Refinamento da Segmentação Fonética em Aplicações de Síntese de Fala

Monique V. Nicodem, Sandra G. Kafka, Rui Seara Jr. e Rui Seara

***Resumo*— Em síntese concatenativa, a fala é produzida a partir da justaposição de segmentos automaticamente selecionados dentre unidades contidas em um banco previamente gravado e segmentado. A fala sintética obtida através dessa técnica apresenta naturalidade influenciada pela eficácia das ferramentas de segmentação. O desempenho de tais ferramentas é freqüentemente favorecido por uma abordagem híbrida resultante da associação de uma modelagem HMM com um processo de refinamento da segmentação. Esse refinamento tem sido realizado com sucesso através do uso de técnicas baseadas em redes neurais. Neste trabalho, é proposto um conjunto de redes cujo desempenho é superior ao das demais topologias de redes neurais apresentadas na literatura. Nesse caso, as redes são treinadas a partir de uma repartição do conjunto de treinamento baseada em fronteiras de fonemas com similaridades entre si.**

***Palavras-chave*— Refinamento da segmentação, Síntese concatenativa, Redes neurais artificiais, HMM.**

***Abstract*— In concatenative synthesis, speech is produced by joining segments automatically selected among units contained in a previously recorded and segmented database. The synthetic speech resulting from such a technique presents naturalness constrained by segmentation tools. The performance of these tools is often enhanced by a hybrid approach resulting from the association of an HMM modeling with a boundary refining process. Such refining has been carried out sucessfully by using techniques based on neural networks. This paper presents a set of networks that outperform other topologies discussed in the literature. These networks are trained by performing a clusterization of the training set based on those phonetic transitions with similarities between them.**

***Keywords*— Boundary refining, Concatenative synthesis, Artificial neural networks, HMM.**

I. Introdução e Formulação do Problema

Os sistemas de conversão texto-fala no estado-da-arte são capazes de transformar qualquer texto escrito em fala. Para realizar essa transformação, diversas técnicas são apresentadas na literatura da área. Dentre as existentes, uma que se destaca consideravelmente por produzir uma fala sintética de maior similaridade com a fala humana é a síntese concatenativa [1].

No processo de síntese concatenativa, a fala é gerada a partir da união de segmentos automaticamente selecionados dentre unidades contidas em um banco previamente gravado e segmentado [2]. Nos sistemas atuais de síntese, esse banco apresenta duração total da ordem de dezenas de horas [3]. Nesses sistemas, um locutor profissional grava um conjunto considerável de sentenças previamente definidas. Após a gravação, as sentenças são submetidas a um procedimento de segmentação visando determinar a localização temporal de cada unidade acústica do banco gravado. Freqüentemente, a unidade acústica básica dos sistemas de síntese é o fonema. A fala sintética obtida através de tal técnica apresenta naturalidade limitada por alguns fatores, dentre os quais podemos citar: variabilidade fonética e prosódica do *corpus* de texto escolhido para gravação, qualidade da voz do locutor e eficácia das ferramentas de segmentação. Desse modo, por exemplo, uma ferramenta de segmentação que apresente pouca precisão na localização das fronteiras entre unidades, freqüentemente, conduz a uma redução de naturalidade da fala sintética. Problemas advindos de uma segmentação pouco precisa levam à inserção na fala sintética de algum fonema indesejado, deleção de um fonema ou redução de sua duração.

Por este motivo, o desenvolvimento de técnicas eficazes de segmentação tem sido atualmente um tema de grande interesse de pesquisas. É importante mencionar que, para tal aplicação, torna-se inviável realizar o processo de segmentação de forma manual, visto que demandaria um tempo considerável de realização para um banco de fala com duração da ordem de horas. Outro problema seria uma provável perda de consistência nas marcas manualmente obtidas, especialmente se o tempo envolvido no processo de segmentação manual fosse extensivamente grande. É importante considerar que a precisão da segmentação é, em geral, avaliada a partir da obtenção do percentual de transições entre fonemas que apresentem erros de segmentação inferiores a 20 ms. Isso porque erros de segmentação inferiores a 20 ms são perceptualmente inaudíveis [4].

