TÉCNICAS DE CODIFICAÇÃO DE FALA BASEADAS EM ANÁLISE POR SÍNTESE

Speech Codification Techniques Based on Analysis-by-Synthesis

ULIANI NETO, Mário
Faculdade de Jaguariúna; Faculdade Politécnica de Campinas; Fundação CPqD

VIOLATO, Ricardo Paranhos Velloso
Fundação CPqD

SIMÕES, Flávio Olmos
Fundação CPqD

COSTA, Bruno Ribeiro
Fundação CPqD

Resumo: O presente artigo apresenta uma revisão bibliográfica de técnicas de codificação de sinais de fala baseadas em um paradigma conhecido na área de processamento de sinais como análise por síntese. Os codificadores de análise por síntese são baseados em um modelo fonte-filtro de produção de fala; a extração dos parâmetros da fonte e do filtro é feita sintetizando-se um sinal de fala e comparando-o com o sinal de fala original que se deseja codificar, escolhendo-se o conjunto de parâmetros que gerou o sinal menos distorcido. Ou seja, a análise é feita através da síntese.

Palavras-chave: Análise por síntese; codificação de fala; modelo fonte-filtro; predição linear.

Abstract: This paper presents a bibliographic review of speech signal coding techniques based on an approach know in the signal processing area as analisys-by-synthesis. Analisys-by-synthesis encoders are based on source-filter speech model; source and filter parameters calculation is made by synthesizing a speech signal and comparing it with the original speech signal to be encoded, parameters are chosen

so as to minimize the distortion error. In other words, the analisys is made by synthesis.

Key-words: Analysis by synthesis; speech codification; source-filter model; linear prediction.

Introdução

Historicamente, as tecnologias de codificação de fala têm sido dominadas pelos codificadores baseados em predição linear [4]. Para atingir uma boa qualidade do sinal de fala, a maioria dos codificadores padronizados (codificadores que se tornaram referência) realizam uma aproximação da forma de onda do sinal de fala, e usam predição linear para explorar a redundância presente na forma de onda.

As técnicas utilizadas nos codificares de fala padronizados não são as únicas existentes: há várias outras técnicas de codificação propostas na literatura que não são utilizadas nos codificadores padronizados. Porém, os codificadores padronizados baseados em predição linear se tornaram a técnica de codificação dominante.

A predição linear recebe este nome pois considera que cada amostra do sinal de fala pode ser aproximada a partir de uma combinação linear de amostras passadas. Os pesos dados às amostras passadas nesta combinação são denominados coeficientes de predição linear e definem o chamado filtro de predição linear, cuja ordem é determinada pelo número de amostras passadas utilizadas.

Os codificadores baseados em predição linear podem ser encarados como uma classe das técnicas chamadas de análise por síntese. Estas são baseadas em um modelo fonte-filtro de produção da fala e o decodificador é composto basicamente por um gerador de excitação e por um filtro de síntese.

Para promover uma visão geral acerca das técnicas de codificação baseadas em análise por síntese, este artigo apresenta uma introdução à teoria acústica de produção da fala e aos conceitos básicos da codificação por modelo fonte-filtro. São apresentadas algumas técnicas de codificação baseadas em análise por síntese e técnicas baseadas em modelos senoidais.

1. Teoria Acústica da Fala

Um sinal de fala é produzido a partir da passagem do ar pelo aparelho fonador humano, constituído basicamente pelos pulmões, laringe e trato vocal. A Figura 1 ilustra os principais componentes do aparelho fonador humano.

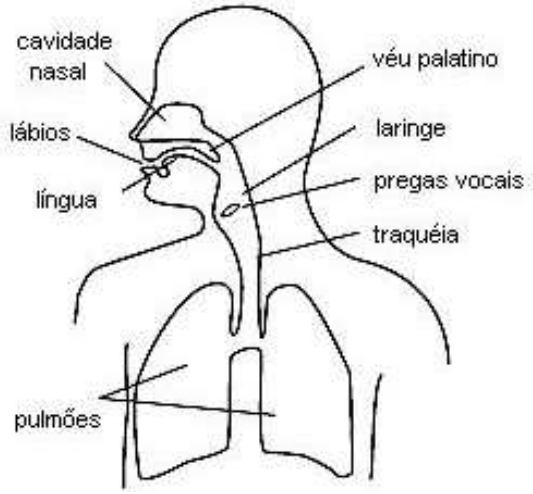

Figura 2: Aparelho fonador humano.

