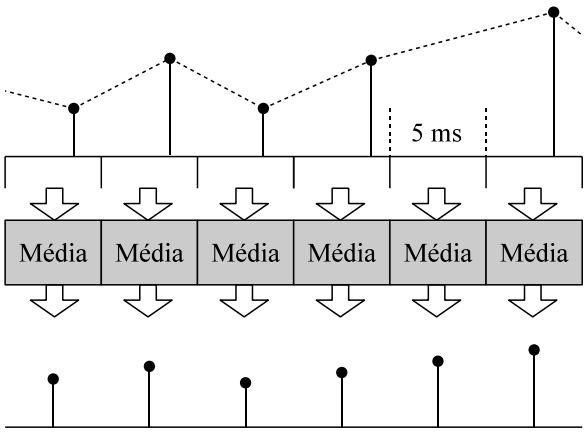

Fig. 5.3: Interpolação dos parâmetros LSF.

Para realizar a síntese, os trechos correspondentes às unidades selecionadas são extraídos do banco e concatenados. Uma possível abordagem para a síntese seria concatenar os parâmetros dos trechos selecionados, decodificando a seqüência de dados codificados. No entanto, observou-se que essa abordagem produz resultados ruins, com grandes distorções na fala sintética. Desse modo, se utiliza uma abordagem diferente, decodificando os trechos do banco antes de concatená-los. Porém, a simples concatenação dos elementos decodificados apresenta qualidade pobre, com a ocorrência de artefatos, tais como cliques, nos pontos de concatenação. Para produzir fala sintética com qualidade adequada, é necessário utilizar uma técnica de concatenação suave, associada à determinação automática dos pontos mais apropriados para concatenação.

O algoritmo utilizado para determinar o melhor ponto de concatenação é baseado no cálculo da correlação em torno dos limites dos segmentos que serão concatenados. Para tal, são definidos dois vetores auxiliares, de dimensão N_c: $\mathbf{x}$, centrado na última amostra do segmento à esquerda e $\mathbf{y}_i$, centrado na i-ésima amostra do segmento à direita. O valor de N_c utilizado é igual à dimensão dos quadros do *codec* iLBC (160 amostras para a taxa de 15,2 kbps e 240 amostras para a taxa de 13,3 kbps). A Fig. 5.4 ilustra a posição

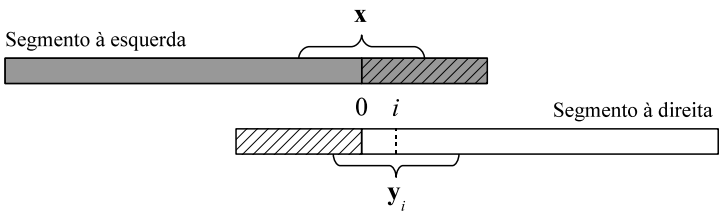

Fig. 5.4: Vetores auxiliares $\mathbf{x}$ e $\mathbf{y}_i$.

desses vetores em relação aos limites dos trechos que serão concatenados. As regiões hachuradas indicam trechos não pertencentes aos segmentos que serão concatenados, mas decodificados para serem utilizados como contexto no cálculo da correlação. Para obter os trechos de contexto, é necessário decodificar um quadro à direita e um quadro à esquerda de cada segmento extraído do banco. Utilizando os vetores auxiliares, a correlação para um dado atraso i é definida como

$$r_{\mathbf{xy}}(i) = \mathbf{x}^{\mathrm{T}}\mathbf{y}_i. \tag{5.5}$$

O ponto de concatenação i_{c} é então determinado verificando o ponto que maximiza o módulo da correlação. Assim:

$$i_{\mathrm{c}} = \arg\max_i \left| r_{\mathbf{xy}}(i) \right|, \tag{5.6}$$

com i variando de $-N_{\mathrm{c}}/2 + 1$ a $-N_{\mathrm{c}}/2$.

Após determinar o ponto de concatenação, os segmentos são concatenados de maneira suave, como ilustrado na Fig. 5.5. O segmento à esquerda é multiplicado por uma janela triangular decrescente de N_{c} amostras, centrada no último ponto desse segmento. Por sua vez, o segmento à direita é multiplicado por uma janela triangular crescente de N_{c} amostras, centrada no ponto de concatenação i_{c}. Esses dois segmentos são então sobrepostos e somados, suavizando a transição entre segmentos.

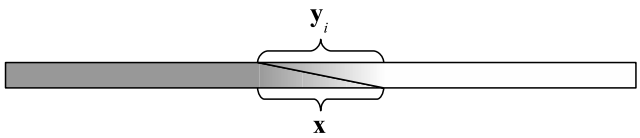

Fig. 5.5: Concatenação suave dos segmentos.

O *codec* iLBC pode operar nas taxas de 15,2 kbps ou 13,3 kbps. Testes preliminares, durante o desenvolvimento, indicaram que a taxa de 13,3 kbps apresenta qualidade notadamente inferior à taxa de 15,2 kbps. Desse modo, o banco de gravações é codificado utilizando essa taxa, incluindo o reaproveitamento dos vetores LSF do banco de parâmetros.

5.4 Sumário e Comentários

O sistema de conversão texto-fala considerado produz fala sintética de muito boa qualidade, mas apresenta grande ocupação de memória. Os dados utilizados pelo algoritmo de seleção de unidades são estruturados no banco de parâmetros e as gravações utilizadas na produção da fala sintética são armazenadas no banco de gravações.