O procedimento automático atualmente mais utilizado no processo de segmentação consiste em alinhar a transcrição fonética de uma dada frase com o correspondente sinal de fala (alinhamento forçado). Nesse caso, para a realização de tal alinhamento, é comum a adoção de técnicas de reconhecimento de fala. Assim, um conjunto de parâmetros extraídos do sinal de fala é tomado como referência para representar cada unidade acústica através de um modelo oculto de Markov (*hidden Markov model* – HMM) [5]. Em seguida, o algoritmo de Viterbi [6] é adotado para o alinhamento e a definição das fronteiras entre segmentos.

O desempenho dos sistemas baseados em HMM pode ser favorecido por uma abordagem híbrida na qual há a associação da técnica HMM com um processo de refinamento da segmentação. Para realizar tal refinamento, algumas abordagens são apresentadas na literatura. Dentre as técnicas existentes,

Monique V. Nicodem, Sandra G. Kafka, Rui Seara Jr. e Rui Seara, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mails: {monique, kafka, seara, ruijr}@linse.ufsc.br.

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Financiadora de Estudos e Projetos (FINEP) e Dígitro Tecnologia Ltda.

as baseadas em redes neurais artificiais (RNA) têm conduzido a resultados satisfatórios. Nesse caso, um banco manualmente segmentado e parametrizado é tomado como referência para o treinamento de um conjunto de redes neurais. Tais redes indicam (usando um único neurônio na camada de saída), para cada quadro do sinal de fala, qual a probabilidade de um dado quadro representar a fronteira entre dois fonemas.

Em [7], uma única rede é adotada para determinar tal probabilidade. Nesse caso, todos os tipos de transições fonéticas são adotados para treinamento da rede. Ainda em [7], é apresentada outra topologia de rede com desempenho superior àquela que usa uma única RNA [4]. Nessa técnica, quatro redes são adotadas para determinar a probabilidade de existência de fronteiras entre fonemas, sendo cada uma dessas redes treinada de acordo com o tipo de transição fonética: vozeada/não-vozeada, vozeada/vozeada, não-vozeada/vozeada e não-vozeada/não-vozeada. Essa melhoria de desempenho indica que uma "clusterização" do conjunto de treinamento conduz a melhorias no desempenho de cada rede treinada. Em [4], esse processo de "clusterização" é realizado de forma automática. Nesse processo, quatro redes são inicialmente treinadas a partir de uma divisão do conjunto de treinamento de forma arbitrária. Para cada difone contido no conjunto de treinamento, verifica-se qual rede treinada conduz ao menor erro na estimativa das probabilidades de fronteiras. Os difones são reclassificados para a rede que apresente menor erro associado ao respectivo difone, sendo posteriormente retreinadas as quatro redes. Esse procedimento se repete de forma iterativa até que as redes apresentem pouca variação de erro. A desvantagem desta última técnica consiste em definir, para a primeira iteração, os tipos de transições que devem fazer parte do conjunto de treinamento de cada rede. O autor não descreve em seu artigo como deve ser feita essa inicialização. Dependendo das condições de inicialização, o desempenho do algoritmo pode ser alterado.

Neste trabalho, visando evitar uma dependência do desempenho do algoritmo com as condições de inicialização e, ao mesmo tempo, obter um melhor desempenho do que o apresentado em [7], o seguinte procedimento é adotado: o processo de "clusterização" e divisão do conjunto de treinamento é realizado a partir de uma inspeção visual do espectrograma de transições entre fonemas com similaridades entre si. Nesse caso, é realizada a divisão do conjunto de treinamento em 36 subconjuntos e posteriormente 36 RNAs são treinadas. O desempenho dessa topologia de rede proposta é um pouco superior ao desempenho das demais topologias de redes apresentadas na literatura (aqui também implementadas) para qualquer número de sentenças de treinamento.

