Compressão de Bancos de Fala para Sistemas de Síntese Concatenativa de Alta Qualidade

Augusto Henrique Hentz e Rui Seara

***Resumo*— Neste artigo, são apresentadas técnicas para reduzir a ocupação de memória de sistemas de conversão texto-fala concatenativos sem comprometer significativamente a qualidade da fala sintética. Para comprimir o banco de gravações mantendo a capacidade de decodificação parcial de segmentos, o que é fundamental para a aplicação em síntese concatenativa, propõe-se o uso do *codec* iLBC, que codifica quadros do sinal de fala de forma independente. O custo de concatenação no processo de seleção de unidades é calculado usando parâmetros LSF quantizados vetorialmente. A aplicação da abordagem proposta em um sistema de conversão texto-fala para o português brasileiro proporciona uma redução de até 76% na ocupação de memória. Avaliações perceptuais indicam que a quantização vetorial dos parâmetros para cálculo do custo de concatenação não causa perda significativa na qualidade da fala sintética.**

***Palavras-Chave*— *Codec* iLBC, compressão de sinais de fala, conversão texto-fala, LSFs.**

***Abstract*— This paper presents techniques to reduce the memory usage in text-to-speech synthesis systems, without significantly affecting the synthetic speech quality. To compress the speech database, while keeping the partial decoding capability, the iLBC codec is used, which allows to encode speech frames independently. The concatenation cost in the unit selection process is evaluated by using vector quantized LSF coefficients. The proposed approach yields up to a 76 % memory reduction in a Brazilian portuguese TTS system. Perceptual evaluations show that vector quantization of the parameters used to calculate the concatenation cost cause no significant loss in the synthetic speech quality.**

***Keywords*— iLBC codec, speech compression, text-to-speech (TTS) conversion, LSFs.**

I. Introdução

Sistemas atuais de conversão texto-fala (*text-to-speech – TTS*), em sua grande maioria, sintetizam o sinal de fala através da concatenação de trechos de gravações extraídos de um banco de dados segmentado e transcrito foneticamente. Para que tais sistemas produzam fala sintética de qualidade satisfatória (inteligibilidade e naturalidade próximas à fala humana), é necessário que o banco de gravações contenha diversos exemplos de um grande número de contextos fonéticos. Isso faz com que o banco tenha grande duração, em geral da ordem de diversas horas, tornando a ocupação de memória de tais sistemas muito elevada. Por exemplo, o Orador, sistema de síntese de fala desenvolvido no LINSE e objeto de estudo deste trabalho, possui um banco de gravações de 28 horas, que ocupa 800 MB quando amostrado à taxa de 8 kHz e quantizado não linearmente através da lei A.

Uma alternativa para reduzir a ocupação de memória de tais sistemas (mantendo a qualidade da fala sintética) é através da compressão do banco de gravações, utilizando alguma técnica de codificação de fala. No entanto, a compressão do banco de gravações, visando síntese de fala, apresenta algumas características particulares que a distinguem das aplicações mais usuais, tais como em sistemas de comunicação em tempo real. A etapa de codificação pode ter alta complexidade computacional, já que não precisa ser realizada em tempo real. Por sua vez, a complexidade da etapa de decodificação deve ser baixa, para não impactar demasiadamente o custo computacional do sistema de síntese. Por último, a característica mais importante é que, no processo de síntese, o banco de gravações não é acessado seqüencialmente, mas de maneira aleatória. Desse modo, o algoritmo utilizado para compressão deve permitir a decodificação também de forma aleatória.

Algoritmos de compressão baseados em predição linear e análise por síntese permitem a codificação do sinal de fala com muito boa qualidade, utilizando reduzidas taxas de bits. Para tal, grande parte desses algoritmos exploram correlações de longo termo do sinal de fala [1]. Uma abordagem bastante comum é o uso de *codebook* adaptativo, que permite utilizar 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, haverá grandes distorções até que o estado do decodificador sincronize com o estado do codificador. Desse modo, *codecs* que tiram proveito de correlações entre quadros consecutivos não são adequados quando é necessário realizar decodificação parcial de dados codificados, tornando-se inviáveis para compressão de um banco de gravações para aplicação em 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 devido à 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*

Augusto Henrique Hentz e Rui Seara, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mails: {augusto, seara}@linse.ufsc.br.