Grande parte da ocupação de memória no banco de parâmetros se deve aos coeficientes MFCC utilizados na caracterização do envelope espectral das unidades. Os coeficientes MFCC são substituídos por coeficientes LSF, quantizados vetorialmente. O banco de gravações é codificado utilizando o *codec* iLBC, permitindo acesso aleatório a qualquer trecho do banco. Os coeficientes LSF armazenados no banco de parâmetros são reaproveitados na codificação do banco de gravações.

O resultado e o desempenho das técnicas apresentadas é discutido no Capítulo 6.

6 Experimentos e Resultados

Este capítulo apresenta uma avaliação das técnicas de redução de ocupação de memória discutidas anteriormente. As técnicas são aplicadas a um sistema de conversão texto-fala desenvolvido no LINSE, sendo então avaliadas considerando os seguintes aspectos:

1. Capacidade de redução de memória
2. Efeito na qualidade da fala sintética.

O sistema TTS considerado possui um banco de gravações com duração total de 28 horas. As gravações realizadas em um estúdio por um locutor profissional (do sexo masculino) foram adquiridas a uma taxa de amostragem de 44,1 kHz e quantizadas linearmente em 16 bits. As gravações são transcritas foneticamente, identificando as unidades do banco. As unidades são então caracterizadas por parâmetros contextuais e acústicos determinados através de técnicas automáticas, com posterior conferência por especialista. Na construção do banco, as gravações são reamostradas para 8 kHz e convertidas para quantização não-linear em 8 bits, segundo a curva da lei A. As gravações e os parâmetros são então armazenados em arquivos, utilizando um formato próprio do TTS.

As técnicas de redução de memória discutidas anteriormente são aplicadas de forma independente aos bancos de gravações e de parâmetros. Sendo assim, é possível obter quatro versões distintas do sistema de síntese através das combinações dos bancos originais e compactados. As possíveis versões do sistema de síntese são denominadas:

1. **MFCC + PCM.** Bancos de parâmetros e gravações originais.
2. **LSF + PCM.** Banco de parâmetros compactado e banco de gravações original.
3. **MFCC + iLBC.** Banco de parâmetros original e banco de gravações compactado.
4. **LSF + iLBC.** Bancos de parâmetros e gravações compactados.

Na seqüência, os resultados obtidos nas avaliações de redução de memória e qualidade são apresentados e discutidos.

6.1 Redução de Ocupação de Memória

O objetivo principal deste trabalho é a redução da ocupação de memória em sistemas de síntese de fala. Conforme mencionado, o banco do sistema considerado é dividido em dois segmentos, o banco de parâmetros e o banco de gravações. Nessa seção, o resultado das técnicas propostas, considerando a capacidade de redução de memória, é discutido.

A ocupação de memória do banco de parâmetros se deve, em grande parte, aos coeficientes MFCC utilizados para caracterizar o envelope espectral dos quadros do banco. Originalmente, esses coeficientes são armazenados como um vetor de 12 números em ponto flutuante com precisão simples, resultando na ocupação de 384 bits por quadro. Utilizando parâmetros LSF quantizados vetorialmente, o envelope espectral dos quadros do banco pode ser adequadamente representado utilizando apenas 24 bits (redução de 93 %). O banco de parâmetros contém ainda outras informações, tais como os parâmetros contextuais das unidades que não são comprimidas pela técnica proposta. Ainda assim, o uso dos parâmetros LSF quantizados vetorialmente permite reduzir a ocupação de memória do banco de parâmetros de 735 MB para 147 MB (redução de 80 %).

A redução do uso de memória no banco de gravações é obtida codificando as gravações com o *codec* iLBC. Avaliações preliminares indicaram que esse *codec*, utilizando taxa de 15,2 kbps, é adequado para a aplicação considerada, proporcionando capacidade de acesso aleatório às unidades do banco, sem causar grandes distorções. Além disso, os parâmetros LSF armazenados no banco de parâmetros podem ser reaproveitados com sucesso na codificação do banco de gravações, proporcionando redução na taxa de bits básica do iLBC. Codificando as gravações com *codec* iLBC modificado para reaproveitar os coeficientes LSF do banco de parâmetros, é possível reduzir o uso de memória do banco de gravações de 800 MB para 180 MB (redução de 76 %).

A ocupação de memória do TTS considerando as diferentes versões do sistema de síntese é ilustrada na Fig. 6.1. Quando são utilizados os dois bancos compactados, a ocupação total de memória do TTS é reduzida de 1535 MB para 327 MB (redução de 79 %). As versões em que apenas um dos bancos é compactado possuem ocupação de memória total semelhante, em torno de 900 MB. No entanto, é importante notar que o banco de parâmetros deve ser mantido na memória principal do sistema, de modo que a compressão do banco de parâmetros é mais vantajosa do que a compactação do banco de gravações. O efeito da compressão dos bancos na qualidade do TTS é discutido na Seção 6.2.