O desempenho da técnica de refinamento da segmentação aqui proposta supera o das demais apresentadas na literatura; no entanto, o foco deste artigo é a determinação de um tamanho apropriado do conjunto de treinamento, o que em nosso conhecimento não é abordado em trabalhos anteriores de refinamento da segmentação baseado em redes neurais. A determinação desse tamanho evita que um número excessivo e desnecessário de sentenças sejam manualmente segmentadas, resultando em uma economia do tempo requerido para segmentação manual, que é, além de tudo, um processo exaustivo e demorado. Através de experimentos realizados, é possível verificar que o aumento do conjunto de treinamento implica uma melhoria de desempenho até um certo número de sentenças. A partir desse ponto, há uma saturação de desempenho.

II. Segmentação por Modelagem HMM

Para realizar a segmentação, é necessário primeiramente obter a transcrição fonética do sinal de fala. Essa transcrição é tomada como referência para a obtenção do modelo HMM e a realização de um alinhamento forçado com o sinal de fala. Para efetuar o treinamento do modelo, duas metodologias podem ser consideradas: uma baseada em um pequeno conjunto de sentenças manualmente segmentadas e outra em um conjunto maior de treinamento sem qualquer informação sobre a segmentação manual. Neste último caso, os modelos são treinados utilizando o algoritmo de reestimação de Baum-Welch [8][1], algoritmo no qual o modelo é obtido por uma maximização de verossimilhança. Caso contrário, um modelo que represente apropriadamente a segmentação manual é obtido através de um algoritmo de Viterbi[2]. Nesse caso, com a informação da segmentação manual, verifica-se um desempenho superior de segmentação [4].

III. Refinamento da Segmentação

Após a obtenção da localização de cada fonema a partir da segmentação baseada em HMM, as fronteiras são refinadas por RNAs. Para o treinamento dessas redes, um banco de fala é inicialmente parametrizado e tais parâmetros são tomados como entrada das RNAs. Neste trabalho, os parâmetros de entrada são similares aos apresentados em [4]. Assim, são fornecidos como parâmetros: 13 MFCCs de quatro quadros consecutivos, as taxas de cruzamento por zero (ZCR - *zero crossing rate*) do primeiro e quarto quadros, uma taxa de transição de característica espectral (SFTR - *spectral feature transition rate* [10]) e a distância de Kullback-Leibler simétrica (SKLD - *symmetrical Kullback-Leibler distance* [11]), totalizando 56 parâmetros.

A medida SFTR é obtida como segue. Considere $\mathbf{x}(n) = [x_1(n) \quad x_2(n) \quad \dots \quad x_l(n)]^{\mathrm{T}}$ o vetor de coeficientes MFCC do *n*-ésimo quadro, onde $x_i(n)$ representa o *i*-ésimo coeficiente MFCC e l, o número total de coeficientes por quadro. A medida em questão é dada por

$$s(n) = \sum_{i=1}^{l} \left[\frac{\sum_{m=-M}^{M} m x_i(n+m)}{\sum_{m=-M}^{M} m^2} \right]^2 . \qquad (1)$$

Aqui, é adotado um valor de M igual a 2.

A distância SKLD é expressa como

$$D_{SKL}(n) = \int_0^{\pi} [P_n(\omega) - P_{n+1}(\omega)] \log \left[\frac{P_n(\omega)}{P_{n+1}(\omega)} \right] d\omega \qquad (2)$$

[1]A reestimação de Baum-Welch é realizada com a função HRest do HTK [9].

[2]O modelo baseado em informações da segmentação manual é obtido com a função HInit do HTK [9].