Este trabalho foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e pela empresa Dígitro Tecnologia Ltda.

adaptativo do *codec* G.729 por um *codebook* fixo de sinais de excitação, treinado utilizando gravações do próprio banco do sistema de síntese .

Neste trabalho, propõe-se a utilização do *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 desse *codec* é a independência entre quadros, permitindo reduzir distorções decorrentes da perda de pacotes, podendo assim ser usada 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 metadados, também ocupa grande quantidade de memória, principalmente devido a parâmetros espectrais armazenados para cada unidade (fonema) do banco de gravações. Reduzir a ocupação de memória desse conjunto de dados é outro objetivo deste trabalho.

Este artigo está organizado como segue. Na Seção II, são revisitados os conceitos fundamentais utilizados, tais como características da representação LSF dos coeficientes de predição linear e alguns detalhes sobre o *codec* iLBC. A Seção III apresenta a arquitetura do sistema de síntese de fala considerado, bem como as alterações propostas visando reduzir a ocupação de memória. Os resultados experimentais são mostrados e discutidos na Seção IV. Finalmente, a Seção V apresenta as conclusões do trabalho.

II. Revisão de Conceitos Fundamentais

A. Representação LSF dos Coeficientes da Predição Linear

Predição linear é uma técnica para redução de redundância, utilizada sobretudo em codificação de sinais de fala visando compressão [6]. O conceito fundamental dessa técnica consiste em obter uma aproximação $\hat{s}(n)$ de um dado sinal de fala $s(n)$ através da combinação linear de suas M amostras passadas. Assim,

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

onde os coeficientes a_i caracterizam o preditor linear. É possível representar os coeficientes de predição linear de diferentes maneiras. Uma delas, bastante usada por sua robustez à quantização, é a representação dos a_i por parâmetros LSF (*line spectral frequency*). Esses parâmetros são obtidos a partir das raízes dos polinômios $P(z)$ e $Q(z)$, derivados do polinômio de predição linear $A(z) = 1 + a_1 z^{-1} + \cdots + a_M z^{-M}$ através das seguintes relações:

$$P(z) = A(z) + z^{-M-1} A(z)$$
$$Q(z) = A(z) - z^{-M-1} A(z).$$

Uma propriedade importante desta representação 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 de precedência:

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

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. 1. 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 [7].

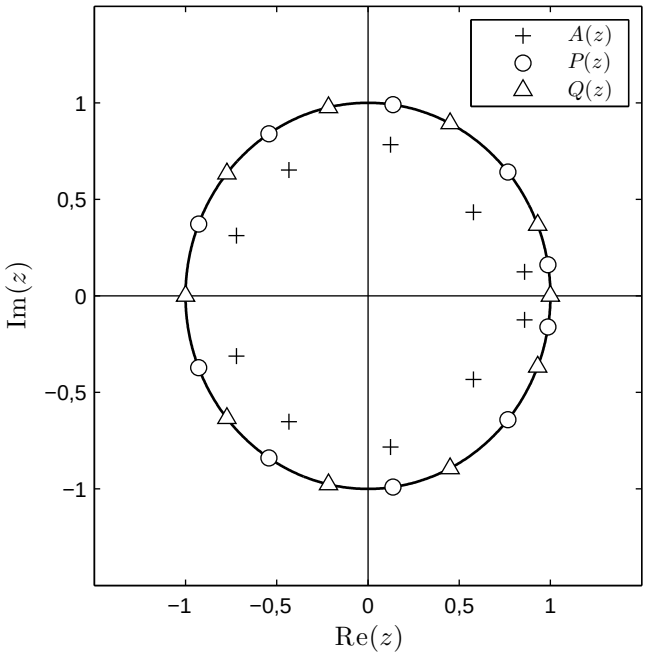

Fig. 1. Propriedade de intercalamento dos coeficientes LSF.