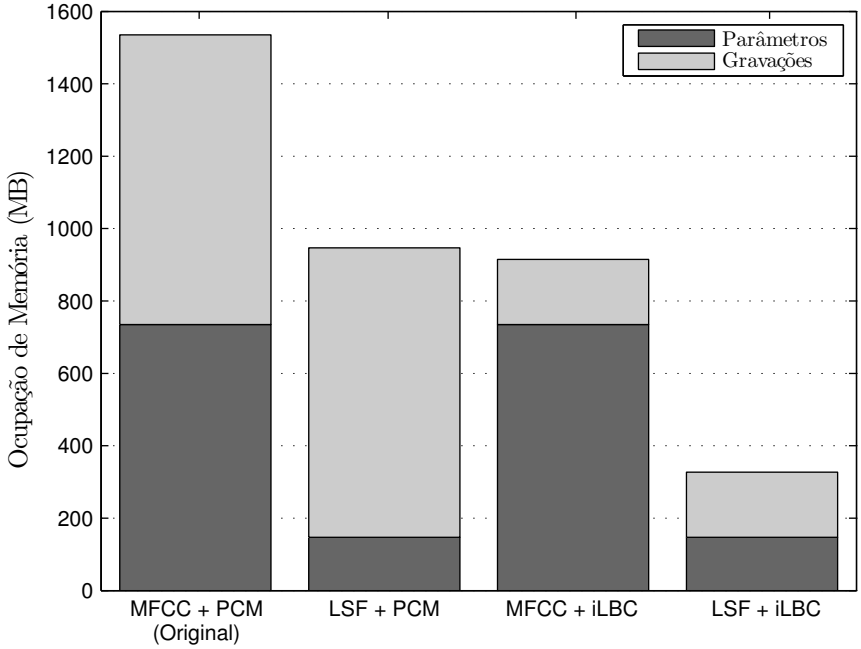

Fig. 6.1: Ocupação de memória do sistema TTS, considerando as diferentes versões do sistema de síntese.

6.2 Avaliação da Qualidade

Nesta seção são descritos os experimentos realizados para verificar o efeito das técnicas de redução de memória na qualidade da fala sintética produzida pelo sistema TTS. Como não há métodos objetivos confiáveis para avaliar a qualidade de sistemas de conversão texto-fala, é necessário realizar um experimento de avaliação perceptual. O experimento realizado é discutido a seguir em detalhes e seus resultados são apresentados e comentados.

6.2.1 Descrição do Experimento

O objetivo dessa avaliação é comparar a qualidade obtida utilizando as diferentes versões do sistema de síntese. Há diferentes métodos de avaliação perceptual que podem ser utilizados para avaliar a qualidade de diferentes sistemas [25]. Para que os resultados das diferentes avaliações possam ser mais facilmente comparados, é importante que as avaliações sejam de mesma natureza. Sendo assim, optou-se por realizar avaliações de qualidade absoluta (*absolute category rating – ACR*) para todas as versões do sistema de síntese.

Avaliações perceptuais de qualidade absoluta consistem em classificar a qualidade de gravações com uma nota absoluta, ou seja, sem comparar cada gravação a uma referência. A qualidade de um sistema é caracterizada pela nota média de um conjunto de avaliações,

denominada *mean opinion score* (MOS). Os critérios de avaliação utilizados no teste considerado são apresentados na Tabela 6.1.

Tabela 6.1: Critérios de avaliação utilizados no teste perceptual

Classificação	**Nota**	**Descrição**
Excelente	5	Qualidade muito boa. Inteligibilidade e naturalidade muito próximas às da fala natural.
Boa	4	Pouco ou nenhum prejuízo à inteligibilidade. Presença de pequenas distorções, perceptíveis porém não desagradáveis. Naturalidade levemente prejudicada.
Razoável	3	Fala ainda inteligível, mas com presença de distorções desagradáveis. Fala, em geral, pouco natural.
Pobre	2	Presença de distorções desagradáveis, prejudicando a inteligibilidade. Naturalidade bastante comprometida.
Ruim	1	Fala severamente distorcida. Inteligibilidade muito prejudicada. Fala extremamente artificial.

Para realizar o teste, foram selecionadas 25 gravações contendo frases lidas pelo locutor presente no banco. As 25 frases foram então sintetizadas, utilizando as quatro versões do sistema de síntese descritas anteriormente. É importante ressaltar que as gravações lidas não foram utilizadas na construção do banco, garantindo que nenhum de seus segmentos tenha sido utilizado na síntese. A energia das gravações sintetizadas foi normalizada, tomando como referência a energia das gravações lidas. Esse procedimento é necessário para que o volume das gravações seja consistente ao longo do teste.

Cada avaliador classificou a qualidade de 25 gravações, sendo 5 sintetizadas com cada versão do sistema de síntese e mais 5 gravações contendo frases lidas. As gravações foram apresentadas aos avaliadores em ordem aleatória. Além do mais, a versão do sistema utilizada para sintetizar uma dada frase também foi escolhida aleatoriamente. As frases não foram repetidas, fazendo com que todos os avaliadores ouvissem as mesmas 25 frases.

Um problema que ocorre em testes perceptuais é o forte efeito subjetivo das avaliações, ou seja: diferentes avaliadores podem ter interpretações significativamente diferentes de conceitos como “ótimo” ou “ruim”, gerando resultados inconsistentes no teste. Para melhorar a consistência das avaliações, um documento de instruções descrevendo os objetivos do teste e os critérios de avaliação foi apresentado aos avaliadores. O documento

continha uma tabela semelhante à Tabela 6.1, mostrando as classificações e descrições correspondentes. Além dessa tabela, foram apresentadas duas gravações, para proporcionar referências de qualidade "excelente" e "ruim". As gravações contêm a mesma frase e foram obtidas da seguinte forma:

1. **Excelente.** Consiste de uma gravação de fala natural.

2. **Ruim.** Obtida por uma versão abandonada do sistema de síntese, com qualidade muito ruim. A qualidade dessa gravação é muito pior do que a obtida pelas versões do sistema de síntese utilizadas no teste, em qualquer situação.

