Extensão Artificial de Largura de Banda para Sinais de Fala Usando Classificação Fonética

Ênio dos Santos Silva e Rui Seara

***Resumo*— Este artigo apresenta uma nova estratégia para implementar sistemas de extensão artificial de largura de banda (*artificial band width extension* - ABWE) para sinais de fala aplicados à telefonia em redes de pacotes. Na literatura, diversas abordagens são propostas, entretanto, ainda não há um procedimento adequado que consiga representar satisfatoriamente segmentos de fala com energia concentrada em altas frequências. Visando melhorar o desempenho desses sistemas, uma nova estratégia baseada em classificação fonética é aqui discutida. Os resultados obtidos são avaliados e o realce na qualidade dos sinais de fala reconstruídos confirmam a eficácia da estratégia proposta.**

***Palavras-Chave*— Classificação fonética, codificação de fala, extensão artificial de largura de banda, realce de voz.**

***Abstract*— This paper presents a new strategy for implementing artificial band width extension (ABWE) systems of speech signals applied to telephony over packet networks. In the open literature, several approaches are presented for theses systems, however, there is still no procedure that can satisfactorily represent speech segments with concentrated energy at high frequencies. Aiming to improve the performance of these systems, a new strategy based on phonetic classification is discussed here. The obtained results show an enhancement in the quality of reconstructed speech signals, confirming the effectiveness of the proposed strategy.**

***Keywords*— Phonetic classification, speech coder, artificial band width extension, speech enhancement.**

I. Introdução

A extensão artificial de largura de banda (*artificial band width extension* - ABWE) é uma técnica que sintetiza componentes de alta frequência em um sinal de fala digital. Nos últimos anos, essa técnica ganhou evidência por estar contida em um dos mais difundidos *codecs* de telefonia: o *codec* G.729 [1], em sua Recomendação ITU-T G.729.1 [2]. A necessidade de extensão de largura de banda surgiu devido à baixa qualidade do sinal de fala na rede de telefonia pública (*public switched telephone network* - PSTN) que, historicamente, por razões econômicas e para evitar interferências entre canais (*cross-talk*), adotou, como padrão para transmissão, sinais de banda estreita (*narrowband* - NB) [3]. Nesse cenário, os sinais são amostrados a 8000 Hz e contêm componentes de frequências limitados entre 300 e 3400 Hz. Tal limitação de largura de banda provoca perda de qualidade nos sinais de fala, tornando-os "abafados", "sem brilho" e com degradação de naturalidade e inteligibilidade [4].

Em [2], [5] e [6], a utilização de codificadores de banda larga (*wideband* - WB) vem sendo discutida e adotada em alguns sistemas de comunicações. Em um futuro próximo, é inevitável a extinção das PSTNs, como já vem ocorrendo por meio da adoção de sistemas VoIPs (voz sobre IP) para chamadas de longa distância (DDD/DDI) [3]; no entanto, a PSTN ainda é uma das redes mais difundidas em todo o mundo e sua modernização para WB demandaria um enorme esforço, o qual seria, em curto prazo, economicamente inviável [3]. Portanto, durante os próximos anos, preveem-se apenas migrações graduais para terminais de WB e, por um certo período de transição, redes de telefonia mistas NB e WB irão coexistir [4]. Assim, para contornar as limitações da comunicação em NB, a extensão artificial de largura de banda pode ser vista como uma alternativa interessante. Nesse contexto, os componentes de frequência não transmitidos pelos codificadores NB devem ser sintetizados artificialmente através da estimação de parâmetros do modelo fonte-filtro de produção do sinal da fala, no qual são consideradas duas etapas: de estimação do sinal de excitação e de estimação do envelope do trato vocal [1], [3], [4]. Essa técnica proporciona melhorias na qualidade do sinal de fala, tornando-o mais próximo de um sinal WB, requerendo apenas alterações nos terminais receptores (*far-end*), fazendo o sistema compatível com a maior parte das redes de telefonia existentes [3].

Nos últimos anos, diversos sistemas usando diferentes estratégias vêm sendo propostos para a realização de ABWE [6]. Em [2], [5] e [7], a ABWE é obtida através da transmissão de informações extras ao *far-end* (*side information*). Entretanto, a implementação de tais estratégias resulta em aumento na taxa de bits do sinal, bem como alterações nos terminais transmissores (*near-end*) e receptores (*far-end*) da rede de comunicação. Para contornar tais problemas, [8] e [9] sugerem estratégias que não necessitam de *side information*. No entanto, tais estratégias levam a um alto índice de ocorrência de ruídos musicais e impulsivos devido à falta de um tratamento adequado na escolha e clusterização de parâmetros discriminativos do trato vocal.

