Compressão de Sinais de Fala Baseada em Segmentos Etiquetados Foneticamente

Carlos Meneses Ribeiro
ISEL-CEDET/INESC

Isabel Trancoso
IST/INESC

INESC, R. Alves Redol nº 9 Lisboa
01.3100314 cmr@inesc.pt

RESUMO

De entre as técnicas de codificação de sinais de fala, a codificação fonética é aquela que permite atingir os mais baixos débitos binários. Esta eficiência deve-se ao facto deste tipo de codificação se basear numa cadeia de reconhecimento (emissor) e síntese (receptor). Os codificadores fonéticos segmentam o sinal através de um reconhecedor fonético, sendo transmitido o índice do segmento reconhecido e a respectiva informação prosódica para o receptor, para sintetizar o sinal de fala.
A "reconhecibilidade" do orador é um dos problemas principais da codificação fonética, dado o tipo e a quantidade de informação transmitida. Este artigo descreve uma metodologia de adaptação ao orador no âmbito dos codificadores fonéticos, que permite codificar sinais de fala com um débito binário de 560 bit/s. Descrevem-se ainda os testes de inteligibilidade e reconhecibilidade do orador efectuados para avaliação do codificador, cujos resultados sugerem que as melhorias introduzidas no sinal sintetizado, pela metodologia de adaptação ao orador, não se traduzem apenas na reconhecibilidade do orador, mas também na inteligibilidade e na qualidade geral. De forma a tirar partido da correlação intra-orador, é também proposto um método incremental de adaptação, que permite um compromisso entre o débito binário e a qualidade.

INTRODUÇÃO

A comunidade de investigadores em codificação de fala tem investido um grande esforço para reduzir o débito binário, não só para conseguir comunicar em canais de baixa capacidade mas também para multiplexar mais sinais no mesmo canal de comunicação. Da mesma forma, o advento das tecnologias multimédia e a necessidade de armazenamento de grandes quantidades de informação, exige a necessidade de reduzir o débito binário na representação de sinais de fala, já que este determina o espaço requerido na unidade de armazenamento.
A cerca de 2400 bit/s, os codificadores representam os sinais de fala de um modo totalmente paramétrico e exploram a redundância exibida pelos sinais de fala, sendo conhecidos algumas métodos de codificação produzindo boa qualidade [2][4]. De modo a tirar partido desta redundância, os sinais são processados por tramas, com uma duração típica entre os 10 e os 30 ms, em que se considera o sinal quase-estacionário. Para muito baixo débito binário, é necessário explorar a correlação entre tramas, existindo apenas um codificador normalizado, pela OTAN em 1995 [5], com um débito binário de 800 bit/s. Abaixo deste débito binário, ainda é necessário um grande esforço de investigação de modo a produzir fala de boa qualidade, sendo a codificação fonética, por tirar partido das características da própria linguagem, uma das técnicas que demonstra maior potencial.
O sinal de fala é composto por um conjunto de sons articulados, produzidos pelo aparelho fonador por variação da pressão do ar. Estas variações são detectadas pelo ouvido e transmitidas ao cérebro, que as interpreta. Os sons, desprovidos de sentido quando isolados, distinguem pela sua associação elementos constituintes de níveis superiores da linguagem: sílabas, palavras e frases. O conteúdo de uma mensagem falada, no sentido estrito, não é mais do que a sequência fonética de sons, praticamente não se distinguindo em termos de informação de uma mensagem escrita, como nas legendas de um filme. O número de fonemas em cada língua ronda normalmente os 20 a 60, pelo que estes são codificados com um máximo de 6 bits. A uma média de 15 fonemas por segundo, resulta um débito binário médio de 90 bit/s para codificar a informação fonética, sendo este valor dependente do número de segmentos fonéticos produzido pelo orador de entrada. O valor obtido deve no entanto ser interpretado como um limite inferior, uma vez que, neste sentido estrito, não são levadas em consideração as características do orador nem a prosódia da fala natural que os codificadores devem tentar reproduzir. O emissor é então constituído por um reconhecedor fonético, que converte o sinal acústico numa sequência de segmentos fonéticos, e o receptor, constituído por um sintetizador fonético, converte a sequência fonética num sinal acústico. O esquema de blocos de um codificador deste tipo é apresentado na figura 1. A informação transmitida inclui a sequência de segmentos fonéticos reconhecidos e também a informação de carácter prosódico (duração dos segmentos fonéticos, energia e frequência fundamental).