Idealmente, um número muito grande de sessões de avaliação deve ser realizado para garantir a confiabilidade do teste subjetivo. No entanto, esse processo é muito custoso e demorado, de modo que não foi possível realizar diversas sessões de avaliação. Sendo assim, para que algumas avaliações pouco confiáveis não contaminem o resultado do teste, foram consideradas apenas as sessões de avaliação que atendem aos seguintes critérios:

1. Nenhuma gravação do grupo de controle (frases lidas) avaliada com nota menor do que 5 (excelente).

2. Nenhuma gravação avaliada com nota 1 (ruim). Esse critério é justificado pois a gravação apresentada como exemplo de qualidade ruim apresenta distorções muito mais severas do que qualquer gravação produzida pelas versões do sistema de síntese avaliadas.

Os resultados do teste perceptual são apresentados e discutidos na seqüência.

6.2.2 Análise dos Resultados

Ao todo, foram realizadas 20 sessões de avaliação com avaliadores distintos, das quais 10 foram desconsideradas, de acordo com os critérios discutidos anteriormente. As sessões são divididas em dois grupos, a saber:

1. **Conhecedores.** Avaliadores se consideram familiares com sistemas de conversão texto-fala.

2. **Leigos.** Avaliadores não se consideram familiares com sistemas de conversão texto-fala.

A Tabela 6.2 apresenta a distribuição das sessões de avaliação nos dois grupos, considerando as sessões realizadas e as sessões aproveitadas para análise. Para cada versão do

sistema de síntese avaliada, a distribuição das notas é apresentada através de histogramas. São mostradas as distribuições de notas para os dois grupos e a distribuição total. Finalmente, se apresenta uma comparação das notas médias das versões avaliadas. Os resultados do grupo de controle são omitidos, pois todas as avaliações consideradas possuem nota máxima para as gravações desse grupo.

Tabela 6.2: Distribuição das avaliações

Grupo	Avaliações Realizadas	Avaliações Analisadas
Conhecedores	9	6
Leigos	11	4
TOTAL	20	10

As notas do sistema original (banco de parâmetros utilizando coeficientes MFCC e banco de gravações codificado em PCM) são mostradas na Fig. 6.2. Há uma grande quantidade de notas 4 e 5 (classificação boa e excelente, respectivamente). Um aspecto interessante dessa avaliação é a diferença na distribuição das notas para os dois grupos. Avaliadores leigos classificaram a qualidade do sistema como razoável em mais situações do que classificaram como excelente. Já a classificação dada por avaliadores familiares com sistemas TTS tem comportamento inverso: a quantidade de classificações como excelente é muito maior do que a quantidade de classificações como razoável.

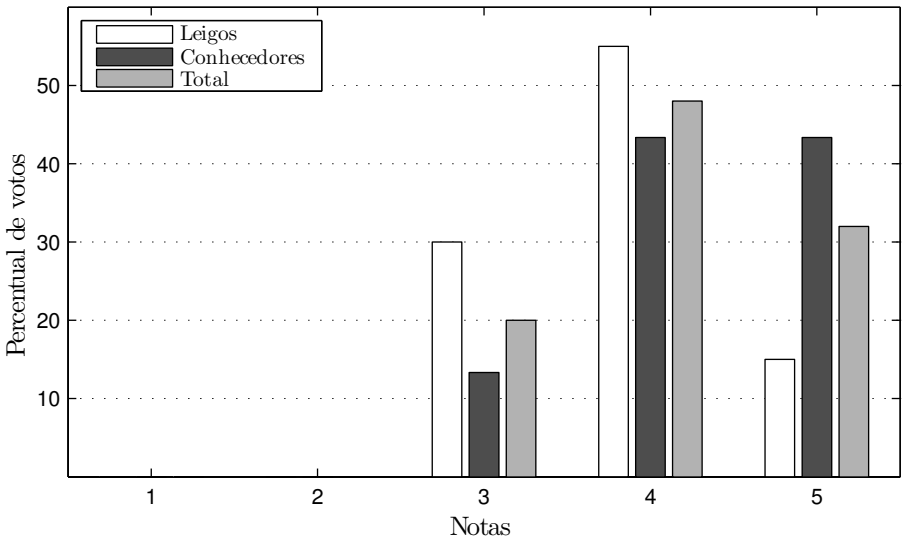

Fig. 6.2: Notas da avaliação perceptual, considerando os bancos de parâmetros e gravações originais (MFCC + PCM).

Na Fig. 6.3 estão apresentadas as notas considerando o banco de gravações original, associado ao banco de parâmetros com compressão. Comparando com as notas do sistema

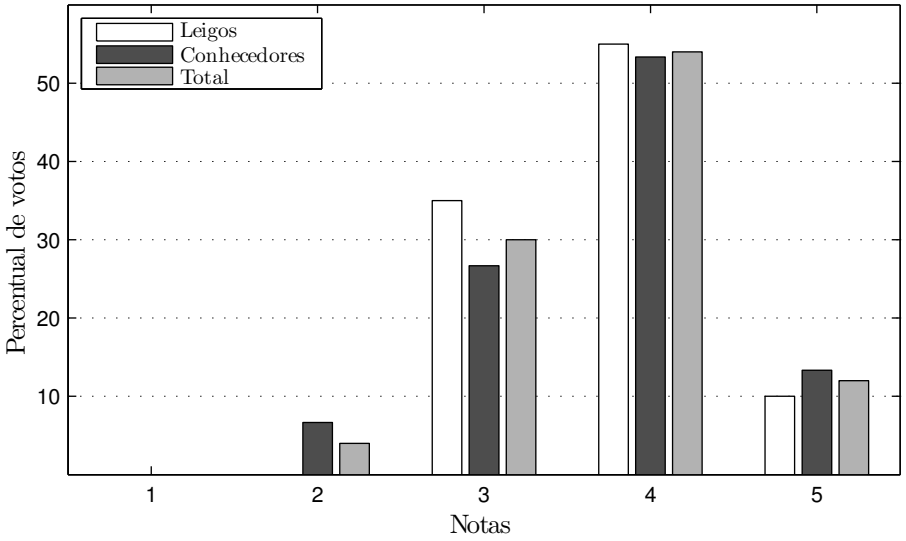

Fig. 6.3: Notas da avaliação perceptual, considerando banco de parâmetros compactado e banco de gravações original (LSF + PCM).