Os pulmões controlam a intensidade do fluxo de ar que passa pela laringe, onde se encontram as pregas vocais. Estas últimas são formadas por dois pares de músculos. Durante a respiração normal, as pregas estão relaxadas e abertas; no processo de produção de voz, ao contrário, as pregas ficam tensas e vibram com a passagem do ar.

A taxa de vibração das pregas vocais é chamada de frequência fundamental (*F0*) do sinal de voz. Nos sons mais agudos, as pregas estão mais contraídas e, portanto, vibram mais depressa; já nos sons mais graves, a taxa de vibração é menor. O comprimento das pregas também influi na taxa de vibração; é por isso que as mulheres, cujas pregas vocais são geralmente mais curtas que as dos homens, possuem um tom de voz normalmente mais agudo. A região do trato vocal onde se encontram as pregas vocais é chamada de glote e, por isso, a onda sonora ali produzida forma os chamados pulsos glotais.

O trato vocal é formado pela faringe, véu palatino, cavidade nasal, palato, língua, dentes e lábios, chamados genericamente de articuladores. Ele funciona como uma *caixa de ressonância,* que atenua ou amplifica certas frequências do pulso produzido na glote. As inúmeras configurações adotadas pelos articuladores

possibilitam a geração dos diversos sons da língua, modificando a onda acústica advinda da laringe.

Deve-se ressaltar que nem todos os sons são formados a partir de pulsos glotais. Nos chamados sons não-vozeados, não ocorre vibração das pregas vocais e o trato vocal recebe da laringe um fluxo de ar turbulento. No caso em que há vibração das pregas vocais, os sons são chamados de vozeados. Há ainda os chamados são mistos, compostos por parte sonora e não sonora.

De forma resumida, as características do sinal de fala em um dado instante dependem da configuração momentânea do aparelho fonador do falante, ou seja, da abertura dos lábios e da mandíbula, da posição da língua, da taxa de vibração das pregas vocais etc. Ao proferir uma sentença, o falante modifica continuamente a configuração de seu aparelho fonador, de forma a produzir uma sequência de sons que transmite uma mensagem ao ouvinte.

2. Conceitos Básicos da Codificação por Modelo Fonte-Filtro

Dadas as colocações da seção anterior, pode-se dizer que o sinal de fala é gerado por uma fonte ou sinal de excitação (fluxo de ar advindo da laringe) que passa por um filtro (trato vocal). Isso serviu de inspiração para a criação de codificadores de fala baseados em um modelo fonte-filtro de produção da fala. A Figura 2 ilustra esse modelo.

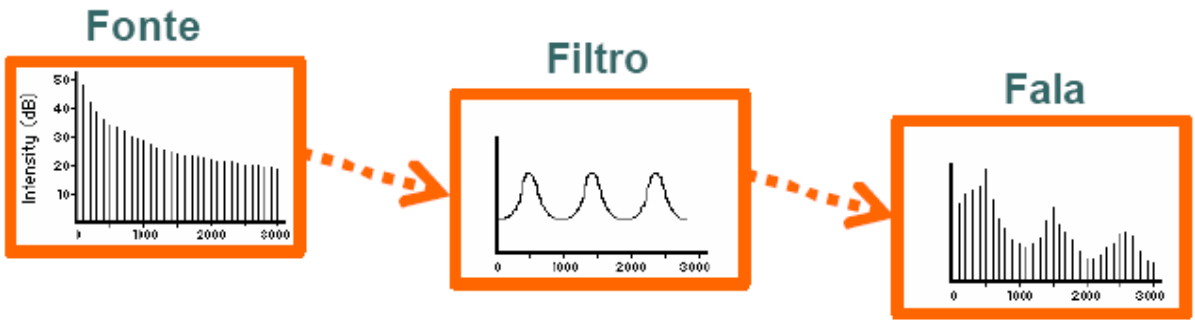

Figura 3: Visão pictórica do modelo fonte-filtro de produção da fala.

No domínio da frequência, o sinal de excitação obtido a partir da vibração das pregas vocais pode ser aproximado por uma sequência de pulsos igualmente espaçados, uma vez que se trata de um sinal periódico. Estes pulsos são chamados

de harmônicas e estão localizados nas frequências múltiplas inteiras de *F0*, como mostra a imagem da esquerda da Figura 2.