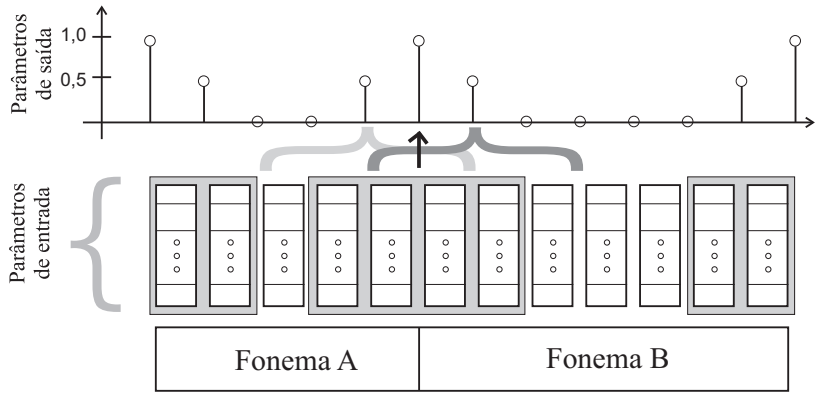

Fig. 1. Procedimento para escolha do valor de saída da rede.

onde $P_n(\omega)$ e $P_{n+1}(\omega)$ representam os envelopes espectrais (transformadas de Fourier) dos dois quadros centrais consecutivos (n e $n+1$, respectivamente).

A cada conjunto de 56 parâmetros de entrada é associado um dentre os seguintes valores de saída alvo: 0; 0,5 e 1. O valor nulo indica que os quatro quadros de análise não compreendem regiões de transição entre fonemas. O valor 1 aponta uma transição entre os dois quadros à direita e os dois quadros à esquerda. Por fim, o valor 0,5 indica a existência de uma transição entre o terceiro e o último quadro ou entre o primeiro e o segundo quadro [4]. Esse procedimento é ilustrado na Fig. 1.

É importante mencionar que, como no conjunto de treinamento está contido um número maior de vetores cuja saída é "0" (sem transição), para evitar que essa rede tenha uma tendência de gerar valores nulos, deve haver um cuidado especial no modo de apresentação dos parâmetros para a rede. Nesse caso, para uma dada época, essa apresentação é realizada procurando balancear a quantidade de vetores de mesma saída. Para tal, pode haver repetições dos vetores de entrada cuja saída é 0,5 ou 1.

Após o treinamento da rede, os 56 parâmetros de um dado arquivo de fala podem ser adotados como entrada da rede treinada. Assim, para cada conjunto de parâmetros, é obtida uma saída correspondente (escore). O ponto de maior escore em uma dada região de busca é considerado como o novo ponto de transição entre fonemas. Nesse caso, a região de busca utilizada se estende da metade de um fonema até a metade do fonema seguinte, conforme ilustrado na Fig. 2.

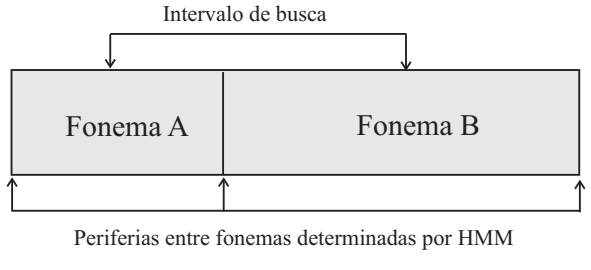

Fig. 2. Região de busca.

IV. Refinamento da Segmentação Dependente de Contexto

Sabe-se que os parâmetros acústicos de um dado segmento fonético são influenciados pelos seus contextos vizinhos. Em virtude disso, uma estimativa mais precisa das fronteiras entre fonemas pode ser alcançada quando inseridas informações fonéticas no processo de refinamento. Assim, o uso de um método de refinamento dependente de contexto, no qual uma rede especializada é selecionada de acordo com o tipo de transição fonética, pode ser adotado para melhorar o desempenho da segmentação. No entanto, ainda existem discussões sobre quais transições considerar em cada rede especializada. Em [7] e [4], são apresentadas algumas estratégias para particionar os tipos de transições entre fonemas. Entretanto, neste trabalho, resultados indicam que esses métodos apresentam ainda alguns problemas de desempenho.

Assim, a segmentação do espaço acústico é aqui realizada com base na inspeção visual do espectrograma dos segmentos de fala. Dessa maneira, aqueles tipos de transições detectadas como similares entre si são agrupados em uma única rede. Assume-se, então, que o sistema visual humano comporta-se como um classificador cujo desempenho supera o das demais técnicas automáticas de classificação, tal como a adotada em [4]. Na Tabela I, são apresentadas as classes de fonemas consideradas para definir os tipos de transições com seus respectivos símbolos [12].