original, há menos notas classificando o sistema como excelente (nota 5) e maior quantidade de notas classificando o sistema como razoável (nota 3). Em um pequeno número de situações, a qualidade do sistema foi considerada pobre por avaliadores familiares com sistemas TTS. Apesar disso, a grande quantidade de classificações avaliando a qualidade do sistema como boa (nota 4) indica que a redução de qualidade causada pela compressão banco de parâmetros é pequena.

As notas para as gravações sintetizadas utilizando o banco de gravações compactado, associado ao banco de parâmetros original estão mostradas na Fig. 6.4. Percebe-se que a compressão do banco de gravações causa maior perda na qualidade do que a compressão no banco de parâmetros, já que nessa situação há um aumento no número de gravações classificadas como pobres (nota 2) e nenhuma gravação considerada excelente (nota 5). A quantidade de notas classificando o sistema como razoável é igual à quantidade de notas classificando como bom para o grupo de avaliadores leigos. O grupo de avaliadores conhecedores apresenta maior quantidade de notas classificando o sistema como razoável.

Na Fig. 6.5, estão apresentadas as notas considerando compressão aplicada aos dois bancos do sistema de síntese. Percebe-se uma redução adicional na qualidade da fala sintética em relação à situação anterior, utilizando o banco de gravações original. Nessa situação, a distribuição de notas para os dois grupos também é diferente: para o grupo de avaliadores leigos, a quantidade de notas classificando o sistema como pobre é muito maior do que a quantidade de notas classificando como bom, enquanto para o grupo de avaliadores familiares com sistemas TTS, não há muita diferença na quantidade de notas para as duas classificações.

As notas médias das diferentes configurações do sistema de síntese considerado estão mostradas na Fig. 6.6. Em geral, a opinião dos dois grupos é consistente, indicando uma pequena redução da qualidade do sistema com a substituição do banco de parâmetros original pelo banco compactado e uma redução maior de qualidade com o uso de compressão no banco de gravações. Avaliadores familiares com sistemas de síntese de fala perceberam maior redução de qualidade com a troca do banco de parâmetros. O grupo de avaliadores leigos praticamente não percebeu mudança nessa situação.

A título de ilustração, as notas médias considerando todas as sessões de avaliação são apresentadas na Fig. 6.7. Embora as notas para todas as versões do sistema de síntese sejam um pouco menores, percebe-se um comportamento semelhante ao das sessões que atenderam aos critérios apresentados anteriormente. Há uma pequena redução de qualidade quando o banco de parâmetros é compactado e a compressão do banco de gravações causa uma redução maior na qualidade do sistema.

6.3 Comentários

As técnicas propostas no trabalho proporcionam uma redução substancial no uso de memória do sistema TTS considerado. Quando apenas o banco de parâmetros é compactado, a redução de qualidade é pequena. No entanto, quando o banco de gravações é compactado, há uma maior degradação na qualidade do sistema.

Na Tabela 6.3 estão apresentadas a ocupação de memória final do sistema e a nota média do teste perceptual para as quatro variações do sistema de síntese. O uso do banco

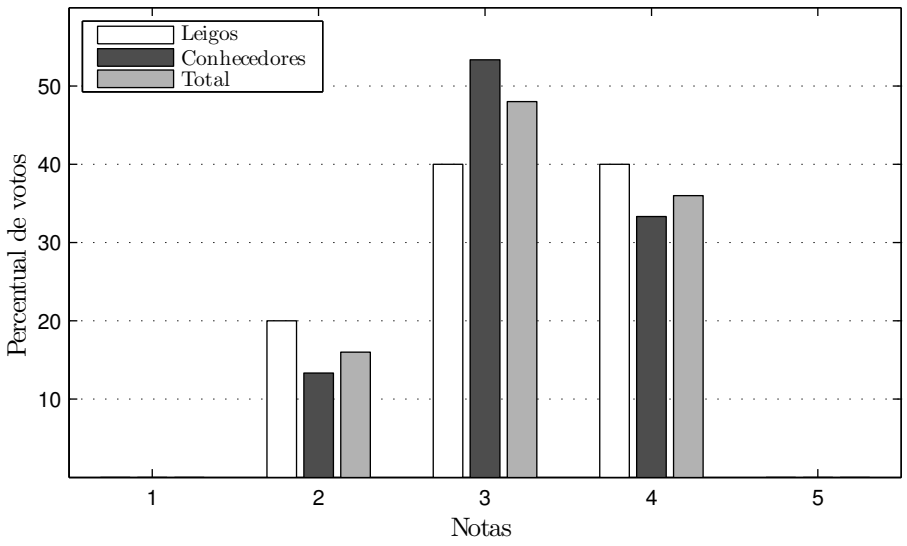

Fig. 6.4: Notas da avaliação perceptual, considerando bancos de parâmetro original e banco de gravações comprimido (MFCC + iLBC).