A conformação instantânea do trato vocal define um filtro, que pode ser representado em termos de sua resposta em frequência, como mostra a imagem central da Figura 2. As frequências que são amplificadas por esse filtro são chamadas de formantes (frequências de ressonância do trato vocal), correspondendo a frequência central e largura de banda dos picos mostrados na curva da figura.

Por fim, o resultado obtido é o sinal de fala, cujo contorno é chamado de envoltória espectral, e que pode ser caracterizado simplificadamente pelo valor de *F0* e pelos formantes. As Figuras 3 e 4 ilustram a diferença entre esses conceitos.

Na Figura 3, é ilustrada a diferença entre os espectros de um mesmo som da língua (mesmo envoltória espectral), por exemplo a vogal “a”, com diferentes valores de *F0*.

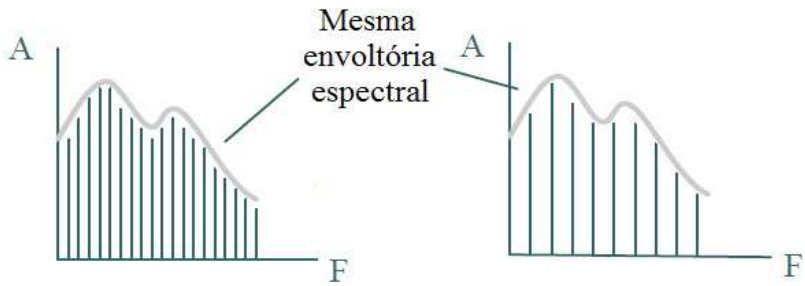

Figura 4: Espectros de um mesmo som da língua com diferentes valores de F0. O primeiro é mais agudo (F0 maior) e o segundo mais grave (F0 menor).

Já na Figura 4, é ilustrada a diferença entre sons diversos da língua (envoltória espectral diferente), mas com o mesmo *F0* (mesmas frequências harmônicas). A figura da esquerda representa uma vogal (e.g. a vogal “a”), enquanto a figura da direita representa uma vogal diferente (e.g. a vogal “i”).

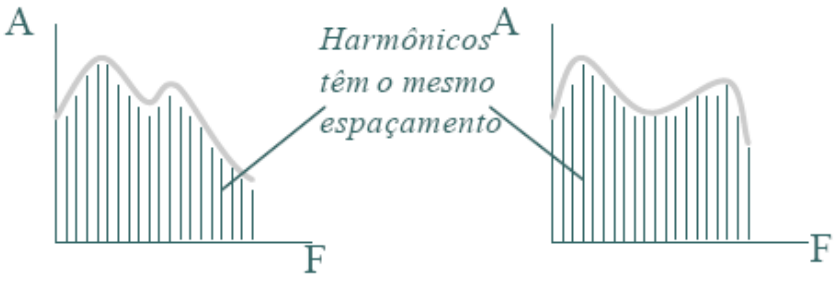

Figura 5: Espectros de sons diversos da língua com o mesmo F0.

Nos codificadores baseados no modelo fonte-filtro, primeiramente o sinal de fala é dividido em trechos de curta duração, geralmente entre 10 ms e 20 ms, denominados *quadros*. Nesse intervalo de tempo, considera-se que o sinal de excitação e a configuração do trato vocal não variam. Cada quadro então é analisado isoladamente, extraindo-se do sinal a informação da fonte e do filtro. A forma como essa separação é feita depende do modelo adotado, sendo comum o emprego de um modelo de predição linear.

Dessa forma, não é necessário transmitir ou armazenar a forma de onda do sinal de fala. Ao invés disso, são necessários apenas um índice para o sinal de excitação (fonte), e os parâmetros requeridos para gerar o que passaremos a chamar de filtro de síntese.

3. Técnicas de Codificação Baseadas em Modelo Fonte-Filtro

Os codificadores de análise por síntese são baseados no modelo fonte-filtro de produção de fala. Essas técnicas receberam esse nome porque o cálculo dos parâmetros da fonte e do filtro é feito sintetizando-se um sinal de fala e comparando-o com o sinal de fala original que se deseja codificar, escolhendo-se o conjunto de parâmetros que gerou o sinal menos distorcido. Ou seja, a análise é feita através da síntese. A Figura 5 apresenta uma ilustração simplificada de um codificador baseado em análise por síntese.