Apesar da crescente evolução dos sistemas de ABWE, ainda não se dispõe de qualquer procedimento consolidado apresentando desempenho satisfatório na estimação de WB, principalmente nos casos em que o sinal de fala contém energias concentradas em altas frequências, i.e., maiores do que 5000 Hz. Nesses casos, nota-se geralmente artefatos indesejados no sinal de fala reconstituído [10], [11], [12] e [13].

Neste trabalho, visando compatibilidade com *codecs* NB e WB, que são amplamente adotados no mercado de telefonia, tal como Recomendação ITU-T G.729 [1] e G.729.1 [2], a estratégia aqui desenvolvida para ABWE é baseada nos procedimentos descritos em [6]. Para propiciar independência de *side information*, bem como redução de ocorrências e atenuação dos efeitos indesejáveis no sinal de fala reconstruído, são propostas alterações nas etapas de representação e estimação do trato vocal do sistema proposto em [2], para o qual sugerimos acrescentar uma etapa de classificação fonética. Essa etapa é responsável por um tratamento específico em diferentes classes de sinais de fala, garantindo, dessa forma, uma melhor representação e, consequentemente, uma melhoria

Ênio dos Santos e Rui Seara, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil, e-mails: enio@linse.ufsc.br; seara@linse.ufsc.br. Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

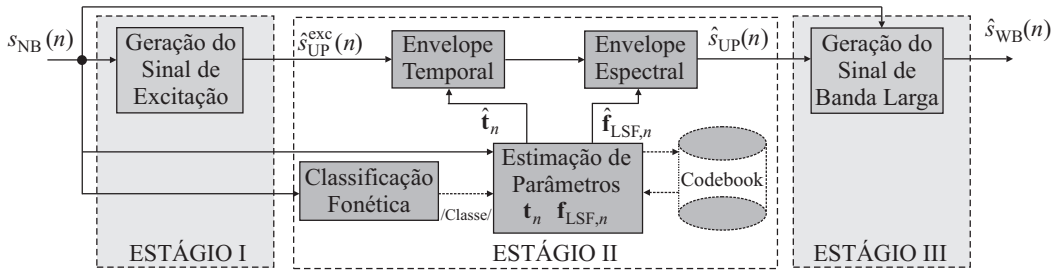

Fig. 1. Diagrama de blocos do sistema ABWE.

na clusterização dos parâmetros discriminativos do trato vocal, resultando em uma síntese mais "limpa" e agradável dos sinais WBs.

II. Algoritmo de ABWE

A idéia básica de um sistema de ABWE é recriar artificialmente componentes de alta frequência, convertendo um sinal NB em um sinal WB, isto é, restaurando as propriedades acústicas pertinentes à WB. Essa técnica, que aparentemente conflita o critério de Nyquist, assume que ambos sinais, NB e WB, sejam gerados pelo mesmo modelo fonte-filtro de produção da fala [4].

Sendo o sinal de fala quase-estacionário a cada segmento de curta duração (aproximadamente 20 ms) [14], parâmetros WB do modelo fonte-filtro, tais como o sinal de excitação e o envelope do trato vocal, são estimados através de informações implícitas contidas no sinal NB [3]. Dessa forma, modelos matemáticos do processo de produção do sinal de fala podem ser obtidos em intervalos de tempo periódicos e, assim, torna-se possível estimar os componentes de alta frequência não presentes no sinal transmitido através da PSTN.

A solução ideal, representada por (1), para um sistema de ABWE, é a obtenção de uma função f, tal que f determine os parâmetros de banda alta UP, $y \in S_{\mathrm{UP}}$, a partir de parâmetros NB, $x \in S_{\mathrm{NB}}$, onde S_{UP} e S_{NB} representam, respectivamente, os espaços de dados dos parâmetros UP e NB. Dessa forma, com a combinação de ambos espaços, S_{UP} e S_{NB}, é possível uma estimação dos parâmetros de banda larga $S_{\mathrm{WB}} \supset (S_{\mathrm{NB}} \cup S_{\mathrm{UP}})$. Assim,

$$\begin{array}{rcl} f : S_{\mathrm{NB}} & \rightarrow & S_{\mathrm{UP}} \\ x & \mapsto & y = f(x) \end{array} \quad (1)$$

Neste artigo, adota-se como função f, para estimação do espaço de dados S_{UP}, uma função f_{CB} de mapeamento por *codebook* [15].