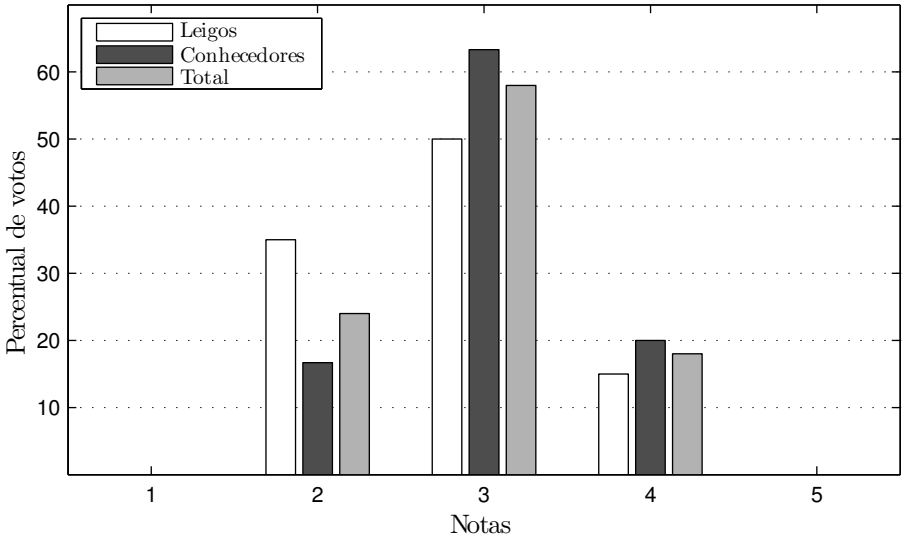

Fig. 6.5: Notas da avaliação perceptual, considerando os bancos de parâmetros e gravações compactados (LSF + iLBC).

de parâmetros compactado proporciona considerável redução de memória, com pequena redução de qualidade. Além do mais, a compressão do banco de parâmetros proporciona redução no uso de memória principal do sistema. Dessa maneira, há um grande benefício para o desempenho do TTS, sem grande redução na qualidade. Utilizar apenas o banco de gravações compactado não faz muito sentido, pois a ocupação de memória total é semelhante à do sistema que utiliza o banco de parâmetros compactado, com maior redução de qualidade. Além do mais, com o banco de gravações original, a ocupação de memória principal ainda é elevada. Apesar de causar uma redução sensível na qualidade da fala sintética, a versão do sistema de síntese com compressão nos dois bancos possui

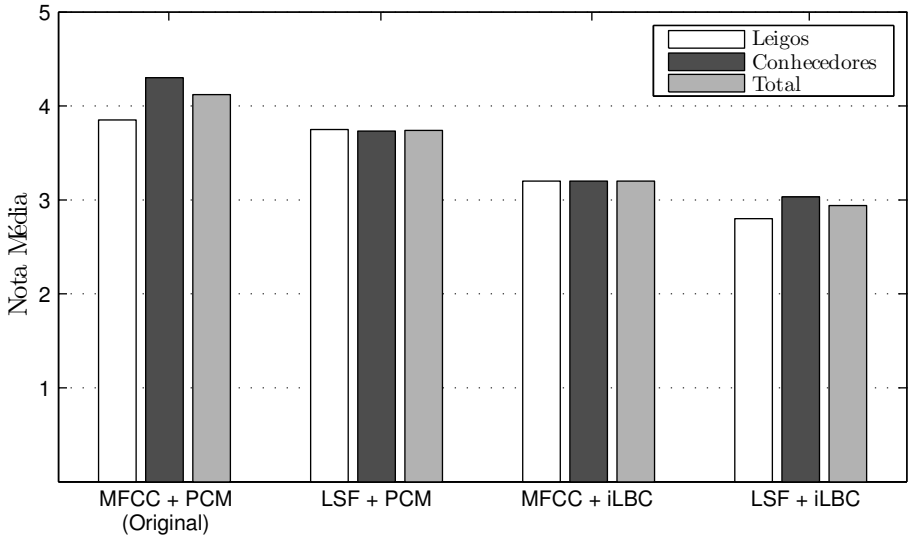

Fig. 6.6: Resultado médio da avaliação perceptual, comparando as diferentes abordagens.

uma carga de memória consideravelmente menor do que a do sistema original, podendo até viabilizar o uso do sistema considerado em aplicações com baixa disponibilidade de memória principal e de armazenamento em memória secundária.

Tabela 6.3: Redução de ocupação de memória

Situação	Ocupação de memória [MB]			Nota média
	Parâmetros	Gravações	Total	
MFCC + PCM	735	800	1535	4,16
LSF + PCM	147	800	947	3,80
MFCC + iLBC	735	180	915	3,20
LSF + iLBC	147	180	327	2,96