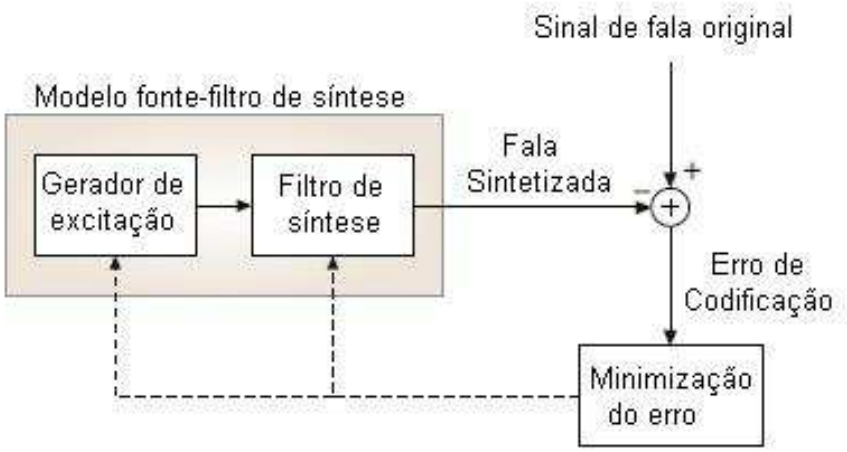

Figura 6: Codificador baseado em análise por síntese.

O sinal de fala é analisado quadro a quadro. Por isso, o sinal de excitação é quantizado também quadro a quadro, obtendo-se blocos de amostras que formam um vetor de tamanho fixo *K*. Se cada amostra for representada por *r* bits, cada vetor de excitação é composto por *Kr* bits. Assim, o sinal de excitação de cada quadro pode ser quantizado em um dos 2^{Kr} vetores possíveis.

Na prática, o filtro de síntese é derivado diretamente do sinal de entrada e diversos vetores de excitação candidatos são submetidos ao filtro de síntese, conforme um algoritmo de otimização. O sinal de fala sintetizado resultante para cada vetor de excitação é comparado com o sinal de entrada e aquele que minimizar determinado critério de distorção tem seu vetor de excitação selecionado para ser transmitido ou armazenado.

O decodificador é composto basicamente pelo gerador de excitação e pelo filtro de síntese (exatamente como no codificador) que são os componentes destacados na Figura 5. Ao receber os parâmetros do filtro de síntese e o sinal de excitação, o sinal de fala sintetizado é gerado e, considerando ausência de erros de transmissão, esse sinal é idêntico ao que foi avaliado no codificador e, portanto, seu erro de codificação (sua distorção) é conhecido.

3.1 Técnicas baseadas em predição linear

Predição linear (LPC – *Linear Prediction Coding*) é uma ferramenta utilizada em diversas áreas, tais como filtragem adaptativa, economia e geofísica. Em processamento de fala, a predição linear é talvez a forma mais comum de análise do sinal e desempenha um papel fundamental em diversas aplicações (reconhecimento de fala, reconhecimento de locutor, compressão, modelagem etc.) [6]. Isso por que o sinal de fala pode ser muito bem modelado por meio da predição linear.

A predição linear recebe esse nome por considerar que cada amostra do sinal de fala pode ser aproximada (*predita*) a partir de uma combinação *linear* de amostras passadas. Os pesos dados às amostras passadas nesta combinação são denominados coeficientes de predição linear e definem o chamado filtro de predição linear, cuja ordem é determinada pelo número de amostras passadas utilizadas [5].

Retomando os conceitos da seção anterior, técnicas de codificação baseadas em predição linear funcionam da seguinte forma: inicialmente faz-se a análise de

predição linear [4][10][11] do sinal que se deseja codificar, obtendo-se os coeficientes do filtro de predição. Com esse filtro e o sinal de entrada, pode-se obter o sinal estimado pelo preditor e, assim, calcular o erro de predição (diferença entre o sinal original e o sinal predito).

Esse sinal de erro é também chamado de resíduo. Uma propriedade dos filtros de predição linear é que se o sinal de resíduo for usado como sinal de excitação do filtro, o sinal obtido é indistinguível do sinal original. Ou seja, com a predição linear obtém-se tanto o sinal de excitação quanto o filtro necessário para a síntese do sinal de fala.

