Extensão Artificial de Largura de Banda Aplicada em Reconhecimento Automático de Fala

Ênio dos Santos Silva e Rui Seara

***Resumo*— Este trabalho apresenta uma nova estratégia para a implementação de sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR) aplicados à rede pública de telefonia (*public switched telephone network* - PSTN). O estado da arte reporta que sistemas de ASR que decodificam sinais de banda estreita (*narrowband* - NB) apresentam desempenho inferior a sistemas que operam com sinais de banda larga (*wideband* - WB). Visando o aprimoramento de sistemas de ASR aplicados a PSTN, as etapas de extração de atributos do sinal da fala, bem como a etapa de construção do modelo acústico (MA), são desenvolvidas baseadas em sinais sintéticos de WB estimados a partir do realce de sinais de NB, utilizando extensão artificial de largura de banda (*artificial bandwidrh extension* - ABWE). Resultados de taxa de erro de reconhecimento são avaliados e comprovam a eficácia da estratégia proposta.**

***Palavras-Chave*— Extensão de largura de banda, realce do sinal de fala, reconhecimento automático de fala.**

***Abstract*— This paper presents a new strategy for implementation of automatic speech recognition (ASR) systems applied to the public switched telephone network (PSTN). The state of the art reports that ASR systems, that decode narrowband (NB) signals, exhibit a lower performance than the ones operating with wideband (WB) signals. In order to improve PSTN-ASR systems, WB signals synthetically estimated [using NB signals enhanced by artificial bandwidth extension (ABWE)] are used to obtain more discriminating attributes of speech signals, as well as to achieve better performance of acoustic models (AM). Results of word error rate (WER) are presented confirming the effectiveness of the proposed strategy.**

***Keywords*— Artificial bandwidth extension, speech enhancement, automatic speech recognition.**

I. Introdução

Atualmente, o mercado de telefonia dispõe de diversos serviços interativos, tais como os serviços presentes em sistemas automatizados de *help desk*, portais de voz, gerenciamento de diálogos em unidades de resposta audível (URA) e outros tipos de atendimento via *call centers* [1], como ilustrado na Fig. 1. Tais serviços auxiliam o acesso de usuários da rede pública de telefonia (*public switched telephone network* - PSTN) a informações via navegação em menus interativos. Esses menus podem ser acessados com a ajuda do teclado telefônico ou diretamente através de comandos de fala, nos quais o usuário é atendido por um assistente virtual auxiliado por reconhecimento automático de fala (*automatic speech recognition* - ASR) aplicado à PSTN [2]. Tendo em vista que os serviços comandados por fala são altamente dependentes do desempenho do ASR, as pesquisas sobre modelagem estatísticas desses sistemas continuam em franca evolução e se destacam como um tópico ativo de pesquisa em processamento digital de sinais [3]–[5].

Ênio dos Santos e Rui Seara, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica e Eletrônica, 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).

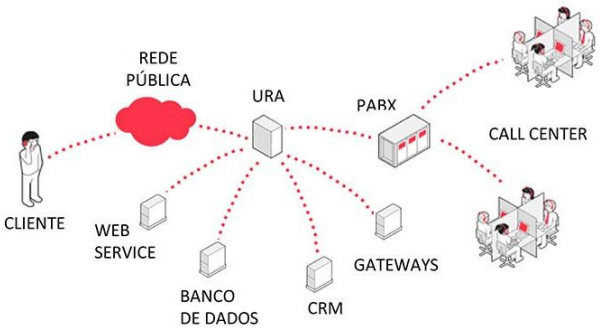

Fig. 1. Exemplos de serviços disponíveis na rede telefônica.

Resumidamente, um sistema típico de ASR é composto por duas etapas principais, o *front-end* e o *back-end*. A etapa de *front-end* recebe o sinal de entrada e efetua a extração de vetores de observação contendo informações codificadas da fala. A etapa de *back-end*, composta por decodificador, dicionário fonético, modelo acústico e modelo de linguagem, é responsável pela decodificação dos vetores de observação e pelo processo de "busca" das informações linguísticas, por exemplo, fonemas e/ou palavras, contidas no sinal de fala [6].