Como as raízes de $P(z)$ e $Q(z)$ estão sobre a circunferência de raio unitário, é suficiente representá-las por seus ângulos. Além disso, como as raízes se apresentam em pares complexos conjugados, os coeficientes LSF são caracterizados pelo ângulo das raízes de $P(z)$ e $Q(z)$ localizadas no semiplano superior. Essa representação dos coeficientes LPC é bastante utilizada, devido aos seguintes aspectos:

i) Ao contrário dos coeficientes de $A(z)$, os parâmetros LSF têm faixa dinâmica limitada, o que os torna mais adequados para a quantização.
ii) Desde que a propriedade de intercalamento seja mantida, erros de quantização não tornam o filtro de síntese $1/A(z)$ instável.
iii) Parâmetros LSF podem ser interpolados. Essa característica é aproveitada em *codecs* que utilizam a estratégia de divisão de quadros em subquadros. Assim, os coeficientes LSF de cada subquadro são obtidos através da interpolação dos coeficientes calculados no quadro correspondente.

B. Codec iLBC

O iLBC é um *codec* robusto a situações em que há perda de quadros. Para obter tal robustez, o iLBC não explora as correlações de longo termo entre quadros distintos. Tal característica é importante para nossa aplicação, pois permite a decodificação parcial de trechos do banco de gravações.

O *codec* pode comprimir sinais de fala de banda estreita (amostrados em 8 kHz) em duas taxas: 15,2 kbps (utilizando quadros de 20 ms) ou 13,33 kbps (quadros de 30 ms). Para tal, o sinal é dividido em quadros de análise, que são então codificados de forma independente. Assim, ao contrário de outros algoritmos de compressão de fala, os parâmetros contidos em um dado quadro codificado pelo iLBC são suficientes para reconstruir uma réplica do sinal original sem grandes distorções [4]. Detalhes sobre a implementação desse *codec* podem ser encontrados em [5].

III. Redução da Ocupação de Memória do TTS

Nesta seção, são apresentadas as propostas para reduzir a ocupação de memória do sistema de síntese de fala.

A. Arquitetura do Sistema TTS

Considera-se, neste trabalho, um sistema de conversão texto-fala para o português brasileiro, baseado na técnica de síntese concatenativa por seleção de unidades. O fluxograma do sistema TTS considerado é ilustrado na Fig. 2. O texto a ser sintetizado passa por um módulo de processamento lingüístico, que o transforma em uma seqüência de fonemas. Essa seqüência de fonemas passa por um módulo de seleção de unidades, responsável pela escolha das unidades do banco mais apropriadas para a síntese. Finalmente, as unidades selecionadas são extraídas do banco de gravações e concatenadas, dando origem ao sinal de fala sintetizado.

O banco de metadados indicado no fluxograma da Fig. 2 é um conjunto de informações de unidades do banco de gravações utilizadas para determinar a seqüência mais apropriada para o processo de síntese. Nesse banco, estão contidas informações sobre contexto (posição da unidade na frase, se a unidade é parte de uma sentença interrogativa ou exclamativa, etc.), contorno de *pitch* e parâmetros espectrais. Para a obtenção da melhor seqüência a ser buscada, determina-se um custo de concatenação, baseado em fatores de contexto, diferença de *pitch* e distância espectral, esta última definida como a distância euclidiana entre os vetores de parâmetros espectrais das unidades. O banco de metadados ocupa aproximadamente 700 MB, de modo que a ocupação total de memória do sistema, incluindo o banco de gravações de 800 MB, é da ordem de 1,5 GB. A ocupação de memória do banco de metadados se deve em grande parte aos parâmetros espectrais, compostos por 12 coeficientes mel-cepstrais (*mel-frequency cepstral coefficients – MFCCs*) para cada quadro do banco de gravações e armazenados na forma de um vetor de números em ponto flutuante com precisão simples (32 bits). Os quadros correspondem aos períodos de *pitch* do sinal de fala em trechos vozeados, enquanto, nos demais trechos, apresentam duração fixa de 10 ms. Como as informações