O desafio dos codificadores LPC está em implementar formas eficientes para transmissão ou armazenamento dos coeficientes do filtro e, principalmente, do sinal de excitação. A seguir, são apresentadas algumas das técnicas de codificação mais importantes baseadas nesse paradigma.

3.1.1 Code Excited Linear Prediction (CELP)

A partir de sua criação na década de 80, o codificador CELP [7] e suas variações se tornaram a técnica dominante dos padrões de codificação de fala. Isso fica evidente pela família de codificadores que se criou a partir dele, mostrada na Figura 6 [4]. Neste relatório é apresentado apenas o CELP original.

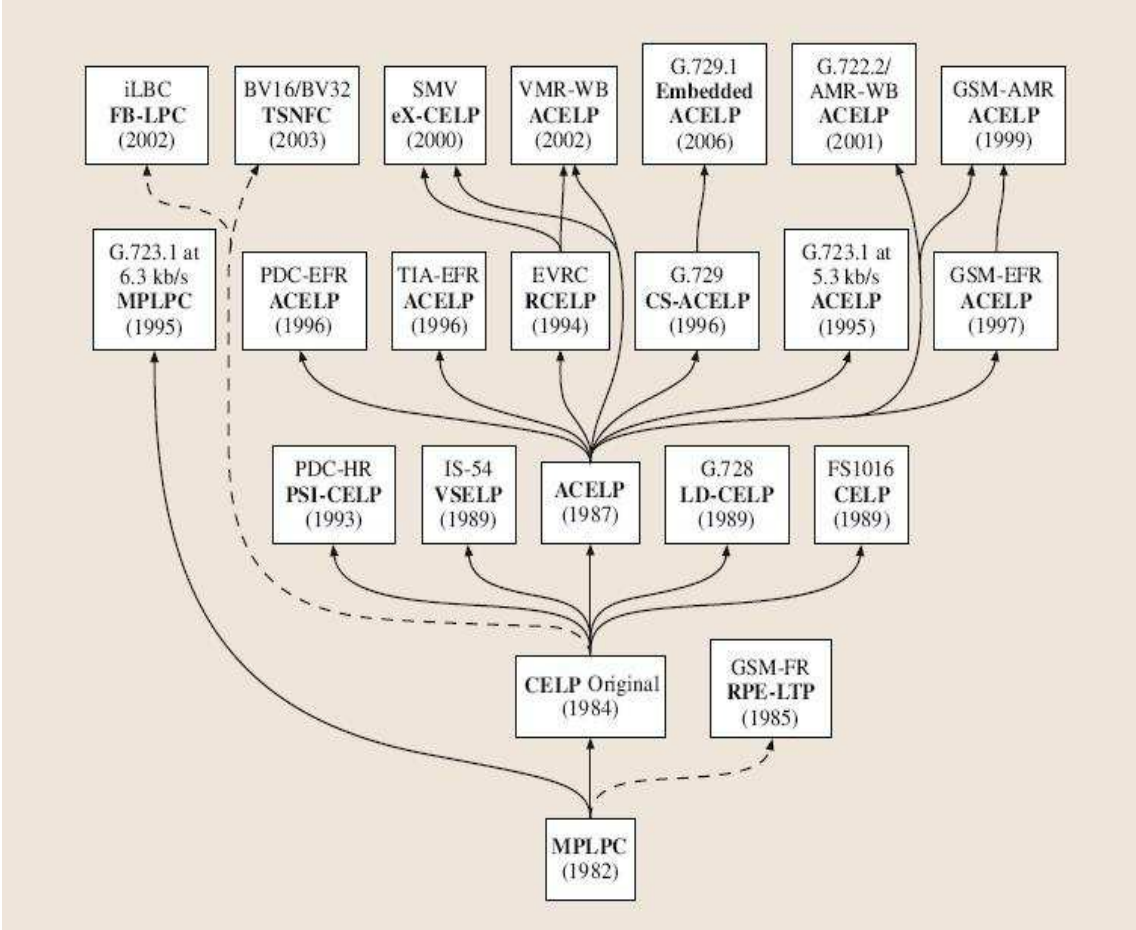

Figura 7: Família de codificadores CELP [4].