Os sistemas de ASR aplicados a redes de telefonia podem ser implementados de acordo com três diferentes metodologias: reconhecimento de fala embarcado (*embedded speech recognition* - ESR), que contêm as etapas de *front-end* e *back-end* integradas ao aparelho telefônico; reconhecimento de fala distribuído (*distributed speech recognition* - DSR), em que o *front-end* é integrado ao aparelho telefônico, tendo o *back-end* hospedado em um servidor externo; e o reconhecimento de fala em rede (*network speech recognition* - NSR), em que tanto o *front-end* quanto o *back-end* são hospedados em um servidor remoto e o aparelho telefônico apenas envia o sinal de fala para ser processado nesse servidor [7].

Visando a obtenção de melhores taxas de reconhecimento, sistemas de ASR desenvolvidos para *desktops* utilizam sinais de fala em banda larga (*wideband* - WB) amostrados geralmente a taxas maiores do que 16 kHz. Entretanto, sistemas aplicados à PSTN utilizam sinais de banda estreita (*narowband* - NB) amostrados a 8 kHz e devido às características inerentes ao canal telefônico da PSTN (largura de banda de 300 a 3400 Hz) [8], além da perda de naturalidade e inteligibilidade, esses sinais quando comparados com sinais de fala de WB (com largura de banda entre 50 e 7000 Hz) também apresentam perdas de desempenho em sistemas de ASR [6].

Em [5], [8] e [9], para contornar as limitações da

comunicação em NB, a extensão artificial de largura de banda (*artificial bandwidrh extension* - ABWE) é adotada como uma alternativa interessante capaz de proporcionar melhorias na qualidade dos sinais de fala, tornando-os mais próximos aos sinais de WB e, consequentemente, mais agradáveis aos ouvidos dos usuários da PSTN.

Portanto, visando explorar os benefícios da ABWE, assim como os recursos computacionais em servidores remotos, neste trabalho, sugerimos a inclusão de um sistema de ABWE como uma etapa antecessora ao *front-end* de um ASR, utilizando uma estrutura NSR. Nesse contexto, o NSR surge como uma alternativa promissora por apresentar como principal vantagem o escalamento de recursos computacionais disponíveis em um servidor [10], tornando possível um processamento independente dos componentes de *hardware* dos aparelhos telefônicos.

Neste artigo, uma estratégia para sistemas de ASR baseado em NSR é proposta e o desempenho do ASR com ABWE é discutido. A eficácia da estratégia proposta é verificada através de avaliações objetivas da taxa de erro de palavras (*word error rate* - WER).

II. Fundamentos de reconhecimento automático de fala

O objetivo de um sistema de ASR é estimar satisfatoriamente as informações contidas em um sinal de fala. O principal propósito desse sistema é a conversão de um sinal de fala em texto. Esta seção apresenta uma breve introdução dos fundamentos de um ASR.

A. Arquitetura

Um sistema típico de ASR é composto por cinco blocos principais: *front-end*, dicionário fonético, modelo acústico, modelo de linguagem e decodificador, sendo que os quatro últimos blocos compõem uma estrutura usualmente chamada *back-end*, conforme ilustrado na Fig. 2.

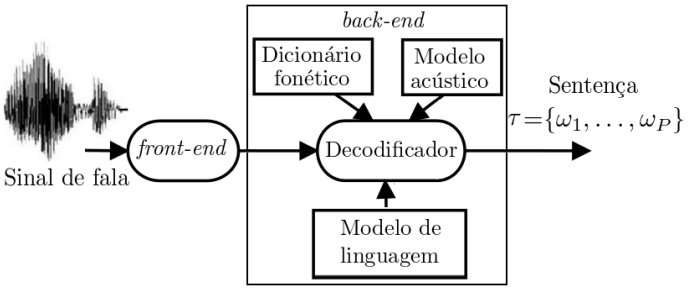

Fig. 2. Diagrama de blocos de um sistema de ASR.