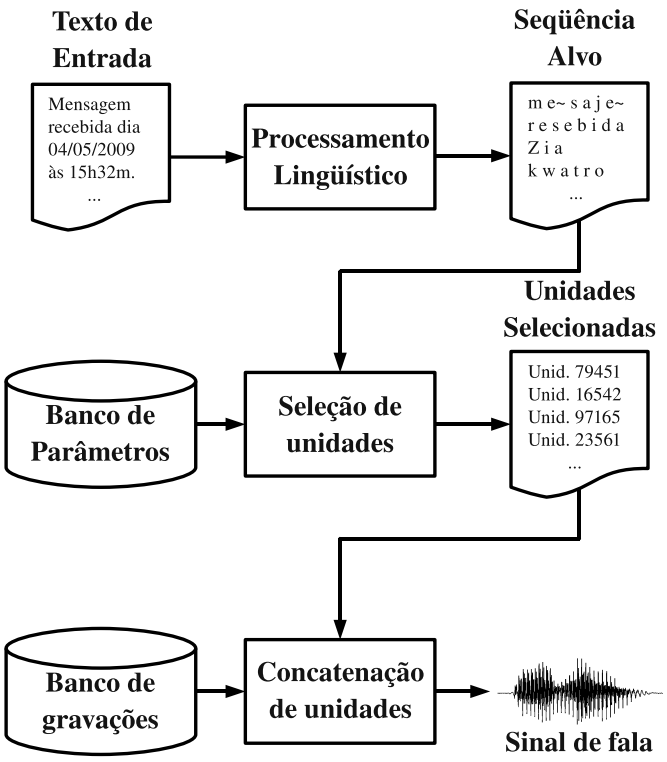

Fig. 2. Fluxograma do sistema de conversão texto-fala.

presentes no banco de metadados são utilizadas no processo de seleção de unidades para síntese, elas são acessadas com uma freqüência muito maior do que os dados do banco de gravações. Desse modo, para um melhor desempenho do sistema de síntese, é necessário que o banco de metadados seja mantido na memória principal do sistema, que proporciona acesso mais rápido do que o armazenamento em disco. Assim, reduzir a ocupação de memória do banco de metadados passa a ser um ponto de grande importância.

B. Alterações Propostas

Nesta seção, são apresentadas as alterações propostas à arquitetura do sistema de conversão texto-fala considerado, para reduzir sua ocupação de memória.

Propõe-se substituir os 12 coeficientes MFCC do banco de metadados por 10 coeficientes LSF. Como é sabido, os parâmetros LSF são mais apropriados para serem quantizados, pois possuem reduzida faixa dinâmica. Dessa forma, visando reduzir substancialmente a ocupação de memória associada a esses coeficientes, utiliza-se quantização subvetorial, com subvetores de dimensão 3, 3 e 4 quantizados com 8 bits cada. É interessante notar que, como os parâmetros espectrais são quantizados vetorialmente, torna-se possível pré-calcular todas as possíveis distâncias espectrais, de modo que o cálculo da distância entre quadros no processo de seleção de unidades possa ser substituído pelo acesso a uma tabela. Além disso, como os parâmetros LSF armazenados são de mesma ordem dos utilizados no *codec* iLBC, é possível reduzir ainda mais a ocupação de memória do banco de gravações, aproveitando

os coeficientes armazenados no banco de metadados para sua codificação. Assim, é necessário converter os vetores de coeficientes do banco de metadados (calculados a cada período de *pitch*) em vetores de coeficientes para a taxa de quadros do *codec* iLBC. Tal conversão pode ser realizada de forma eficiente através de interpolação linear. Desse modo, é necessário introduzir uma alteração ao *codec* iLBC, substituindo os blocos de análise LPC do codificador por um módulo de interpolação dos parâmetros do banco de metadados. Naturalmente, é também necessário incluir um módulo equivalente no decodificador.