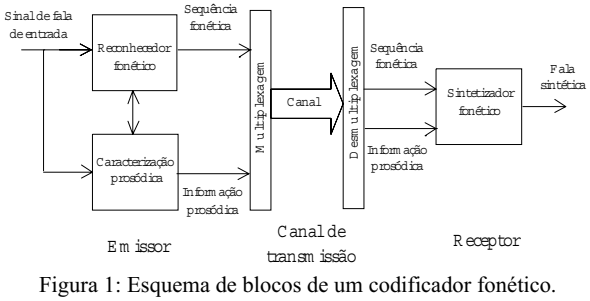

Figura 1: Esquema de blocos de um codificador fonético.

Os codificadores fonéticos tiram partido da representação do espectro, num livro de código acedido pelo índice do segmento fonético reconhecido. No entanto, se não forem

tomadas algumas precauções, o sinal sintetizado pode ter uma grande dependência com o(s) orador(es) que gerou o *corpus* com que o livro de código foi treinado. Este problema, relacionado com a "reconhecibilidade" do orador, ou seja, a capacidade de reconhecer um orador através da sua voz, implica, em termos de codificação fonética, a derivação e transmissão, para além da frequência fundamental, de parâmetros que caracterizem o orador e com que se possa reproduzir uma aproximação do mesmo tipo de voz.

As secções seguintes serão dedicadas à descrição dos blocos constituintes do codificador desenvolvido. Começa-se por descrever o codificador fonético ainda sem capacidades de adaptação ao orador. Segue-se a secção dedicada à descrição da metodologia de adaptação ao orador. A secção seguinte é dedicada aos testes perceptuais realizados para aferir o codificador e, finalmente, apresenta-se as principais conclusões e discute-se o trabalho futuro.

CODIFICADOR FONÉTICO

No codificador proposto, tal como noutros *vocoders* [12] baseados em predição linear, o transmissor efectua, trama a trama, a análise LPC (*Linear Predictive Coding*) e extrai a informação de carácter prosódico como a energia e a frequência fundamental. No codificador fonético desenvolvido, a informação do preditor linear e da energia são aplicados a um reconhecedor de fonemas baseado em modelos de Markov não-observáveis (HMM-Hidden Markov Models), de modo a segmentar o sinal e produzir um índice fonético. Utilizámos modelos com 3 estados e 3 misturas por estado, e um vector de entrada com 30 coeficientes (14 mel-cepstra, 14 delta mel-cepstra, energia e delta energia). Foram gerados 53 modelos, correspondentes a segmentos fonéticos distintos, treinados com o *corpus* de fala EUROM.1 [6], utilizando 8 dos 10 oradores do grupo dos *Poucos*, e os ficheiros correspondentes às *Passagens*. Estes sinais, que correspondem a cerca de 32 minutos de fala depois de removidos os silêncios, foram etiquetados manualmente. Para melhorar a taxa de reconhecimento, foram gerados modelos separadamente para oradores do sexo masculino e do sexo feminino e integrado um modelo de bigramas de segmentos fonéticos, que leva em consideração a grande correlação entra a sequência fonética. A taxa de reconhecimento obtida é de 67%, tendo o segmentador sido testado com 2 oradores, um de cada sexo, que não fizeram parte do conjunto de treino. Esta taxa pode ainda ser melhorada treinando modelos de segmentos fonéticos que levem em conta o contexto em que são produzidos (*trifones*), mas a utilização do contexto à direita pode ser proibitivo pelo consequente aumento do atraso.

O índice do segmento fonético reconhecido e a respectiva duração são então transmitidos, juntamente com o valor da frequência fundamental (vibração das cordas vocais) e da energia. No receptor, o índice fonético, juntamente com os índices dos segmentos anterior e posterior de modo a levar em conta o contexto, são utilizados para obter de um livro de código a informação da envolvente espectral, representada através de palavras de código compostas por uma matriz de coeficientes LSF (*Line Spectrum Frequencies*), com tantas colunas como o número de tramas do trifone armazenado e com tantas linhas como a ordem de predição linear (nº de coeficientes LSF por trama). O livro de código foi gerado com o mesmo conjunto de sinais de fala já utilizados para treinar os modelos HMM. Quando um trifone não se encontra presente no livro de código, este é substituído pelo mesmo segmento fonético com um contexto parecido.