O *front-end* extrai segmentos (quadros) a partir do sinal de fala e parametriza cada segmento em um vector $\mathbf{o}$ de dimensão L. Supõe-se aqui que T quadros são organizados em uma matriz $\mathbf{O}$, de dimensão $L \times T$, para representar uma sentença. O modelo de linguagem de um sistema de ASR fornece a probabilidade $p(\tau)$ de ocorrer uma sentença $\tau = \{\omega_1, \ldots, \omega_P\}$ de P palavras. Conceitualmente, o decodificador visa encontrar a sentença τ^* que maximize a probabilidade *a posteriori* dada por

$$\tau^* = \arg\max_{\tau} p(\tau|\mathbf{O}) = \arg\max_{\tau} \frac{p(\mathbf{O}|\tau)p(\tau)}{p(\mathbf{O})} \qquad (1)$$

onde $p(\mathbf{O}|\tau)$ representa a verossimilhança acústica entre a matriz de observação $\mathbf{O}$ e as palavras da sentença τ. Essa verossimilhança é determinada por um modelo acústico previamente treinado. Visto que $p(\mathbf{O})$ não depende de τ, (1) é equivalente a

$$\tau^* = \arg\max_{\tau} p(\mathbf{O}|\tau)p(\tau). \qquad (2)$$

Devido ao grande número de possíveis sentenças, (2) não pode ser computada independentemente para cada sentença candidata. Portanto, os sistemas de ASR geralmente usam estruturas de dados, tais como árvores lexicais hierárquicas, quebrando sentenças em palavras e palavras em unidades básicas como fones ou trifones [6]. O mapeamento das palavras para as unidades básicas e vice-versa é realizado através de um dicionário fonético.

Em resumo, após o treinamento dos modelos, o ASR na fase de teste usa o *front-end* para converter o sinal de entrada em atributos discriminativos e o *back-end* para estimar a sentença τ mais compatível ao contexto linguístico e ao sinal de entrada $\mathbf{O}$. Dessa forma, quanto melhor a qualidade do sinal de fala, maior o desempenho esperado do ASR. Nesse contexto, sistemas de ASR que utilizam sinais de fala de WB, com frequência de amostragem igual a 16 kHz, apresentam em média desempenho 5% superior a sistemas que utilizam sinais amostrados em 8 kHz, como é o caso dos sinais provenientes da PSTN [6].

B. Extração de Atributos

Tendo em vista a obtenção de atributos ótimos discriminativos do sinal de fala, diversas alternativas para parametrizar as formas de onda desses sinais são utilizadas [6]. Dentre elas, a parametrização utilizando coeficientes cepstrais em escala mel (*mel frequency-cepstral coefficients* - MFCC) vem mostrando-se bastante eficaz e é comumente usada no bloco de *front-end* do ASR [6]. Essa técnica de análise utiliza a escala mel, expressa por

$$m = M(f) = 1125 \cdot \ln\left(1 + \frac{f}{700}\right) \qquad (3)$$

onde f representa os componentes de frequência (em Hz).

O procedimento de extração de atributos consiste em dividir o espectro do sinal em B bandas com frequências centrais igualmente espaçadas na escala mel. Essas bandas de frequências são distribuídas através de bancos de filtros triangulares e, para cada banda, são computados os valores do logaritmo da energia e a transformada discreta do cosseno (*discrete cosine transform* - DCT). Os valores resultantes compõem os coeficientes MFCC, como ilustrado no diagrama da Fig. 3.

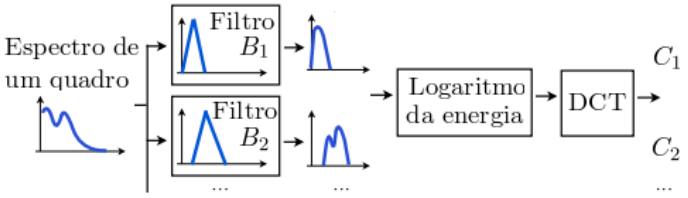

Fig. 3. Diagrama de blocos ilustrativo da extração de atributos MFCC.

Assim, a distribuição (adequada) dos componentes de frequência nas B bandas é fundamental para o cálculo dos atributos ótimos discriminativos. Portanto, a largura de banda do espectro do sinal é diretamente proporcional à qualidade dos atributos MFCC. Então, por apresentarem

uma maior faixa de frequência, sinais de WB possuem atributos de maior resolução na análise MFCC quando comparados a atributos obtidos a partir de sinais de NB [5].