A principal contribuição desse codificador é o emprego de um dicionário (*codebook*) para quantização vetorial (VQ – *Vector Quantization*) do sinal de excitação, permitindo sua codificação com poucos bits. No codificador CELP, o sinal de fala é analisado em blocos de 40 amostras. O dicionário empregado contém 1024 vetores de 40 amostras cada e, portanto, o sinal de excitação pode ser codificado com apenas 10 bits.

O filtro de síntese é constituído, na verdade, por dois filtros distintos, um de predição de longo prazo (*long-term*) e outro de predição de curto prazo (*short-term*). Além disso, o resíduo da predição passa por um filtro de ponderação perceptual [4]. A estrutura do codificador CELP original está mostrada na Figura 7.

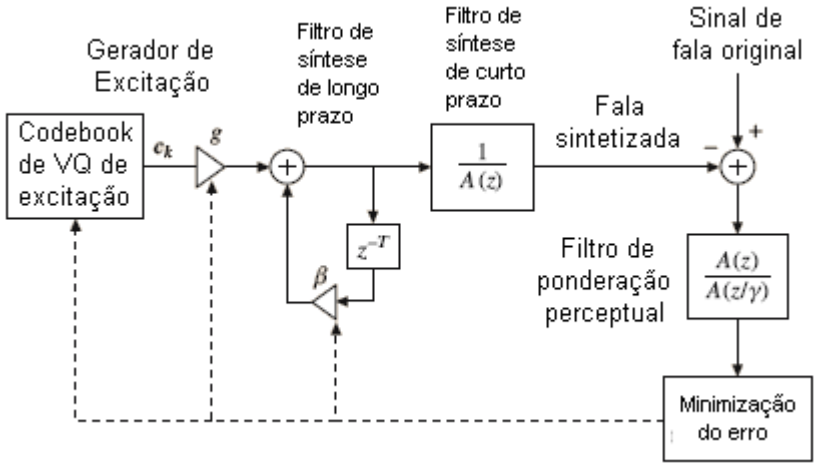

Figura 8: Estrutura do codificador CELP original.

No entanto, o codificador apresentado na Figura 7 requer uma complexidade computacional muito elevada para a busca no *codebook*. Por isso, nos anos subsequentes à publicação do CELP, foram propostos diversos codificadores estruturados especialmente para permitir uma busca rápida.

Como não cabe aqui entrar em detalhes, genericamente esses codificadores dividem o *codebook* em um *codebook* fixo e um *codebook* adaptativo e re-arranjaram a estrutura dos filtros, de forma a obter um codificador matematicamente equivalente, mas computacionalmente mais eficiente.

3.1.2 Mixed Excitation Linear Prediction (MELP)

Conforme visto na seção 1, o sinal de fala pode ser vozeado ou não vozeado. Nos trechos vozeados, o sinal de excitação pode ser bem representado por uma sequência de pulsos, ou um trem de impulsos, com frequência igual a *F0*. Nos trechos não-vozeados, o sinal de excitação pode ser modelado simplesmente como um ruído. A Figura 8 ilustra um modelo simplificado de síntese LPC.

Do sinal de entrada são extraídos, portanto: (i) se o sinal é vozeado ou não e, em caso afirmativo, a frequência fundamental (*F0*), (ii) os coeficientes do filtro LPC e (iii) um ganho para regular a amplitude do sinal sintetizado. Esses parâmetros são otimizados para obter o sinal sintetizado mais próximo do sinal de entrada. O decodificador executa as operações (i), (ii) e (iii), mostradas na Figura 8, obtendo assim o sinal desejado.