O género do orador é, talvez, a primeira característica da fala que um ouvinte humano é capaz de identificar. A dimensão do tracto vocal influencia a localização dos formantes e a espessura e comprimento das cordas vocais influenciam a frequência fundamental, sendo as causas principais desta distinção, que divide o espaço dos sinais de fala. Assim, os oradores do sexo feminino têm, tipicamente, formantes e frequências fundamentais mais elevados que os oradores do sexo masculino. Uma pré-identificação automática do género pode, por isso, assistir com sucesso a algumas aplicações do processamento de fala. No contexto da codificação fonética, foram gerados dois livros de código com informação espectral, um a partir de oradores do sexo masculino e outro a partir de oradores do sexo feminino e utilizada a identificação automática do género para escolher o livro de código mais adequado às características do orador de entrada.

De modo a não transmitir mais informação, foi desenvolvido um detector de género baseado na frequência fundamental. As frequências de ocorrências dos valores médios da frequência fundamental em cada segmento fonético, para o conjunto de treino já utilizado na geração dos livros de código, são apresentadas na figura 2, separadamente para cada género, verificando-se uma grande separação entre as duas classes. Utilizando a regra de decisão de Bayes, obtém-se um valor para a frequência de separação entre classes de 157 Hz, sendo a identificação de cada classe obtida de:

$$\begin{cases} \text{se } \overline{F}0 < 157\text{Hz} & \text{identifica - se género masculino} \\ \text{se } \overline{F}0 \geq 157\text{Hz} & \text{identifica - se género feminino} \end{cases} . \quad (1)$$

Para os segmentos não vozeados, é mantida a última decisão tomada. O identificador de género foi testado com 48 oradores, 24 de cada género, do grupo dos *Muitos Oradores* do *corpus* EUROM.1, correspondendo a 18 minutos de fala, após removidos os silêncios. A taxa de identificação obtida é de 91%. De modo a tornar mais robusta a decisão, optou-se por utilizar a decisão parcial nos N últimos segmentos fonéticos vozeados e uma regra de decisão por maioria na decisão global. Desta forma, é filtrado o aparecimento esporádico de (N-1)/2 segmentos fonéticos, com uma decisão parcial incorrecta. A taxa de identificação obtida para um valor de N=7 é de 95%. Note-se que, sendo estes erros esporádicos ao longo de uma frase, podem não afectar significativamente a qualidade final dos sinais sintetizados, nomeadamente em segmentos correspondentes a consoantes, em que as características dos oradores são perceptualmente menos importantes.

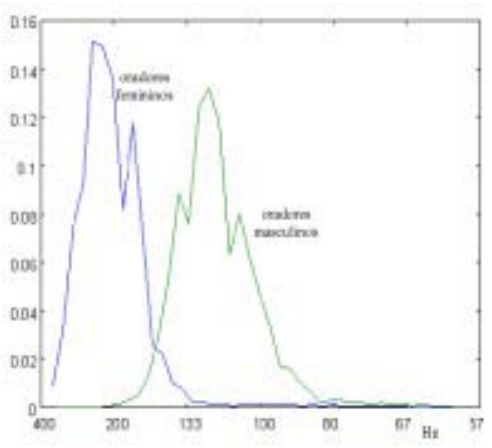

Figura 2: Frequências de ocorrência do valor médio da frequência fundamental, calculadas por cada segmento fonético.

A palavra de código correspondente ao género identificado é então modificada temporalmente para se ajustar à duração recebida e os coeficientes LSF, depois de convertidos para coeficientes de predição linear, são utilizados, conjuntamente com a informação da energia e da frequência fundamental, na reconstrução do sinal. Foram implementados dois tipos de sintetizadores. O primeiro é baseado no modelos dos *vocoders* LPC [12] e o segundo é baseado no modelo harmónico [3], derivando as amplitudes das harmónicas a partir informação da envolvente espectral.