Sinais de NB podem ser realçados através de algoritmos de ABWE que possibilitem a estimação de componentes de frequência adicionais capazes de ampliar a largura de banda do espectro. Assim, os componentes de frequência originais de NB e os estimados de WB podem ser redistribuídos eficientemente nos B bancos de filtros triangulares (da análise MFCC). A Fig. 4 ilustra o processo de distribuição dos espectros de NB e de WB nos filtros triangulares correspondentes às B bandas de frequência.

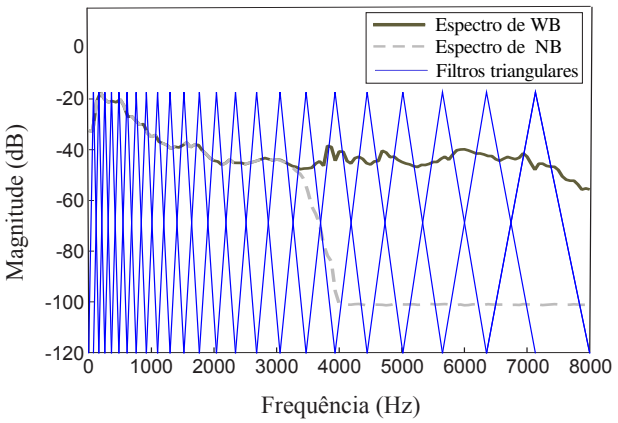

Fig. 4. Distribuição dos espectros de NB e de WB nos B bancos de filtros triangulares.

III. Fundamentos de extensão artificial de largura de banda

O objetivo de um sistema de ABWE é realçar o sinal de fala de NB, tornando-o mais agradável aos ouvintes e fazendo com que sua qualidade subjetiva se "assemelhe" a um sinal de WB. Isso é possível através da estimação dos componentes de frequência acima de 3, 4 kHz. Assim, a ideia básica de um sistema de ABWE é sintetizar artificialmente componentes de alta frequência, convertendo um sinal de NB em um sinal de WB, isto é, estimando as propriedades acústicas pertinentes à WB [5], [8].

A. *Sistema de* ABWE

A Fig. 5 ilustra o diagrama de blocos do sistema de ABWE discutido em [8] e aqui adotado. O diagrama tem como base uma estrutura de três estágios de processamento, descritos como segue:

1) Estágio I. Estimação do sinal de excitação $\hat{s}^{\text{exc}}_{\text{UP}}(n)$ através de predição linear (*code-excited linear prediction* - CELP) [9].
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 de WB.

Os estágios que compõem o sistema ABWE são descritos com detalhes em [8].

IV. Estratégia NSR usando ASR com ABWE

No cenário de telefonia, sistemas NSR são adequados à utilização dos recursos computacionais de um servidor remoto. Assim, técnicas de processamento de sinais, tal como o realce da fala, podem ser melhor exploradas. Nesse contexto, este artigo propõe uma estratégia em que o sinal de NB s_{NB}, fornecido pela PSTN, é realçado através de procedimentos de ABWE, gerando assim um sinal de WB sintético $\hat{s}_{\text{WB}}$ na entrada do ASR, como ilustrado na Fig. 6.

Portanto, o sistema NSR será composto pelos blocos de ABWE, seguido dos blocos do ASR: *front-end* e *back-end*.

A. *Construção de Codebooks para* ABWE

No estágio II do sistema de ABWE, são estimados os envelopes temporais e espectrais de banda alta (UP) que caracterizam o trato vocal no processo de geração da fala. Esses envelopes são representados pelos seguintes vetores de parâmetros

$$\mathbf{t}_n = [t_n(1), \ldots, t_n(16)]^{\text{T}} \tag{4}$$

e

$$\mathbf{f}_{\text{LSF},n} = [f_{\text{LSF},n}(1), \ldots, f_{\text{LSF},n}(19)]^{\text{T}} \tag{5}$$

onde o vetor $\mathbf{t}_n$ representa o envelope temporal contendo energias logarítmicas de 16 subquadros (1, 25 ms cada) [9] e o vetor $\mathbf{f}_{\text{LSF},n}$ contém os componentes LSFs que caracterizam o envelope espectral.

Como ilustrado na Fig. 7, o processo de treinamento do sistema de ABWE consiste na construção do conjunto de *codebooks* correspondentes aos envelopes do trato vocal no processo de geração da fala [8].