Na Tabela II, são mostradas as classes de transições consideradas. A primeira, por exemplo, representa a transição de vogais orais para consoantes plosivas surdas. Isso significa que todos os quadros do sinal de fala adotado no treinamento e contidos no intervalo que se inicia na metade de uma vogal oral até a metade de uma consoante plosiva surda são adotados no treinamento de uma rede específica para este padrão de transição. Dessa maneira, um total de 36 redes são treinadas. É importante mencionar que a literatura da área apresenta uma técnica de refinamento baseada em contexto. Entretanto, os contextos adotados são diferentes dos aqui considerados e essa técnica não se baseia em redes neurais [13].

Após o treinamento das redes, este conjunto de 36 redes é usado para refinar o ponto de segmentação. Assim, no intervalo de busca sob consideração, a nova posição é atribuída àquela

TABELA I
CLASSES DE FONEMAS

Símbolo	Fonemas
AFR	Africadas
ASO	Africadas sonoras
ASU	Africadas surdas
CN	Consoantes nasais
FRI	Fricativas
FRA	Fricativas alveolares e palatais
FRL	Fricativas labiodentais
FRC	Fricativas de coda surdas
FRV	Fricativas velares
LAT	Laterais
LTA	Laterais alveolares
LTP	Laterais palatais
PLO	Plosivas
PSO	Plosivas sonoras
PSU	Plosivas surdas
VS	Vogais e semivogais
VSN	Vogais e semivogais nasais
VSO	Vogais e semivogais orais
TAP	Taps

TABELA II
TIPOS DE TRANSIÇÕES ENTRE FONEMAS

1º fonema	2º fonema	1º fonema	2º fonema
VSO	PSO	TAP	FRI
VSO	PSU	TAP	CN
VSO	FRL	TAP	AFR
VSO	FRA	FRC	PLO
VSO	CN	FRC	FRI
VSO	LAT	FRC	CN
VSO	TAP	FRC	AFR
VSO	FRV	PSO	VS
VSO	ASU	PSU	VS
VSO	ASO	FRL	VS
VSN	PSO	FRA	VS
VSN	PSU	CN	VS
VSN	FRI	LTA	VS
VSN	CN	LTP	VS
VSN	TAP	FRV	VS
VSN	FRV	TAP	VS
VSN	AFR	ASU	VS
TAP	PLO	ASO	VS

região cujo respectivo conjunto de parâmetros assume o maior valor de saída da rede que caracteriza sua transição. Esse ponto de segmentação só é alterado quando esse valor máximo superar um limiar previamente estabelecido. O limiar aqui adotado é de 0,75.

É importante salientar que aquelas transições não englobadas na Tabela II não estão sujeitas ao refinamento da segmentação. Um caso freqüente e não abordado nessa tabela é a transição entre fonemas vocálicos. Neste trabalho, a segmentação entre fonemas vocálicos não é refinada visto que há uma grande coarticulação entre esses fonemas, o que dificulta, inclusive para um ouvinte experiente, identificar onde termina um fonema e inicia o próximo. Assim, há pouca consistência na segmentação manual desses fonemas. Esse problema pode conduzir a uma redução da precisão de segmentação nessas transições.

V. RESULTADOS EXPERIMENTAIS

Para avaliar o procedimento de refinamento da segmentação aqui proposto, dois *corpora* são considerados. O primeiro é constituído por um total de 450 sentenças gravadas pelo locutor profissional de nosso sistema de síntese em uma sala acusticamente isolada. Essas sentenças são foneticamente transcritas e manualmente segmentadas por um lingüista experiente. A quantidade de difones distintos e o número total de difones contidos nesse corpus (para distintos subconjuntos) são apresentados na Tabela III.