A. Proposta de Nova Estratégia para Extensão de Banda

O procedimento de ABWE adotado aqui é baseado no processo de codificação e decodificação de sinais WB propostos em [2]. Entretanto, a estratégia implementada em [2] lança mão de *side information*, tais como fase e energia dos segmentos do sinal de fala para estimar os sinais de excitação e os envelopes temporais e espectrais correspondentes à banda alta do sinal. Para tornar a codificação independente de *side information* e para melhorar o desempenho do decodificador em cenários de transmissões NB, este trabalho propõe uma nova estratégia de processamento, incluindo uma etapa de classificação fonética e adotando coeficientes LSF (*line spectrum frequency*) [15], ao invés de coeficientes de fase e energia, para a representação do envelope espectral.

A Fig. 1 ilustra o diagrama de blocos do sistema de ABWE modificado. Assim como em sua versão original [2], esse diagrama tem como base uma estrutura de três estágios de processamento, descritos como segue:

1) Estágio I. Estimação do sinal de excitação através de predição linear (*code-excited linear prediction* - CELP) [1].
2) Estágio II. Filtragem do sinal de excitação resultante do primeiro estágio através dos envelopes temporal e espectral do trato-vocal estimados a partir de uma consulta a um conjunto de *codebooks* baseados em classificação fonética.
3) Estágio III. Cálculo de ganho e pós-processamento para a estimação do sinal WB.

Os estágios que compõem o sistema ABWE serão descritos com detalhes nas seções seguintes.

III. Estágio I - Estimação do Sinal de Excitação

A cada quadro (de aproximadamente 20 ms) do sinal de fala NB $s_{\mathrm{NB}}(n)$, recebido como entrada no sistema de ABWE da Fig. 1, são estimados sinais de excitação de banda alta $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n)$. A estimação de $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n)$ é obtida através de um modelo ideal de produção de fala quase-estacionário, no qual devem ser satisfeitos os seguintes critérios [6]:

- O sinal de excitação deve apresentar espectro plano.
- Para sinais de fala vozeados, a excitação deve conter harmônicos da frequência fundamental do sinal, F_0.
- Para sinais de fala não-vozeados, a excitação deve ser um ruído branco.
- Sinais de fala mistos (vozeados e não-vozeados) devem apresentar uma razão sinal-ruído (SNR) variável.
- A contribuição de parâmetros vozeados não deve ser dominante na banda de alta frequência.

Em [1] e [2], os critérios descritos anteriormente são satisfeitos e a mesma estratégia de estimação do sinal de excitação é também aqui adotada. A Fig. 2 ilustra o diagrama de blocos do processo de estimação do sinal de excitação, no qual parâmetros de energia, ganho e *pitch* são obtidos através de procedimentos utilizados no CELP e aplicados para estimar o sinal de excitação $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n)$, cuja composição é uma mistura de sinais não-vozeado $\hat{s}^{\mathrm{exc,nv}}_{\mathrm{UP}}(n)$, gerados a partir de um ruído branco, com sinais vozeados $\hat{s}^{\mathrm{exc,v}}_{\mathrm{UP}}(n)$, gerados a partir de um trem de pulsos periódicos com frequência fundamental F_0 [2].

IV. Estágio II - Geração de Envelopes Temporais e Espectrais

Neste estágio do sistema de ABWE, são estimados os envelopes temporais e espectrais de banda alta que caracterizam

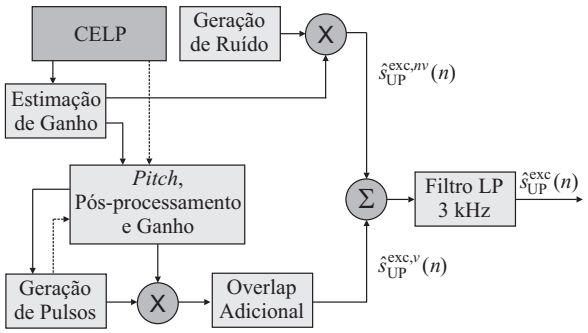

Fig. 2. Diagrama de blocos para geração do sinal de excitação.