O conjunto de parâmetros a ser transmitidos compreende o índice do segmento fonético reconhecido, a sua duração e a informação da energia, da frequência fundamental (F0) e da decisão de vozeamento (*V/UV*). Os codificadores foram treinados com o mesmo conjunto de sinais que serviram para gerar o livro de código e testados com os 48 oradores que serviram no teste de identificação do género. O número médio de segmentos fonéticos encontrado é de 14,3 segmentos por segundo, sendo os métodos de codificação detalhados em [7] e resumidos seguidamente:

- **Índice fonético** - Os 53 segmentos fonéticos são codificados com um código de Huffman, sendo obtido, no *corpus* de teste, um débito binário médio de 76 bit/s. Utilizando um modelo de bigramas de segmentos fonéticos obtém-se um débito de 59 bit/s.
- **Duração dos segmentos** - A duração dos segmentos fonéticos é codificada com um código de Huffman, sendo obtido um débito binário médio de 56 bit/s.
- **Energia** - A energia é transmitida por trama de 22,5 ms, sendo utilizado um quantificador LBG de 32 níveis e o índice codificado com um código de Huffman, resultando num débito de 200 bit/s. Se esta informação for transmitida uma vez por segmento fonético resulta um débito binário médio de 71 bit/s.
- **Frequência fundamental e *V/UV*** - A informação da frequência fundamental e da decisão de vozeamento é também transmitida por trama de 22,5 ms, resultando num débito de 311 bit/s. Utilizando um código diferencial resulta um débito médio de 128 bit/s e transmitindo este parâmetro uma vez por segmento, resulta um débito médio de 84 bit/s.

O débito binário médio mais baixo é atingido pela transmissão de todos os parâmetros uma vez por segmento fonético, correspondendo a 270 bit/s, e o máximo débito binário a 643 bit/s. Entre estes dois extremos, podem ser considerados diversos esquemas de codificação, de que resulta um compromisso entre o débito binário, por um lado e, por outro lado, a qualidade e robustez na presença de erros de canal.

ADAPTAÇÃO AO ORADOR

A estratégia de adaptação que seguimos é baseada na minimização de uma distância espectral D entre o segmento de entrada X e a respectiva palavra de código Y, após o ajustamento temporal, transformando esta última numa matriz adaptada Y_{adap},

$$D = distância\left(X, Y_{adap}\right). \tag{2}$$

Como critério de distância, utilizámos a soma dos erros quadráticos D_j entre cada j-ésimo vector linha de `X`, correspondente à trajectória temporal do j-ésimo coeficiente LSF e o respectivo vector linha de Y_{adap}, sendo a transformação produzida a partir de:

$$LSFadp_{ji} = \alpha_j + \beta_j \, LSFmc_{ji}, \tag{3}$$

em que *LSFmc* representam os valores dos coeficientes LSF da matriz de código. A distância D_j para cada coeficiente virá:

$$D_j = \sum_{i=1}^{L_R} \left(LSFori_{ji} - \left(\alpha_j + \beta_j LSFmc_{ji}\right)\right)^2. \tag{4}$$

Minimizando D_j em relação a α_j e β_j virá, depois de alguma manipulação matemática:

$$\beta_j = \frac{\sum_{i=1}^{L_R} \left(LSFori_{ji} LSFmc_{ji}\right) - \overline{LSFori}_j \, \overline{LSFmc_j}}{\sum_{i=1}^{L_R} \left(LSFmc_{ji}\right)^2 - \left(\overline{LSFmc_j}\right)^2}, \tag{6}$$

$$\alpha_j = \overline{LSFori}_j - \beta_j \, \overline{LSFmc}_j. \tag{7}$$

Os valores adaptados são monitorizados para evitar a instabilidade do filtro de predição linear, forçando a uma distância mínima de 80 Hz entre coeficientes LSF consecutivos. Para além de uma melhor reconhecibilidade do orador, a adaptação deverá, por princípio, melhorar a qualidade do sinal sintetizado. As razões para esta melhoria devem-se ao facto da adaptação ser baseada na minimização de uma distância espectral, pelo que: (1) se minimizam os efeitos dos erros na classificação dos segmentos e na definição das fronteiras dos segmentos; (2) se adequam as palavras de código à frequência fundamental do orador de entrada; (3) se minimizam os saltos dos formantes entre segmentos fonéticos consecutivos que sejam adaptados.