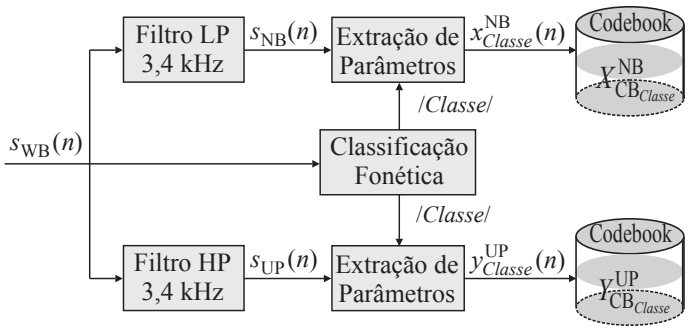

Fig. 7. Etapa do processo de treinamento de *codebooks*.

O processo de treinamento aqui utilizado considera as técnicas de *codebooks* duais de banda estreita $X^{\text{NB}}_{\text{CB}}$ e de banda alta $Y^{\text{UP}}_{\text{CB}}$ [8], criados a partir do agrupamento dos vetores de parâmetros $\mathbf{x}^{\text{NB}}$ e $\mathbf{y}^{\text{UP}}$ em classes fonéticas dispostas de acordo com a Tabela I [8]. Desse modo, um treinamento supervisionado é realizado e o agrupamento dos vetores de parâmetros $\mathbf{x}^{\text{NB}}$ e $\mathbf{y}^{\text{UP}}$ torna-se mais discriminativo.

Assim, cada *codebook* é associado a uma classe fonética específica ϕ, isto é,

$$\begin{matrix} Y^{\text{UP}}_{\text{CB}_\phi} = \text{E}\{Y^{\text{UP}} | \phi = Cl_i\} \\ X^{\text{NB}}_{\text{CB}_\phi} = \text{E}\{X^{\text{NB}} | \phi = Cl_i\} \end{matrix}, \quad \forall i \in \{1, 9\}. \tag{6}$$

B. *Estimação da Classe Fonética e dos Parâmetros do Trato Vocal para* ABWE

Para a estimação das classes fonéticas $\phi \in \{Cl_1, \ldots, Cl_9\}$, o algoritmo de árvore de decisão J48 é utilizado [11].

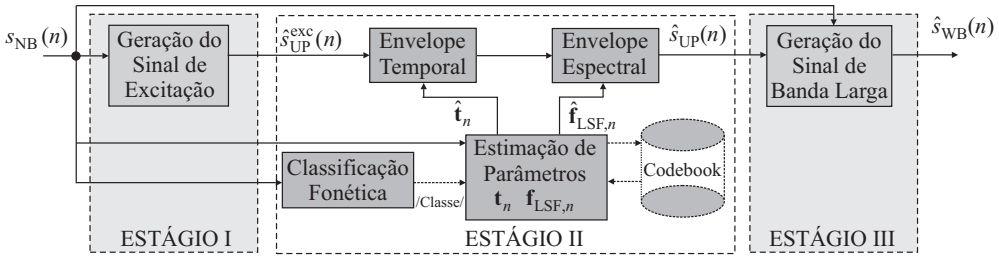

Fig. 5. Diagrama de blocos do sistema ABWE.

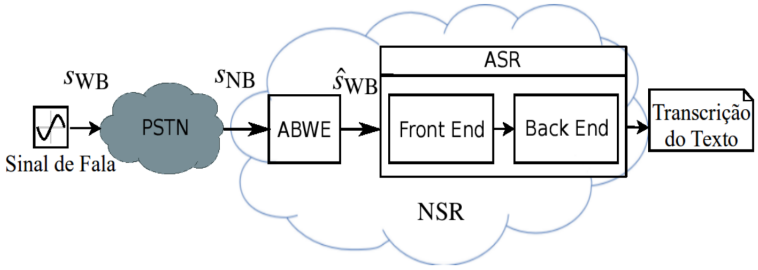

Fig. 6. Diagrama do ASR com ABWE em um sistema NSR.