TABELA I
DISTRIBUIÇÃO DE CLASSES FONÉTICAS PARA SINAIS DE FALA

Classes				ex.	Descrição
2	3	5	9		
A1	B1	C1	D1	/z/	Fricativas vozeadas alveolares
			D2	/v/	Fricativas vozeadas labiodentais
			D3	/j/	Fricativas vozeadas palatais
		C2	D4	/V/	Demais fonemas vozeados
	B2	B3	D5	/f/	Fricativas não-vozeadas labiodentais
			D6	/s/	Fricativas não-vozeadas alveolares
			D7	/x/	Fricativas não-vozeadas palatais
		C4	D8	/U/	Demais fonemas não-vozeados
A2	B3	C5	D9	/Sil/	Silêncio

o trato vocal no processo de geração da fala. Esses envelopes são representados pelos seguintes vetores de parâmetros

$$\mathbf{t}_n = [t_n(1), \ldots, t_n(16)]^{\rm T} \tag{2}$$

e

$$\mathbf{f}_{{\rm LSF},n} = [f_{{\rm LSF},n}(1), \ldots, f_{{\rm LSF},n}(19)]^{\rm T} \tag{3}$$

onde o vetor $\mathbf{t}_n$ representa o envelope temporal contendo energias logarítmicas de 16 subquadros ($1,25$ ms cada) [6] e o vetor $\mathbf{f}_{{\rm LSF},n}$, componentes LSFs que caracterizam o envelope espectral. Como ilustrado na Fig. 1, tais parâmetros são estimados através de consulta a um conjunto de *codebooks* selecionados de acordo com suas classes fonéticas. Esse processo de classificação reagrupa os vetores de parâmetros $\mathbf{t}_n$ e $\mathbf{f}_{{\rm LSF},n}$ em classes fonéticas, permitindo, dessa forma, uma estimação específica para cada classe.

A. Clusterização e Classificação Fonética

O conjunto de sons componentes dos sinais de fala podem ser agrupados, de acordo com suas similaridades acústicas, em classes fonéticas. Tais classes representam um conjunto de características temporais e espectrais singulares. Essas singularidades são específicas de cada conjunto de sinais e garantem maior discriminação entre as diferentes classes fonéticas. Em [10] e [16], são descritas as propriedades acústicas dos sinais de fala. De acordo com a avaliação do comportamento de cada classe fonética quanto à distribuição de energia no domínio espectral, nota-se que, para o conjunto de classes fricativas, os componentes mais discriminativos encontram-se nas altas bandas de frequência, $S_{\rm UP}$, i.e., acima de 3400 Hz e, consequentemente, além da banda de frequência originalmente considerada pela PSTN. Dessa forma, dentre as demais classes fonéticas (veja [17]), a discriminação entre fonemas das classes fricativas torna sua percepção mais difícil para os usuários de PSTNs convencionais. Assim, uma atenção especial é dada para tais classes de fonemas, resultando na proposta da Tabela I, a qual será utilizada como referência para o processo de clusterização dos sinais de fala $s_{\rm NB}(n)$ e $s_{\rm WB}(n)$.

A Tabela I apresenta quatro diferentes tipos de clusterização denominadas A, B, C e D, contendo duas, três, cinco e nove classes. Após o processo de clusterização e classificação fonética, a seleção de *codebooks* NB e WB torna-se mais discriminativa e, consequentemente, mais apropriada para a etapa de treinamento.

B. Extensão Usando Mapeamento via Codebooks

A Fig. 3 mostra o diagrama de blocos da etapa de treinamento de *codebooks*, em que os filtros LP e HP representam filtros passa-baixas e passa-altas, respectivamente. Nessa etapa, o sinal de banda larga $s_{\rm WB}(n)$ é filtrado pelos correspondentes filtros, resultando nos sinais de banda estreita $s_{\rm NB}(n)$ e banda alta $s_{\rm UP}(n)$. A partir desses sinais, são extraídos parâmetros que representam os espaços de dados $x_{Classe}^{\rm NB}$ e $y_{Classe}^{\rm UP}$, de acordo com suas correspondentes classes fonéticas.