O segundo *corpus* (correspondente ao nosso sistema de síntese) é constituído por aproximadamente 11.500 sentenças. Esse *corpus* é tomado como referência para obter um modelo HMM por fonema. Nesse caso, como as sentenças não são manualmente segmentadas, a informação da segmentação manual não é adotada no treinamento do modelo. O treinamento

TABELA III
QUANTIDADE DE DIFONES DISTINTOS E NÚMERO TOTAL DE DIFONES CONTIDOS NO CONJUNTO DE TREINAMENTO

Número de sentenças	Difones distintos	Total de difones
50	985	5.655
100	1.222	12.114
150	1.318	17.507
200	1.370	21.039
250	1.468	28.467
300	1.504	33.630
350	1.568	41.334
400	1.596	48.212

é realizado através do *software* HTK [9]. Assim, as sentenças são segmentadas em quadros, utilizando uma janela de *Hamming* com 25 ms de duração e um deslocamento entre quadros de 10 ms. Em seguida, cada quadro do sinal de fala é parametrizado utilizando 13 coeficientes MFCC (*Mel Frequency Cepstrum Coefficients*) [14] com seus 13 componentes delta e 13 componentes de aceleração, totalizando 39 coeficientes. Adota-se um fator de pré-ênfase igual a 0,97. Após a parametrização, um modelo com cinco estados (sendo três emissores) é considerado para representar cada fonema do português brasileiro (PB). O número de fonemas aqui modelados é de 53. Modelos adicionais para pausas curtas e segmentos de silêncio são também adotados. A topologia desses modelos é *left-to-right*, sendo permitida uma transição direta do estado 2 (primeiro estado emissor) para o estado 4 (terceiro estado emissor). Nesse caso, as matrizes de covariância do modelo são diagonais. O número de misturas gaussianas adotado para modelar a função densidade de probabilidade de cada estado emissor é variado de 2 a 4. A Tabela IV apresenta o percentual de transições entre fonemas com erros

de segmentação inferiores a 20 ms (para 2, 3 e 4 gaussianas), considerando 11.500 sentenças de treinamento e 50 sentenças de teste.

TABELA IV
DESEMPENHO DA SEGMENTAÇÃO NA AUSÊNCIA DE REFINAMENTO DESCONSIDERANDO INFORMAÇÕES DE SEGMENTAÇÃO MANUAL PARA OBTENÇÃO DO MODELO HMM

Número de gaussianas	Percentual
2	79,24%
3	80,49%
4	80,57%

Aquela segmentação responsável pelo melhor desempenho descrito na Tabela IV (obtido com quatro gaussianas) é refinada através das seguintes técnicas: (a) técnica 1 – baseada em uma única rede neural para qualquer tipo de transição [7]; (b) técnica 2 – quatro redes com transição definida pelo vozeamento [7]; (c) técnica 3 – quatro redes com transições automaticamente obtidas [4]; (d) técnica 4 – técnica aqui proposta. Para tais refinamentos, o banco de fala de treinamento é inicialmente parametrizado. Parâmetros mel-cepstrais são obtidos adotando as seguintes configurações: janela de *Hanning* com duração de 20 ms, deslocamento de 10 ms e fator de pré-ênfase de 0,95. As medidas SFTR e SKLD são calculadas a partir dos dois quadros centrais consecutivos. As taxas de cruzamento por zero dos primeiros e quartos quadros são também calculados. Ao final dessa parametrização, os 56 parâmetros resultantes são normalizados por suas respectivas médias e variâncias. Tais parâmetros são tomados como entrada das redes, sendo assim necessário um número de 56 neurônios de entrada para cada rede. Nesse caso, *perceptrons* com três camadas de neurônios (de entrada, oculta e de saída) são treinados utilizando o algoritmo de retropropagação do erro [15], [16]. Nesse caso, um procedimento de busca exaustiva é adotado para escolher a taxa de aprendizado do algoritmo e a quantidade de neurônios na camada oculta (0,6 e 18, respectivamente). A ferramenta computacional considerada nesse treinamento é a biblioteca FANN (*fast artificial neural network*), a qual é escrita em linguagem C [17]. Dentre as quatro técnicas de refinamento previamente mencionadas, a única cujo resultado é dependente das condições de inicialização das redes é a terceira. Assim, para essa técnica, as redes são inicialmente divididas tomando como referência a classificação quanto ao vozeamento dos fonemas vizinhos (não-vozeado/não-vozeado, não-vozeado/vozeado, vozeado/não-vozeado e vozeado/vozeado).