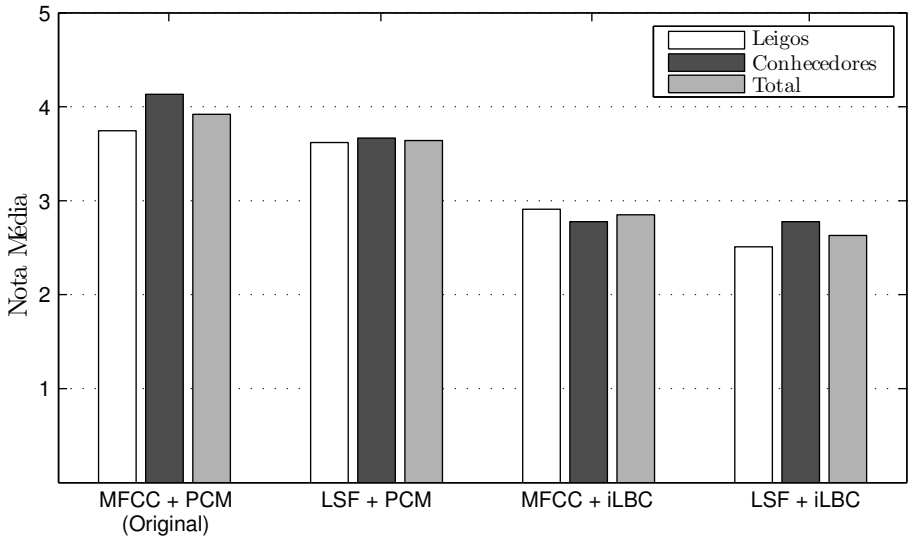

Fig. 6.7: Resultado médio da avaliação perceptual, comparando as diferentes abordagens (considerando todas as avaliações).

7 Considerações Finais

O processo de conversão texto-fala permite produzir fala sintética tomando como entrada um texto qualquer. Esse procedimento é dividido em duas etapas principais: o processamento lingüístico e a síntese do sinal de fala. As funções de processamento lingüístico são responsáveis por converter um dado texto em uma representação lingüística detalhada e estruturada, descrevendo os sons que devem ser gerados para produzir a fala sintética. O sintetizador de fala então produz o sinal de fala artificial, baseado nas informações fornecidas pelas funções de processamento lingüístico.

O processo de síntese de fala pode ser realizado por diferentes métodos, sendo que os melhores resultados têm sido obtidos através da abordagem de síntese concatenativa. Atualmente, os sistemas TTS de estado-da-arte são construídos utilizando grandes bancos de gravações, utilizando algoritmos de seleção de unidades para determinar os trechos que devem ser concatenados. Utilizando bancos de longa duração construídos apropriadamente (com grande diversidade de contextos fonéticos e prosódicos) é possível obter fala sintética com inteligibilidade e naturalidade muito próximas às da fala humana.

Um problema que ocorre em sistemas de conversão texto-fala atuais é a grande ocupação de memória, devido aos grandes bancos utilizados para a síntese. Desse modo, há dificuldade de utilizar tais sistemas em situações com restrição de memória.

O sistema TTS considerado nesse trabalho utiliza um banco de gravações com duração total de 28 horas, associado a um banco contendo parâmetros utilizados para o cálculo de custos no processo de seleção de unidades para a síntese. A longa duração do banco permite que o sistema tenha qualidade muito boa, mas torna sua ocupação de memória bastante elevada, aproximadamente 1,6 GB.

A ocupação de memória do banco de parâmetros é causada, em grande parte, por coeficientes representando o envelope espectral das unidades do banco, utilizados para o cálculo do custo de concatenação no processo de seleção de unidades. O banco de parâmetros deve ser mantido na memória principal do sistema em que o TTS é executado, para proporcionar desempenho adequado. Desse modo, a compressão do banco de parâmetros é muito vantajosa, permitindo a utilização do TTS em equipamentos com capacidade

de memória principal limitada. Também é possível suportar mais de uma voz (ou seja, múltiplos bancos) em equipamentos que atualmente utilizam o sistema TTS. As técnicas propostas nesse trabalho permitiram reduzir em 80 % a ocupação de memória do banco de parâmetros. Além do mais, avaliações perceptuais indicam que a perda de qualidade causada pela compressão do banco de parâmetros é pequena.

Nesse trabalho, a função custo associada aos parâmetros espectrais do banco de parâmetros é considerada como a distância euclidiana entre os vetores de coeficientes espectrais. Uma sugestão para trabalhos futuros é avaliar o impacto de diferentes funções distância na qualidade da fala sintética. O uso de quantização vetorial permite pré-calcular as distâncias de todos os vetores de parâmetros, permitindo o uso de funções distância sofisticadas, com elevado custo computacional, sem prejuízo ao desempenho do sistema.

O banco de gravações pode ser compactado utilizando algoritmos de codificação de fala. Nesse trabalho, optou-se por utilizar algum *codec* já existente. No entanto, a maior parte dos *codecs* disponíveis não é adequada para aplicações em síntese de fala, por não proporcionarem capacidade de acesso aleatório aos dados codificados. Em sua grande maioria, os *codecs* são desenvolvidos para aplicações de comunicação em tempo real, fazendo uso de informações de quadros passados para melhorar a eficiência de codificação. Desse modo, caso algum quadro seja perdido, há a ocorrência de grandes distorções, que se propagam por longos períodos.

Um *codec* existente e apropriado para aplicação em síntese de fala é o iLBC. Esse *codec* foi concebido para aplicações de comunicação de voz sobre IP (VoIP), em que a perda de quadros é comum. Desse modo, o *codec* foi desenvolvido para ser robusto à perda de quadros. Essa característica pode ser aproveitada com vantagem em aplicações de síntese de fala, proporcionando capacidade de acesso aleatório ao banco de gravações. Sendo assim, o banco de gravações do TTS considerado nesse trabalho foi codificado utilizando uma versão modificada do iLBC, reaproveitando os coeficientes espectrais armazenados no banco de parâmetros para reduzir a taxa de codificação. Dessa maneira, foi possível obter uma redução de 76 % na ocupação de memória do banco de gravações.