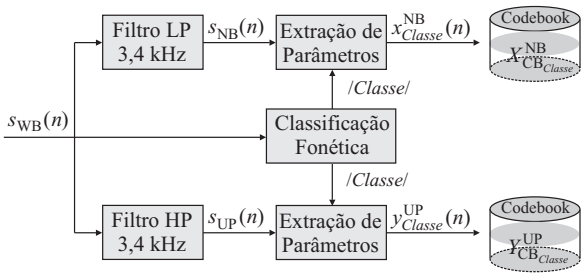

Fig. 3. Etapa do processo de treinamento de *codebooks*.

O processo de treinamento dos melhores *codebooks* para representação dos envelopes do trato vocal inclui o bloco de classificação fonética proposto, visando gerar os diferentes *codebooks* para as classes fonéticas definidas na Tabela I. Em um primeiro momento, esse procedimento é governado por um processo de aprendizagem supervisionada [8], em que a classe fonética seja consultada e os respectivos espaços de dados $x_{Classe}^{\rm NB}$ e $y_{Classe}^{\rm UP}$ sejam gerados de acordo com suas classes correspondentes. Após a discriminação em classes, o algoritmo LGB [15] é utilizado para agrupar os vetores de parâmetros em torno de diferentes centróides e gerar os *codebooks*, $X_{{\rm CB}_{Classe}}^{\rm NB}$ e $Y_{{\rm CB}_{Classe}}^{\rm UP}$, de acordo com as suas correspondentes distâncias euclidianas, $d(S,\hat{S})$. Assim,

$$d(S,\hat{S}) = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|S(e^{j\Omega}) - \hat{S}(e^{j\Omega})\right|^2 d\Omega\right]^{1/2} \tag{4}$$

onde $S(e^{j\Omega})$ caracteriza o espaço de dados das centróides e $\hat{S}(e^{j\Omega})$, o espaço de dados do sinal recebido e analisado. Dessa forma, os *codebooks* gerados possuem maior discriminação entre suas diferentes centróides.

O processo de treinamento aqui utilizado considera as técnicas discutidas em [8] e [15], na qual *codebooks* duais de banda estreita $X_{\rm CB}^{\rm NB}$ e de banda alta $Y_{\rm CB}^{\rm UP}$ são construídos. A construção desses *codebooks* duais se dá através do mapeamento um-para-um de índices equivalentes a *codewords* de banda alta para *codewords* de banda estreita. Assim, índices de mapeamento das *codewords* de banda alta são usados para o mapeamento de seus correspondentes coeficientes em banda estreita [8]. A utilização de índices de banda alta é adotada para

evitar confusabilidade e perda de informações de singularidades que resultassem em imprecisão dos envelopes do trato vocal, principalmente nos casos em que os componentes espectrais sejam predominantemente de altas frequências. Entretanto, na etapa de teste indicada na Fig. 1, durante o processo de estimação de parâmetros, uma imprecisão de singularidades acústicas pode ocorrer, por exemplo, devido a eventuais erros na classificação fonética e escolha do *codebook* mais apropriado para um determinado quadro do sinal de fala. Todavia, tal problema é atenuado devido à robustez da estratégia proposta com respeito a erros de classificação fonética, em que, mesmo ocorrendo erros de classificação, as versões estimadas dos parâmetros de banda alta não estariam tão distantes da versão correta, dado o agrupamento de classes similares.

C. Estimação de Parâmetros e Pós-processamento

Para a estimação dos parâmetros do vetor $\tilde{\mathbf{y}}$ que caracteriza o modelo do trato vocal, uma vez determinada a classe, $Classe$, do quadro do segmento de fala NB em análise, as K *codewords* mais similares aos vetores $\mathbf{f}_{\mathrm{LSF},n}$ e $\mathbf{t}_n$ do *codebook* $Y^{\mathrm{UP}}_{\mathrm{CB}_{Classe}}$ são selecionadas, i.e. $\mathbf{Q}(n|Y^{\mathrm{UP}}_{\mathrm{CB}_{Classe}}) = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_K\}$, e os correspondentes vetores $\hat{\mathbf{f}}_{\mathrm{LSF},n}$ e $\hat{\mathbf{t}}_n$ de banda alta UP são combinados linearmente com um peso w para cada *codeword*. Assim,

$$\hat{\mathbf{y}}_{S_{\mathrm{UP}}}(n) = \sum_{m=1}^{K} w_m^{S_{\mathrm{UP}}} \cdot \mathbf{q}_m(n|Y^{\mathrm{UP}}_{\mathrm{CB}}) \; \forall \; S_{\mathrm{UP}} \supset [\mathbf{f}_{\mathrm{LSF}}, \mathbf{t}] \quad (5)$$

onde $\mathbf{q}_m(n|Y^{\mathrm{UP}}_{\mathrm{CB}})$ representa as *codewords* do espaço de dados de banda alta S_{UP}.