TABELA I
DISTRIBUIÇÃO DAS CLASSES FONÉTICAS

Classes	Descrição
Cl_1	Fricativas vozeadas alveolares
Cl_2	Fricativas vozeadas labiodentais
Cl_3	Fricativas vozeadas palatais
Cl_4	Demais fonemas vozeados
Cl_5	Fricativas não-vozeadas labiodentais
Cl_6	Fricativas não-vozeadas alveolares
Cl_7	Fricativas não-vozeadas palatais
Cl_8	Demais fonemas não-vozeados
Cl_9	Silêncio

A cada quadro de 20 ms, são extraídos vetores de observação contendo os seguintes parâmetros: os 10 primeiros coeficientes de auto-correlação, a taxa de cruzamento por zero, o índice de gradiente, a energia normalizada do quadro, a *kurtosis* local e a centróide espectral, como apresentado em [8]. A partir dos vetores de observação, a árvore de decisão é consultada e a classe fonética correspondente é obtida. A abordagem de agrupamento de classes similares (veja Tabela I) aumenta a robustez quanto a erros de classificação. A robustez é garantida porque, mesmo havendo falhas de classificação, as classes fonéticas estimadas não estarão tão distantes de suas versões verdadeiras. Especificamente, neste artigo, o erro de classificação foi de aproximadamente 25%.

Supõe-se aqui que $\mathbf{q}^{\mathrm{UP}} \supset [\mathbf{t}, \mathbf{f}_{\mathrm{LSF}}]$ representa os vetores que caracterizam o modelo do trato vocal de banda alta. Assim, $\mathbf{q}^{\mathrm{UP}}$ é organizado em uma matriz de *codewords* $\mathbf{Q}(n|Y^{\mathrm{UP}}_{\mathrm{CB}_\phi})$, de acordo com as suas correspondentes classes fonéticas ϕ. Dessa maneira, para a estimação de $\hat{\mathbf{q}}_{UP}$ (uma vez determinada a classe ϕ) do quadro do segmento de fala de NB no instante n, as K *codewords* mais próximas (no sentido euclidiano [8]) aos vetores $\mathbf{f}_{\mathrm{LSF},n}$ e $\mathbf{t}_n$ do *codebook* $Y^{\mathrm{UP}}_{\mathrm{CB}_\phi}$ são selecionadas via *codebook dual* $X^{\mathrm{NB}}_{\mathrm{CB}_\phi}$, isto é,

$$\mathbf{Q}(n|Y^{\mathrm{UP}}_{\mathrm{CB}_\phi}) = \{\mathbf{q}^{\mathrm{UP}}_1(n|X^{\mathrm{NB}}_{\mathrm{CB}_\phi}), \mathbf{q}^{\mathrm{UP}}_2(n|X^{\mathrm{NB}}_{\mathrm{CB}_\phi}), ..., \mathbf{q}^{\mathrm{UP}}_K(n|X^{\mathrm{NB}}_{\mathrm{CB}_\phi})\} \quad (7)$$

e os correspondentes vetores $\mathbf{f}_{\mathrm{LSF},n}$ e $\mathbf{t}_n$ de banda alta são combinados linearmente com pesos w, calculados via distâncias euclidianas [8] para cada *codeword*. Assim, o modelo estimado do trato vocal de banda alta $\hat{\mathbf{q}}_{\mathrm{UP}}$ pode ser determinado através da combinação linear das k *codewords* de $\mathbf{Q}(n|Y^{\mathrm{UP}}_{\mathrm{CB}_\phi})$ com seus correspondentes fatores de pesos w.

$$\hat{\mathbf{q}}_{\mathrm{UP}}(n) = \sum_{m=1}^{K} w_m \cdot \mathbf{q}^{\mathrm{UP}}_m(n|X^{\mathrm{NB}}_{\mathrm{CB}_\phi}) \; \forall \; \mathbf{q}^{\mathrm{UP}}_m \supset [\mathbf{f}_{\mathrm{LSF}}, \mathbf{t}] \quad (8)$$

onde $\mathbf{q}^{\mathrm{UP}}_m$ representa as *codewords* constituídas pelos parâmetros $\mathbf{f}_{\mathrm{LSF}}$ e $\mathbf{t}$ de UP. Posteriormente, o sinal estimado de UP $\hat{s}_{\mathrm{UP}}$, resultante da convolução do sinal estimado de excitação $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}$ com os envelopes temporal e espectral $\hat{\mathbf{t}}_n$ e $\hat{\mathbf{f}}_{\mathrm{LSF},n}$, respectivamente, é combinado com o sinal de NB s_{NB} e, assim, o sinal estimado de WB $\hat{s}_{\mathrm{WB}}$ é obtido. Então,