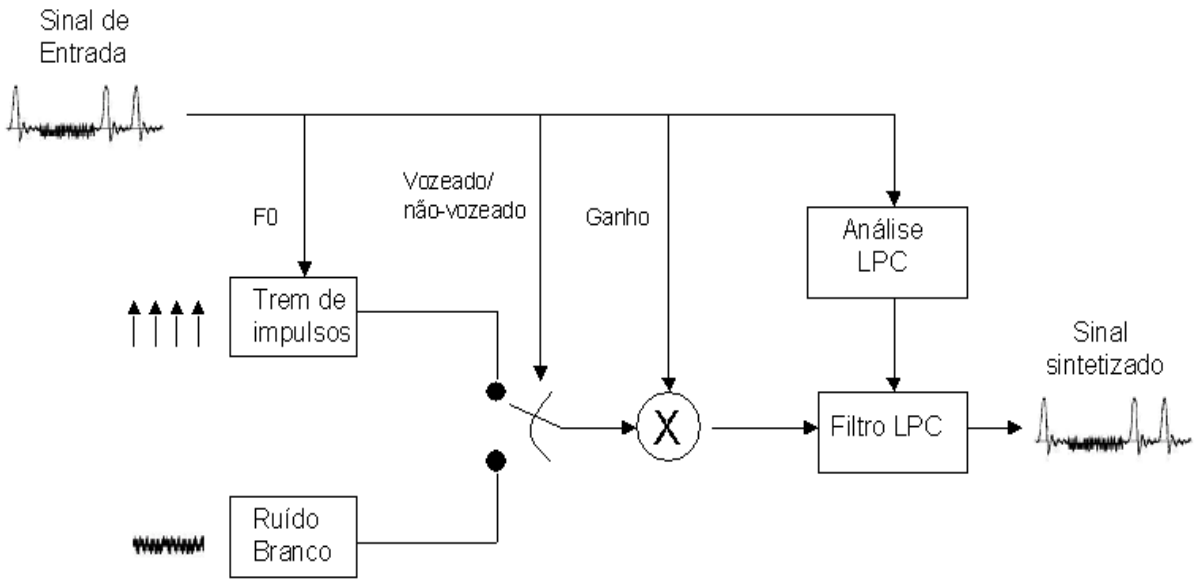

Figura 9: Modelo simplificado de síntese LPC.

Esse modelo apresenta algumas limitações. A maior delas vem do fato de que o sinal de fala nunca é puramente vozeado ou não vozeado. Por isso, a excitação deveria ser constituída sempre de uma "combinação" de pulsos e ruído. Basicamente, essa é a ideia por trás codificador MELP [8].

No entanto, a codificação MELP não se restringe a isso. Antes de serem combinados, o trem de impulsos e o ruído branco são submetidos a filtros passa-banda. Para formar o sinal de excitação do filtro de síntese LPC, o sinal resultante é submetido ainda a uma etapa de aprimoramento espectral (para realizar uma conformação espectral). O ganho só é incorporado ao sinal após o filtro de síntese. Por fim, o sinal é submetido a mais um filtro para produzir o sinal de fala sintetizado.

4. Técnicas Baseadas em Modelos Senoidais

As técnicas baseadas em modelos senoidais assumem que a forma de onda de um sinal qualquer é composta pela soma de um determinado número de senoides com amplitude, frequência e fase variantes no tempo. Os modelos senoidais foram propostos como uma alternativa ao uso da predição linear e são capazes de produzir fala sintetizada perceptualmente indistinguível do sinal original [4].

4.1 ABS/OLA

O modelo ABS/OLA incorpora a técnica de análise por síntese (Analysis-By-Synthesis) para estimar os parâmetros do modelo senoidal e um processo de síntese que utiliza uma técnica conhecida como O*verLap-Add*, surgindo o nome ABS/OLA [9]. Neste modelo, o sinal de fala é aproximado por uma função que depende de uma modulação da envoltória do sinal, introduzida para levar em consideração as variações da energia silábica, depende de um termo que corresponde a uma janela de síntese complementar condicionada a uma restrição, e de um termo correspondente a componentes senoidais, com amplitude, frequência e fase variáveis. Dadas as frequências senoidais, é possível estimar os parâmetros de amplitude e fase ótimos que minimizam o erro quadrático médio entre o sinal analisado e o sinal obtido através do modelo, através de uma aproximação recursiva. As frequências senoidais ótimas são obtidas através de busca exaustiva sobre o conjunto de frequências candidatas uniformemente distribuídas. A síntese é obtida utilizando a técnica de *overlap-add* [9].

4.2 HNM

A técnica de codificação de fala conhecida como HNM (*harmonic-plus-noise model*) é largamente utilizada em sistemas de conversão texto-fala baseados em síntese concatenativa, na etapa de síntese do sinal de fala, devido à sua capacidade de suavizar descontinuidades e de comprimir o sinal.

Esse modelo considera que o sinal de fala é composto pela adição de dois componentes: um componente harmônico (banda inferior) e um componente de ruído (banda superior), o que corresponde a uma separação do espectro da fala em duas bandas.