A qualidade da estimação dos parâmetros como também a qualidade do sinal de fala reconstruído são aprimoradas através de filtragem de média móvel nos parâmetros estimados $\hat{\mathbf{f}}_{\mathrm{LSF},n}$ e $\hat{\mathbf{t}}_n$,

$$\tilde{\mathbf{y}}(n) = \frac{1}{2} \cdot [\hat{\mathbf{y}}_{S_{\mathrm{UP}}}(n) + \tilde{\mathbf{y}}(n-1)] \quad (6)$$

onde $\tilde{\mathbf{y}}(n)$ representa o vetor de parâmetros estimados do quadro atual. Esse procedimento proporciona uma transição mais suave entre os parâmetros de cada quadro, atenuando artefatos presentes no processo de estimação variante no tempo [11].

V. Estágio III - Criação do Sinal de Banda Larga

Neste estágio, como ilustrado no diagrama da Fig. 1, o sinal de banda alta $\hat{s}_{\mathrm{UP}}(n)$, resultante da convolução do sinal estimado de excitação $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n)$ com os envelopes temporais e espectrais do trato vocal $\hat{\mathbf{f}}_{\mathrm{LSF},n}$ e $\hat{\mathbf{t}}_n$, é combinado com o sinal de banda estreita $s_{\mathrm{NB}}(n)$ e, assim, o sinal estimado de banda larga $\hat{s}_{\mathrm{WB}}(n)$ é obtido. Então,

$$\hat{s}_{\mathrm{WB}}(n) = \hat{s}_{\mathrm{UP}}(n) + s_{\mathrm{NB}}(n) \; \forall \; \hat{s}_{\mathrm{UP}}(n) = \hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n) * \tilde{\mathbf{y}}(n) \; . \quad (7)$$

VI. Resultados e Análise de Desempenho

Para análise de desempenho, além do sistema NB convencional da PSTN, do sistema de codificação WB e do sistema de ABWE implementado a partir da estratégia proposta neste artigo, é também utilizado, para efeito de comparação, um sistema com base no procedimento de quantização vetorial no qual é realizado um mapeamento de *codebook*, com 1024 *codewords* para cada quadro de 20 ms do sinal NB, sem que haja qualquer tipo de classificação fonética do sinal de fala [13], [5].

A Fig. 4 mostra os espectrogramas de um sinal WB original enviado pelo terminal *near-end*, de um sinal NB recebido pelo terminal *far-end* sem qualquer tratamento de ABWE e os sinais WB sintetizados através de ABWE sem e com classificação fonética, para este último, considerando nove classes distintas, e *codebooks* contendo 1024 *codewords*.

A estratégia de ABWE aqui proposta também pode ser avaliada através da análise da densidade espectral de potência dos envelopes do trato vocal dos sinais WB estimados e dos sinais em suas versões originais NB e WB. A Fig. 5 mostra um exemplo de envelope espectral obtido a partir do sinal WB original, do sinal NB convencional sem o uso de ABWE e dos sinais WB sintetizados através de ABWE sem e com o auxílio de classificação fonética.

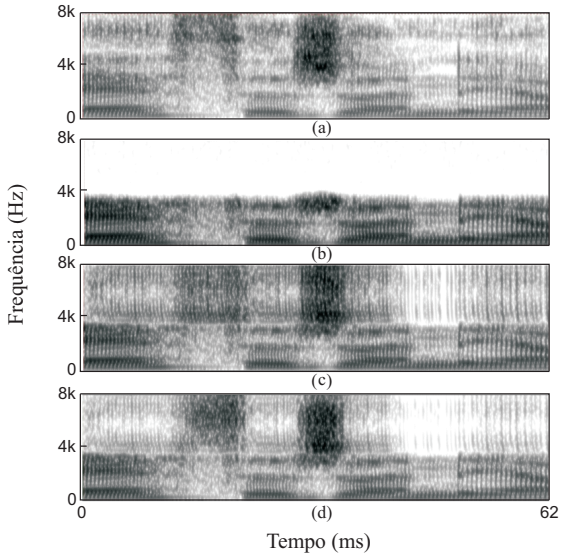

Fig. 4. Espectrogramas: (a) sinal original WB; (b) sinal recebido NB sem ABWE; (c) sinal WB sintetizado através de ABWE sem classificação fonética; (d) sinal WB sintetizado através de ABWE com classificação fonética.