$$\begin{aligned} &\hat{\mathbf{q}}^{\mathrm{UP}}(n) \supset [\hat{\mathbf{t}}_n, \hat{\mathbf{f}}_{\mathrm{LSF},n}] \\ &\hat{s}_{\mathrm{UP}}(n) = [\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n) * \hat{\mathbf{t}}_n] * \hat{\mathbf{f}}_{\mathrm{LSF},n} \\ &\hat{s}_{\mathrm{WB}}(n) = \hat{s}_{\mathrm{UP}}(n) + s_{\mathrm{NB}}(n). \end{aligned} \quad (9)$$

C. Construção de Modelos Estatísticos para o ASR

Em (2), $p(\mathbf{O}|\tau)$ e $p(\tau)$ são determinados a partir dos modelos acústicos e de linguagem MA e ML, respectivamente. A estimação de um modelo acústico satisfatório é considerada a etapa mais desafiadora do projeto de um sistema ASR. Nesse contexto, os procedimentos do ABWE discutidos na seção anterior são responsáveis pelo realce na matriz de observação $\mathbf{O}$, migrando de um espaço de dados de NB para WB. Assim, os atributos MFCC analisados pelo *front-end* contém informações espectrais mais discriminativas. Neste trabalho, utilizamos o software HTK para construir modelos acústicos usando modelos escondidos de Markov (*hidden Markov models* - HMM), de acordo com as etapas descritas em [12]. Para a construção do modelo de linguagem, adotou-se a ferramenta MITLM [13].

Abaixo seguem alguns detalhes sobre as configurações utilizadas:

- Comprimento do quadro igual a 20 ms com sobreposição de 10 ms.

- Coeficientes por quadro contendo valores de energia, 12 coeficientes cepstrais em escala mel e suas primeiras e segundas derivadas.
- MA baseado em HMM contínua de 5 estados e topologia esquerda-direita, constituídas por modelos trifônicos *cross-word* computados a partir de 38 monofones.
- ML baseado em tri-gramas treinados com 1534980 sentenças e utilizando a técnica de suavização de *Kneser-Ney* [13].

V. Resultados e Análise de Desempenho

Na maioria das aplicações de ASR, a figura de mérito usada para avaliar tais sistemas é a taxa de erro de palavra (*word error rate* - WER), definida como

$$WER = \frac{D+R}{W} \times 100\% \tag{10}$$

onde W é o número de palavras na sequência de entrada, e R e D são, respectivamente, o número de erros de substituição e de deleção na sequência de palavra reconhecida quando comparado com a sequência correta. Aqui, foi considerado o reconhecimento de fala contínua para o português brasileiro usando um vocabulário de 65 mil palavras.

Para que seja possível a obtenção de taxas aceitáveis de erro, o estado da arte de ASR, usando MA baseados em HMM, necessita de um banco de dados de áudio (*corpora*) com grande variabilidade acústica, e os ML necessitam de milhões de sentenças para uma modelagem precisa do idioma em questão. Dessa forma, a disponibilidade de *corpora* é um fator primordial para o desenvolvimento de MA e ML precisos. Um *corpora* típico possui arquivos de fala com as suas transcrições associadas, grande quantidade de textos escritos e um dicionário fonético. Para atender tais requisitos, o desenvolvimento da etapa de ASR utiliza os *corpora* de fala e texto disponibilizados em [2] e [14].

Para a análise de resultados, o corpora *LapsBenchmark* [2] é adotado e os resultados das WER são utilizados na avalização de três diferentes sistemas: sistemas de ASR com sinais de NB provenientes da PSTN (NB-PSTN); ASR com sinais de WB (WB-Original); e NSR com sistemas de ABWE (NSR-ABWE). Os resultados obtidos com os dois primeiros sistemas, usando sinais de NB provenientes da PSTN e sinais de WB, são utilizados como referência para a avaliação do desempenho da estratégia de NSR com ABWE proposta neste trabalho.