O limite entre essas bandas é determinado por um limiar de frequência máxima para trechos vozeados. Na banda inferior, o sinal é representado como uma soma de ondas senoidais, cujas amplitudes e frequências variam lentamente. Os valores das amplitudes, das fases e das frequências dessas componentes senoidais são estimadas a partir do sinal de fala original [12].

A banda superior é modelada por um ruído branco gaussiano, que é submetido a um filtro normalizado apenas com polos e é multiplicada por uma função de envoltória [4].

Essa análise é feita para cada quadro do sinal (a uma taxa constante de quadros por segundo), permitindo assim uma representação mais compacta do sinal e a suavização de eventuais descontinuidades que surgiram na concatenação (OLA – *overlap-add*). A Figura 9 ilustra a síntese baseada no modelo HNM.

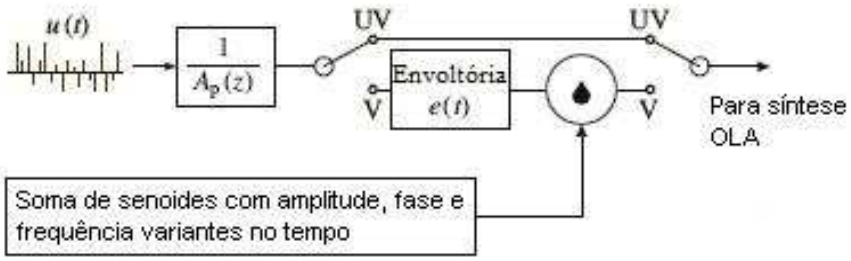

Figura 10: Síntese baseada no modelo HNM.

Considerações Finais

Nesse documento foi apresentada uma visão geral das técnicas de codificação de sinais de fala baseadas em análise por síntese. Esse tipo de codificação baseia-se em um modelo fonte-filtro de produção de fala, e a extração dos parâmetros da fonte de excitação e do filtro do trato vocal é feita sintetizando-se um sinal de fala e comparando-o com o sinal de fala original que se deseja codificar.

Foram apresentadas uma introdução à teoria acústica de produção da fala e aos conceitos básicos da codificação por modelo fonte-filtro, bem como algumas técnicas de codificação baseadas em análise por síntese e técnicas baseadas em modelos senoidais.

Esse documento não se aprofundou em detalhes matemáticos das técnicas apresentadas; desta forma, seu objetivo é fornecer uma visão geral sobre o tema de codificação de fala, levantando alguns pontos de interesse específicos para a

técnica de análise por síntese. Portanto, esse texto deve ser usado como material de consulta inicial e não como um guia para implementação das técnicas aqui citadas.

Bibliografia

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

[15] F. O. Simões. **Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil**. Dissertação de Mestrado, UNICAMP, Maio 1999.

[16] Notas de aula da disciplina F105 - **Física da Fala e da Audição**. Prof. Dr. Marcelo Knobel, IFGW, Unicamp, 2004.

[17] J. Benesty, M. M. Sondhi e Y. Huang, editors. **Springer Handbook of Speech Processing**. Springer, 2008.

[18] J. D. Markel e A. H. Gray Jr. **Linear Prediction of Speech**. Springer, 1976.

[19] B. S. Atal. **The History of Linear Prediction**. In IEEE Signal Processing Magazine, vol. 23, pp. 154–161, March 2006.

[20] M. R. Schoroeder e B. S. Atal. **Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates**. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 937–940, March 1985.

[21] A. McCree e T.P. Barnwell III. **A mixed excitation LPC vocoder model for low bit rate speech coding**. IEEE Transactions on Speech and Audio Processing, vol. 3(4), pp. 242–250, 1995.

[22] R. Crochiere. **A weighted overlap-add method of short-time Fourier analysis/synthesis**. IEEE Trans. Acoust. Speech 28(1), pp. 99–102, 1980.

[23] J. Makhoul. **Linear prediction: A tutorial review**. Proceedings of the IEEE, 63 (5):561–580, abril de 1975.

[24] M. H. Hayes. **Statistical Digital Signal Processing and Modeling**. J. Wiley & Sons, Inc., Nova York, 1996.

[25] Y. Stylianou. **Applying the harmonic plus noise model in concatenative speech synthesis**. IEEE Transactions on Speech and Audio Processing, vol. 9(1), pp. 21–29, 2001.