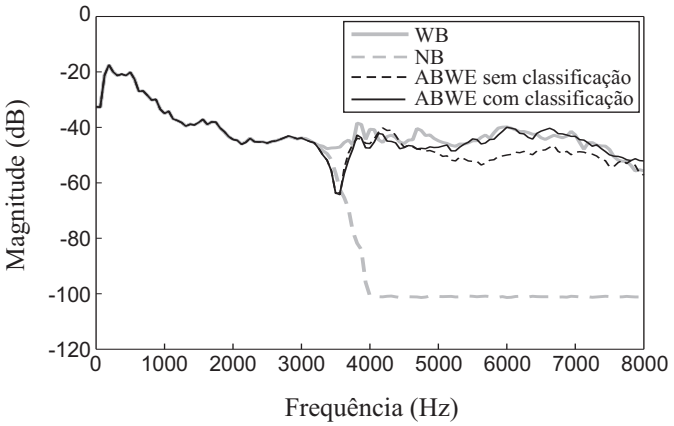

Fig. 5. Envelopes espectrais dos sinais WB, NB e sinais WB sintetizados através de ABWE sem e com classificação fonética considerando 9 classes.

De acordo com as representações dos envelopes estimados (comparados às suas versões originais NB e WB), pode ser constatado que o sinal WB sintetizado através da utilização de classificação fonética com nove classes distintas apresenta uma representação mais fiel à versão WB original.

A. Análise Subjetiva do Sinal de Fala

Nesta etapa, foram considerados testes de avaliação de acordo com as recomendações dadas em [18]. Para tal, foram coletados resultados de testes de 11 ouvintes. Em resumo, é verificada que a estratégia aqui proposta para o algoritmo ABWE é capaz de reduzir a ocorrência de artefatos nos componentes estimados de altas frequências quando comparado com procedimentos convencionais de mapeamento de *codebooks* não-supervisionados. O tratamento específico dos quadros do sinal de fala NB em suas classes fonéticas correspondentes, em especial para as classes fricativas, possibilita um processo de estimação mais fiel aos componentes WB originais. É evidente que, quando comparados os sinais sintetizados com suas versões originais WB, ainda se nota a ausência de maior "brilho" no sinal de fala. Entretanto, quando comparado com sua versão convencional NB sem tratamento ABWE, pode ser verificada uma melhoria significativa da qualidade subjetiva do sinal de fala reconstruído.

B. Análise de Qualidade Usando Medidas Objetivas

Para a avaliação da qualidade do sinal de fala WB sintetizado através da estratégia de ABWE utilizada neste trabalho de pesquisa, são adotadas medidas objetivas que simulam a análise perceptual logarítmica do ouvido humano: a medida W-PESQ (*perceptual evaluation of speech quality*) padronizada pela ITU-T [19], a razão sinal-ruído segmental calculada no domínio da frequência (FwSegSNR) e a distância espectral logarítmica (LSD) [14].

A Tabela II apresenta os resultados das medidas objetivas para os diferentes sistemas descritos no início desta seção e a Fig. 6 ilustra as curvas de desempenho dos sistemas de ABWE em relação às medidas WB-PESQs considerando diferentes números de classes fonéticas com *codebooks* variando entre 8 a 2048 *codewords*.

TABELA II
DESEMPENHO DOS SISTEMAS DE ABWE CONSIDERANDO DIFERENTES MEDIDAS DE QUALIDADE

Sistemas	LSD (dB)	FwSegSNR (dB)	WB-PESQ
WB original	$00,00$	$35,00$	$4,50$
NB sem ABWE	$11,82$	$19,90$	$1,85$
ABWE sem classificação	$07,08$	$19,58$	$2,52$
ABWE 2 classes fonéticas	$07,16$	$19,73$	$2,53$
ABWE 3 classes fonéticas	$06,90$	$21,05$	$2,64$
ABWE 5 classes fonéticas	$06,46$	$23,89$	$3,24$
ABWE 9 classes fonéticas	$06,44$	$23,94$	$3,49$

VII. Conclusões e Comentários Finais

Neste trabalho de pesquisa, uma nova estratégia para implementação de sistemas de ABWE foi apresentada. Essa estratégia utiliza coeficientes LSF para a representação do envelope espectral do trato vocal e inclui um procedimento de classificação fonética para os sinais NB. A estratégia aqui proposta proporciona sinais sintetizados de WB significativamente melhores do que os sinais NB e aqueles sintetizados a partir de ABWE convencionais. Resultados de avaliações subjetivas e objetivas também ratificam tais afirmações.