Dentre os *codecs* disponíveis, considera-se que o iLBC seja o mais apropriado para aplicações em síntese de fala. No entanto, as avaliações perceptuais realizadas mostraram que a codificação do banco de gravações utilizando o iLBC causa uma perda considerável na qualidade da fala sintética, indicando que a codificação do banco de gravações pode ser melhorada. Desse modo, outra sugestão para trabalhos futuros é conceber um *codec* específico para aplicações em síntese de fala. Em geral, os *codecs* com foco em aplicações de comunicação em tempo real são desenvolvidos visando um compromisso entre a qualidade,

a redução na taxa de bits e a complexidade computacional. Para um *codec* utilizado em aplicações de síntese de fala, não há grandes restrições na complexidade computacional da etapa de codificação. Sendo assim, os requisitos para um *codec* visando aplicações em síntese concatenativa são boa qualidade, reduzida taxa de bits e capacidade de acesso aleatório. Uma característica interessante para um possível *codec* para aplicações em síntese de fala é suportar também fala em banda larga (taxa de amostragem de 16 kHz), através de algoritmos de extensão de banda.

Um ponto que deve ser ressaltado é a avaliação de qualidade de sistemas de síntese de fala. Como não existem métodos objetivos para determinação da qualidade de fala sintética, é necessário realizar testes perceptuais. No entanto, tais avaliações carregam um forte efeito subjetivo. Idealmente, um grande número de avaliações, com diferentes pessoas, deve ser realizado para que as diferenças de interpretação dos conceitos sejam suavizadas. No entanto, avaliações perceptuais são demoradas e custosas, de modo que, na maioria das vezes, os testes são realizados com um reduzido número de avaliadores. Sendo assim, é necessário utilizar critérios para desconsiderar avaliações que fujam muito do comportamento esperado, tais como inserir gravações de controle no teste e utilizar gravações de referência.

Apesar de causar reduções perceptíveis na qualidade da fala sintética, as técnicas propostas nesse trabalho são um passo para viabilizar o uso de sistemas de conversão texto-fala em aplicações com reduzida capacidade de memória. Associados ao desenvolvimento dos processadores e memória, avanços na direção de reduzir o custo computacional de sistemas de conversão texto-fala de estado-da-arte possibilitam campos cada vez mais vastos para a aplicação de fala sintética de boa qualidade.

Referências

[1] *Pulse Code Modulation (PCM) of Voice Frequencies*, Document ITU-T Rec. G.711, International Telecommunication Union (ITU), Geneva, Switzerland, 1988.

[2] O. van der Vrecken, N. Pierret, T. Dutoit, V. Pagel, and F. Malfrere, "New techniques for the compression of synthesizer databases," in *Proc. 1997 IEEE International Symposium on Circuits and Systems*, vol. 4, Hong Kong, Jun. 1997, pp. 2641–2644.

[3] C.-H. Lee, S.-K. Jung, and H.-G. Kang, "Applying a speaker-dependent speech compression technique to concatenative TTS synthesizers," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 15, no. 2, pp. 632–640, Feb. 2007.

[4] S. Andersen, W. Kleijn, R. Hagen, J. Linden, M. Murthi, and J. Skoglund, "iLBC - a linear predictive coder with robustness to packet losses," in *Proc. IEEE Workshop Speech Coding*, Tsukuba City, Japan, Oct. 2002, pp. 23–25.

[5] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden, "Internet Low Bit Rate Codec (iLBC)," RFC 3951 (Experimental), Internet Engineering Task Force, Dec. 2004. Disponível em: http://www.ietf.org/rfc/rfc3951.txt

[6] F. S. Pacheco, "Técnicas de Processamento de Sinais para Alteração de Parâmetros Prosódicos Aplicadas a um Sistema de Conversão Texto-Fala para a Língua Portuguesa Falada no Brasil," Dissertação de Mestrado, Universidade Federal de Santa Catarina, Abr. 2001.

[7] M. V. Nicodem, "Detecção e Tratamento de Cliques Naturais em Bancos de Fala Visando Síntese Concatenativa de Alta Qualidade," Dissertação de Mestrado, Universidade Federal de Santa Catarina, Jan. 2006.

[8] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[9] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, 1990.

[10] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. New York: Wiley, 2003.

[11] F. Egashira, "Síntese de Voz a Partir de Texto para a Língua Portuguesa," Dissertação de Mestrado, Universidade Estadual de Campinas, Jul. 1992.

[12] D. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971–995, 1980.

[13] J. Hogberg, "Data driven formant synthesis," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 565–568.

[14] K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[15] E. Moulines and W. Verhelst, "Time-domain and frequency-domain techniques for prosodic modification of speech," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Amsterdam: Elsevier, 1995.

[16] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. International Conference on Acoustics, Speech, and Signal Processing, 1996*, vol. 1, May 1996, pp. 373–376 vol. 1.

[17] *40,32,25,16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*, Document ITU-T Rec. G.726, International Telecommunication Union (ITU), Geneva, Switzerland, 1990.

[18] *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, Document ITU-T Rec. G.729, International Telecommunication Union (ITU), Geneva, Switzerland, 1996.

[19] A. M. Kondoz, *Digital Speech Coding for Low Bit Rate Communication Systems*, 2nd ed. New York: Wiley, 2004.

[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, ser. Prentice-Hall Signal Processing Series. Englewood Cliffs, NJ: Prentice Hall, 1978.

[21] K. Paliwal and W. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Amsterdam: Elsevier, 1995.

[22] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, San Diego, USA, Mar. 1984, pp. 37–40.

[23] A. Buzo, A. J. Gray, R. Gray, and J. Markel, "Speech coding based upon vector quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 562–574, Oct 1980.

[24] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[25] *Methods for subjective determination of transmission quality*, Document ITU-T Rec. P.800, International Telecommunication Union (ITU), Geneva, Switzerland, 1996.