Para a compressão do banco de gravações, propõe-se o uso do *codec* iLBC, com taxa de 15,2 kbps.

IV. Resultados Experimentais

Nesta seção, são apresentados os resultados experimentais da compressão dos bancos de dados do sistema de síntese de fala, como também uma análise da qualidade do sistema modificado.

A. Redução de Ocupação de Memória

A possibilidade de substituição dos coeficientes MFCC por LSFs quantizados vetorialmente no cálculo da distância espectral entre segmentos foi verificada através de testes preliminares. Resultados avaliando o impacto dessa estratégia na qualidade do sistema são discutidos na Seção IV-B. Além do mais, o uso de parâmetros LSF quantizados permite reduzir substancialmente a ocupação de memória do banco de metadados, que passa a ocupar 147 MB ao invés dos 735 MB originais (redução de 80 %). Isso ocorre porque os parâmetros espectrais de cada quadro, antes codificados com $12 \times 32 = 384$ bits, passam a ser codificados em 24 bits (redução de 93 %). É importante salientar que ainda há outras informações contidas no banco de metadados, não comprimidas pela técnica aqui proposta.

A capacidade de decodificação parcial de segmentos do banco codificado com o *codec* iLBC sem distorções, citada anteriormente, é válida. Além disso, foi verificada a possibilidade de reaproveitar os parâmetros LSF do banco de metadados na codificação do banco de gravações. Para tal, basta interpolar os coeficientes, extraídos de acordo com os períodos de *pitch* do sinal de fala, de modo a obter os parâmetros de cada quadro do iLBC. Essa técnica permite uma pequena redução na taxa de bits do *codec*, para 14,4 kbps quando se utilizam quadros de 20 ms e 12 kbps quando os quadros são de 30 ms (as taxas originais são de 15,2 kbps e 13,33 kbps, respectivamente). Utilizando o *codec* iLBC modificado com taxa de 14,4 kbps, a ocupação de memória do banco de gravações é reduzida para 180 MB. Os resultados de redução de memória obtidos para o sistema de síntese de fala estão resumidos na Tabela I.

TABELA I
REDUÇÃO DE OCUPAÇÃO DE MEMÓRIA

Banco	Ocupação de memória (MB)		Redução (%)
	Original	Comprimido	
Metadados	735	147	80
Gravações	800	180	76
TOTAL	1535	327	79

B. Análise da Qualidade do Sistema Modificado

Para verificar o impacto da troca dos parâmetros MFCC por LSFs na qualidade do sistema de síntese, foi realizado um teste perceptual comparativo (*comparison category rating – CCR*) de acordo com a Recomendação ITU-T P.800 [8]. É importante salientar que o objetivo dessa avaliação é verificar somente o efeito da troca dos parâmetros para cálculo de distância espectral, já que a degradação de qualidade causada pelo iLBC é conhecida [4]. Desse modo, utiliza-se o banco de gravações original ao invés do banco compactado. Neste teste, 20 pares de sentenças sintetizadas utilizando os parâmetros MFCC e LSF são apresentados a avaliadores, que julgam a qualidade da segunda gravação em relação à da primeira. Os pares de gravações são reproduzidos em ordem aleatória, sendo que em metade dos pares a primeira gravação é sintetizada utilizando MFCCs e, nos demais, LSFs (quando a primeira gravação é gerada com LSFs, a classificação dada pelo avaliador deve ser invertida). Os avaliadores então classificam a qualidade relativa das gravações utilizando a escala da Tabela II, que também mostra a nota correspondente a cada classificação.