A Tabela V apresenta o percentual de transições entre fonemas com erros de segmentação inferiores a 20 ms, considerando estados emissores modelados por uma mistura de 4 gaussianas (o melhor caso anterior). Além do mais, o número de sentenças de treinamento é variado de 50 a 400 em passos de 50. Verifica-se, por inspeção da tabela, que a técnica proposta apresenta um desempenho superior ao das demais aqui avaliadas.

As 400 sentenças supramencionadas são posteriormente tomadas como referência para obtenção de um modelo HMM para cada fonema. Nesse caso, a modelagem é obtida considerando informações advindas da segmentação manual. A Tabela VI apresenta os percentuais de transições entre fonemas com erros de segmentação inferiores a 20 ms, obtidos na segmentação por HMM considerando 1, 2, 3 e 4 gaussianas (na modelagem de cada estado emissor). O número de sentenças é também variado de 50 a 400 considerando passos de 50. Para cada número de sentenças, o seu respectivo caso de melhor desempenho é destacado (em negrito) na Tabela VI. É importante mencionar que, em um teste realizado com 25 sentenças, não foi possível construir um modelo HMM, pois, nesse caso, o número de observações (ocorrências de cada fonema) é inferior ao mínimo necessário para obtenção do modelo.

TABELA V
DESEMPENHO DA SEGMENTAÇÃO APÓS O REFINAMENTO DESCONSIDERANDO INFORMAÇÕES DE SEGMENTAÇÃO MANUAL PARA OBTENÇÃO DO MODELO HMM

Número	Técnica 1 [7]	Técnica 2 [7]	Técnica 3 [4]	Proposta
50	77,72%	80,23%	81,34%	81,63%
100	82,42%	84,57%	81,23%	84,66%
150	82,31%	84,97%	81,76%	85,78%
200	83,49%	85,80%	81,92%	86,15%
250	83,23%	85,36%	82,05%	86,66%
300	83,69%	86,63%	81,98%	86,87%
350	83,24%	86,48%	81,96%	**87,16%**
400	80,58%	86,96%	78,40%	86,96%

TABELA VI
DESEMPENHO DA SEGMENTAÇÃO NA AUSÊNCIA DE REFINAMENTO CONSIDERANDO INFORMAÇÕES DE SEGMENTAÇÃO MANUAL PARA OBTENÇÃO DO MODELO HMM

	1	2	3	4
50	88,98%	90,00%	**90,15%**	89,88%
100	89,71%	90,35%	**91,43%**	90,69%
150	89,91%	90,62%	**91,26%**	90,95%
200	89,71%	90,84%	90,67%	**91,11%**
250	89,68%	91,14%	91,20%	**91,26%**
300	89,79%	90,81%	91,20%	**91,56%**
350	90,04%	91,14%	91,24%	**91,58%**
400	89,92%	91,15%	**91,61%**	91,35%

É importante destacar que o uso de reestimações após obtenção do modelo baseado em informação manual reduz o desempenho do processo de segmentação. Por exemplo, para 400 sentenças de treinamento e uma gaussiana, o desempenho é reduzido de 90,05% (conforme Tabela VI) para 89,07%.

Por fim, as sentenças segmentadas com informação da segmentação manual no treinamento do modelo HMM são, então, submetidas ao refinamento. Assim, a Tabela VII apresenta os percentuais de transição entre fonemas com erros inferiores a 20 ms resultantes do refinamento.

Verifica-se por inspeção da Tabela VII que, em alguns casos, o desempenho é reduzido após o refinamento da segmentação.