A Tabela II apresenta o desempenho dos sistemas de ASR de referência (NB-PSTN e WB-Original), assim como o desempenho do NSR-ABWE proposto, no qual os sinais de WB são sintetizados através do sistema de ABWE discutido na Seção III-A. Como apresentado na Tabela II, a estratégia de NSR-ABWE proposta aqui apresenta um desempenho intermediário entre os sistemas de ASR usando NB-PSTN e WB-Original, resultando em ganho absoluto de $1,32\%$ na WER quando comparado com a aplicação direta de sinais de NB provenientes da PSTN (NB-PSTN).

VI. Conclusões e Comentários Finais

Neste trabalho de pesquisa, uma estratégia de NSR utilizando ABWE foi apresentada. Essa estratégia implementa o realce do sinal de NB (de entrada) e resulta em um sinal estimado de WB capaz de fornecer (artificialmente) atributos cepstrais mais discriminativos ao ASR. O sinal de WB estimado (de entrada) proporcionou maior riqueza espectral ao treinamento dos modelos acústicos, quando comparado ao treinamento usando suas versões de NB provenientes da PSTN (resultando em MA mais precisos). Resultados de avaliações objetivas apresentados na Tabela II ratificam tais afirmações e confirmam a eficácia da estimação dos novos componentes de frequência do sinal sintetizado de WB $\hat{s}_{\mathrm{WB}}$, bem como a distribuição apropriada desses componentes nas B bandas de frequência da análise MFCC.

TABELA II
Resultados dos Testes Realizados

Duração do corpora	
Treinamento	Teste
4 horas	54 minutos

Desempenho do ASR	
Sistemas	**WER**
WB-Original	27,77 %
NB-PSTN	30,69 %
NSR-ABWE	**29,37 %**

Referências

[1] A. Oliveira, E. Silva, H. Macedo, and L. Matos, "Brazilian portuguese speech-driven answering system," in *Proc. 6th Euro American Conference on Telematics and Information Systems (EATIS)*, Valencia, Spain, May 2012, pp. 1–8.

[2] N. Neto, P. Silva, A. Klautau, and I. Trancoso, "Free tools and resources for brazilian portuguese speech recognition," *J. of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, Mar. 2011.

[3] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword modeling for automatic speech recognition: Past, present, and emerging approaches," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, Nov. 2012.

[4] D. O'Shaughnessy, "Acoustic analysis for automatic speech recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1038–1053, May 2013.

[5] V. F. Patrick Bauer, Johannes Abel and T. Fingscheidt, "Automatic recognition of wideband telephone speech with limited amount of matched training data," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisboa, Portugal, Sept. 2014, pp. 88–93.

[6] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[7] D. Zaykovskiy, "Survey of the speech recognition techniques for mobile devices," in *Proc. 11th Int. Conf. Speech and Comp. (SPECOM)*, St. Petersburg, Russia, Jun. 2006, pp. 88–93.

[8] Ênio Silva, "Extensão artificial de largura de banda usando classificação fonética," Master's thesis, Pós-graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina - UFSC, Florianópolis, SC, Brasil, 2016.

[9] 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.

[10] Y. Sunil and R. Sinha, "Exploration of class specific abwe for robust children's asr under mismatched condition," in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, Jul. 2012, pp. 1–5.

[11] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, pp. 1114–1119, Jun. 2013.

[12] E. Silva, L. Baptista, H. Fernandes, and A. Klautau, "Desenvolvimento de um sistema de reconhecimento automático de voz contínua com grande vocabulário para o português brasileiro," in *Anais XXV Congresso da Sociedade Brasileira de Computação*, São Leopolso, RS, Brasil, Jul. 2005, pp. 2258–2267.

[13] B.-J. Hsu and J. Glass, "Iterative language model estimation: Efficient data structure and algorithms," in *Proc. International Speech Communication Association (INTERSPEECH)*, Brisbane, Australia, Sept. 2008, pp. 841–844.

[14] C. A. Ynoguti and F. Violaro, "A brazilian portuguese speech database," in *Anais XXXI Simpósio Brasileiro de Telecomunicações (SBrT)*, Rio de Janeiro, RJ, Brasil, Set. 2008, pp. 15–20.