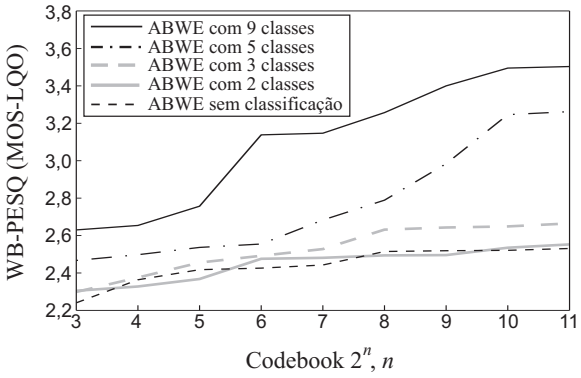

Fig. 6. Curvas de desempenho do algoritmo ABWE utilizando *codebooks* para diferentes números de *codewords*.

Referências

[1] R. ITU, *G.729 : Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*, 1996, ITU-T Rec. G.729 Std., 1996.

[2] ——, *G.729.1 : G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729*, 2006, ITU-T Rec. Rec. G.729.1 Std., 2006.

[3] B. Iser, W. Minker, and G. Schmidt, *Bandwidth Extension of Speech Signals*. New York, USA: Springer, 2008.

[4] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?" *IEEE Commun. Mag.*, vol. 44, no. 5, pp. 106–111, May 2006.

[5] T. K. Patel and P. Shrivastav, "Implementation of ITU-T G.729.ev wideband speech coder," *Int. J. Comp. Tech. Electr. Eng.*, vol. 2, no. 3, pp. 27–32, June 2012.

[6] B. Iser, P. Jax, P. Vary, H. Taddei, and S. Schandl, "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.

[7] M. Ghaderi and M. H. Savoji, "Wideband speech coding using adpcm and a new enhanced bandwidth extension method," in *Proc. IEEE Int. Symposium on Intelligent Signal Process.*, Floriana, Malta, Sept. 2011, pp. 1–4.

[8] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Philadelphia, USA, March 2005, vol. 1, pp. 805–808.

[9] D. M. Mohan, D. B. Karpur, M. Narayan, and J. Kishore, "Artificial bandwidth extension of narrowband speech using gaussian mixture model," in *Proc. Commun. and Signal Process. (ICCSP)*, Calicut, India, Feb. 2011, pp. 410–412.

[10] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve, "The effect of highband harmonic structure in the artificial bandwidth expansion of telephone speech," in *Proc. Int. Speech Commun. Conf. (INTERSPEECH)*, Antwerp, Belgium, Aug. 2007, pp. 2497–2500.

[11] T. Selvi and J. Pragatheeswaran, "Efficient speech enhancement technique by exploiting the harmonic structure of voiced segments," in *Proc. Int. Conf. Recent Trends Inf. Tech. (ICRTIT)*, Chennai, Tamil Nadu, June 2011, pp. 764–769.

[12] U. Koprnagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 86, no. 6, pp. 1296–1306, June 2006.

[13] C. Yagli and E. Erzin, "Artificial bandwidth extension of spectral envelope with temporal clustering," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, vol. 3, pp. 5096–5099.

[14] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Dallas, USA: Boca Raton: CRC Press, 2007.

[15] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Yokohama, Japan, Sep. 94, vol. 5, pp. 1591–1594.

[16] L. M. T. de Jesus, "Acoustic phonetics of european portuguese fricative consonants," Ph.D. dissertation, University of Southampton, 2001.

[17] T. C. Silva, *Fonética e Fonologia do Português - Roteiro de Estudos e Guia de Exercícios*. São Paulo, Br: Editora Contexto, 2010.

[18] R. ITU, *Methods for subjective determination of transmission quality*, 1996, ITU-T Rec. Rec. P.800 Std., 1996.

[19] ——, *P.862.2 : Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, 2007, ITU-T Rec. Rec. P.862.2 Std., 2007.