TABELA VII
DESEMPENHO DA SEGMENTAÇÃO APÓS O REFINAMENTO CONSIDERANDO INFORMAÇÕES DE SEGMENTAÇÃO MANUAL PARA OBTENÇÃO DO MODELO HMM

Número	Técnica 1 [7]	Técnica 2 [7]	Técnica 3 [4]	Proposta
50	84,96%	82,20%	89,81%	86,25%
100	85,76%	87,57%	91,43%	89,84%
150	85,87%	87,92%	91,09%	90,13%
200	86,28%	88,58%	90,83%	90,26%
250	86,62%	88,28%	91,20%	90,88%
300	87,25%	89,56%	91,58%	91,50%
350	86,85%	89,56%	91,54%	92,01%
400	91,61%	90,88%	88,54%	**92,03**%

Isso significa que esse refinamento só é eficaz a partir de um dado número de sentenças de treinamento. Até um certo número, o refinamento obtido pode até prejudicar a precisão da segmentação. Constata-se que a técnica proposta resulta em uma melhoria da segmentação para um número de sentenças igual ou superior a 350. As demais técnicas freqüentemente conduzem a uma redução da precisão de segmentação. A técnica aqui proposta, comparativamente à apresentada em [4], traz a vantagem de ser independente de condições de inicialização.

VI. CONCLUSÕES

Neste trabalho, é proposta uma técnica para refinamento das marcas de segmentação visando a síntese concatenativa de fala. Essa técnica supera o desempenho de outras técnicas apresentadas na literatura, resultando em um aumento do percentual de transições entre fonemas com erros de segmentação inferiores a 20 ms. Essa melhoria de segmentação implica um aumento da qualidade da fala sintética advinda de sistemas de síntese concatenativa. Para trabalhos futuros, pretende-se verificar o desempenho de uma abordagem híbrida na qual a abordagem proposta por Lee [4] é inicializada com o conjunto de redes aqui proposto.

REFERÊNCIAS

[1] F.-C. Chou, C.-Y. Tseng, and L.-S. Lee, "An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis," *Speech Communication*, vol. 48, no. 1, pp. 45–56, Jan. 2006.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'96)*, Atlanta, USA, May 1996, pp. 373–376.

[3] H. Kawai, T. Toda, J. Ni, et al., "Ximera: A new TTS from ATR based on corpus-based technologies," in *Proc. ISCA Tutorial and Research Workshop on Speech Synthesis (SSW'04)*, Pittsburg, USA, June 2004, pp. 179–184.

[4] K.-S. Lee, "MLP-based phone boundary refining for a TTS database," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 3, pp. 981–989, May 2006.

[5] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

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

[7] D. T. Toledano, "Neural network boundary refining for automatic speech segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'00)*, Istanbul, Turkey, June 2000, pp. 3438–3441.

[8] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: IEEE Press, 2000.

[9] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.1)*. Cambridge University, 2001.

[10] C. R. N. Athaudage and M. Lech, "On optimal modeling of speech spectral transitions," in *Proc. Int. Conf. Information, Communications, Signal Processing (ICICS'03)*, Singapore, Dec. 2003, pp. 1330–1334.

[11] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 1, pp. 39–51, Jan. 2001.

[12] T. C. Silva, *Fonética e Fonologia do Português: Roteiro de Estudos e Guia de Exercícios*. São Paulo: Contexto, 1999.

[13] L. Wang, Y. Zhao, M. Chu, F. K. Soong, J. Zhou, and Z. Cao, "Context dependent boundary model for refining boundaries segmentation of TTS units," *IEICE Trans. Information and Systems*, vol. E89-D, no. 3, pp. 1082–1091, Mar. 2006.

[14] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'01)*, vol. 1, Salt Lake City, USA, May 2001, pp. 73–76.

[15] S. Haykin, *Redes Neurais*, 2^a Ed. Porto Alegre: Bookman, 2001.

[16] F. M. de Azevedo, L. M. Brasil e R. C. L. de Oliveira, *Redes Neurais com Aplicações em Controle e em Sistemas Especialistas*. Florianópolis: Bookstore, 2000.

[17] S. Nissen, A. Spilca, and A. Zabot, "Fast artificial neural networks (FANN)," 2007, [Online]. Available at: http://leenissen.dk/fann/.