TABELA II
CLASSIFICAÇÕES E NOTAS DO TESTE CCR

Qualidade relativa	Nota
Muito melhor	3
Melhor	2
Pouco melhor	1
Semelhante	0
Pouco pior	−1
Pior	−2
Muito pior	−3

O teste foi realizado por 13 avaliadores (6 familiares com sistemas de síntese de fala e 7 não familiares), que escutaram as gravações utilizando fones de ouvido tendo a opção de reproduzir cada par de gravações tantas vezes quantas desejarem. O resultado das avaliações é mostrado na Fig. 3. Percebe-se que, na maioria dos casos, o sistema proposto foi classificado como tendo qualidade semelhante (nota 0) ou pouco pior (nota −1). Os valores médios das notas são mostrados na Tabela III. As notas maiores do que −0,5 indicam que, na média, a qualidade do sistema proposto é considerada semelhante à do sistema original.

TABELA III
NOTAS MÉDIAS DO TESTE CCR

Avaliadores	Nota média
Familiares com TTS	−0,40
Não Familiares com TTS	−0,46
TOTAL	−0,43

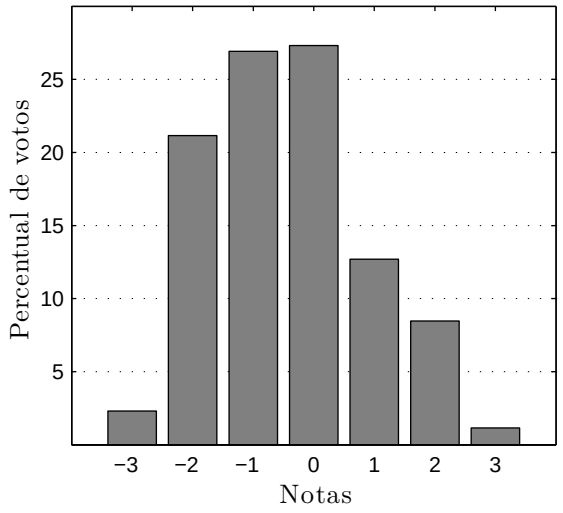

Fig. 3. Resultado da avaliação da qualidade do sistema proposto em relação à do original.

V. Comentários e Conclusões

O objetivo deste trabalho é a redução da ocupação de memória em um sistema de síntese de fala baseado em seleção de unidades. A ocupação de memória do sistema considerado se dá principalmente pelo banco de gravações e pelo conjunto de parâmetros espectrais utilizado no processo de seleção de unidades para a síntese.

Os parâmetros espectrais, originalmente MFCCs, foram substituídos por LSFs quantizados vetorialmente. Tal substituição proporcionou redução de 80% na ocupação de memória do banco de metadados, com fala sintética de qualidade semelhante à gerada pelo sistema original.

O banco de gravações foi comprimido com o *codec* iLBC, por permitir decodificação parcial de qualquer trecho do banco, característica fundamental para o processo de síntese concatenativa. Ainda, foi possível reaproveitar os coeficientes LSF armazenados no banco de metadados para a compressão do banco de gravações. As alterações propostas proporcionaram uma redução de 79 % na ocupação de memória do sistema considerado. Avaliações perceptuais indicam que a substituição dos coeficientes MFCC por LSFs quantizados vetorialmente no cálculo do custo de concatenação não causa perda significativa na qualidade da fala sintética.

Para continuação deste trabalho, propõem-se estudos para suportar bancos de voz em banda larga (taxa de amostragem de 16 kHz) e avaliação de outros tipos de distância espectral entre unidades do banco, tal como a distorção espectral de Itakura-Saito.

Referências

[1] J.-H. Chen and J. Thyssen, "Analysis-by-synthesis speech coding," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2008, pp. 351–392.

[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 Int. Symp. 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 Trans. Acoust., Speech, Language Process.*, 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. [Online]. Available: http://www.ietf.org/rfc/rfc3951.txt

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

[7] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 9, San Diego, USA, Mar 1984, pp. 37–40.

[8] "ITU-T recommendation P.800: Methods for subjective determination of transmission quality," International Telecommunication Union, Tech. Rep., August 1996.