Os coeficientes de adaptação α_j e β_j, j=1,...,p devem ser calculados no emissor e transmitidos para o receptor, para aí ser efectuada a adaptação. O número de coeficientes de adaptação, 2xp por segmento correspondente às vogais ou glides, é no entanto demasiado elevado para, em conjunto com os outros parâmetros transmitidos, manter o débito binário aceitável num codificador deste tipo. Perceptualmente, a característica mais importante a manter é a média de cada coeficiente, dado esta determinar as frequências e larguras de banda médias dos formantes. O valor de β_j controla as variações em relação ao valor médio dos formantes, sendo perceptualmente menos importante, pelo que foi colocado a 1, ou seja, mantém-se as trajectórias representadas na palavra de código. Com esta aproximação sub-óptima, os coeficientes de adaptação α_j representam a diferença entre as médias de cada coeficiente LSF no segmento original e as médias na matriz correspondente da palavra de código, diferença essa que designaremos de *informação específica do orador*. Para quantificar esta informação, foi treinado um quantificador LBG, com 4 níveis de quantificação para cada coeficiente, sendo necessário transmitir 2xp bits por vogal ou glide. O número médio de vogais ou glides por segundo no *corpus* de teste é de 5,8, o que corresponde a um acréscimo no débito binário de 116 bit/s.

Os coeficientes LSF de ordem mais elevada, relacionados com os formantes de frequências mais altas, têm uma importância perceptual menos relevante. Seguindo este princípio, a adaptação ao orador agora proposta, pode adaptar os primeiros coeficientes LSF, utilizando os coeficientes de ordem mais elevada sem efectuar a adaptação. A adaptação nos 7 primeiros coeficientes mostrou, em testes informais, ser um bom compromisso entre a diminuição do débito binário (30% de redução) e a preservação da capacidade de adaptação ao orador.

De modo a reduzir o débito binário imposto pela adaptação, pode-se explorar a correlação intra-orador, tirando partido da informação específica do orador, transmitida anteriormente. Uma forma simples de atingir este objectivo é transmitindo a informação específica apenas a primeira vez que determinado segmento ocorre e reutilizando estes valores nas próximas ocorrências do mesmo segmento, ou, de um modo equivalente, substituindo a palavra de código pela sua versão adaptada. Esta metodologia, embora simples, tem dois inconvenientes: primeiro, a informação intra-orador não é suficientemente bem modelada utilizando apenas uma instância de adaptação por cada segmento fonético, pelo que deve ser implementado algum método de média para melhorar a estimação e, segundo, deve ser introduzido um procedimento de inicialização, de modo a ser aceite um novo orador no sistema. Ambos os problemas são resolvidos através da adaptação *incremental* das palavras de código, descrita com pormenor em [8]. Esta metodologia altera a média de cada coeficiente LSF através de:

$$\overline{LSFmc_j} = (1-\mu)\overline{LSFmc_j} + \mu\,\overline{LSFori_j}, \qquad (8)$$

em que μ controla a velocidade de adaptação. A redução do débito binário é conseguida transmitindo a informação específica do orador, α_j, j=1,...,p, quando a distância espectral entre os valores médios dos coeficientes LSF do segmento de entrada e os valores médios dos coeficientes LSF da palavra de código, exceder um limiar pré-definido *Th*. Este procedimento requer mais um bit por vogal ou glide para codificar a presença ou ausência de adaptação. Para além da condição anterior, é ainda testada a convergência, ou seja, se a distância entre os valores médios do segmento de entrada e os da palavra de código diminui após a adaptação.

Quando a informação específica do orador é transmitida, a adaptação ao orador processa-se da mesma forma aquando da adaptação total, seguindo-se a convergência da palavra de código em direcção ao orador de entrada. Quanto maior for o limiar de comparação Th, menos frequentemente as palavras de código serão actualizadas, com o correspondente ganho em débito binário, mas também com a degradação da reconhecibilidade do orador. Para pequenos valores de *Th*, as palavras de código serão actualizadas com maior frequência e, tanto o débito binário como a reconhecibilidade do orador, se aproximarão dos produzidos pela versão do codificador com adaptação total ao orador.

Os valores de μ e de *Th* foram ajustados de modo a produzir, em 30 segundos de fala, uma redução de 50% na transmissão da informação específica do orador, sendo a reconhecibilidade do orador julgada, em testes informais, muito perto da produzida com adaptação total.

AVALIAÇÃO DO CODIFICADOR

Os codificadores de fala que funcionam a débitos binários muito baixos, apenas tentam preservar as propriedades mais relevantes do sinal original, sem o tentar reproduzir de um modo detalhado. São os casos da preservação do espectro de potência de curta duração e da estrutura fina do espectro, perceptualmente mais importantes que a definição exacta da fase. Para este tipo de codificadores, embora existam medidas de qualidade objectivas, baseadas no espectro de curta duração, os métodos subjectivos ainda são dos mais utilizados. Estes métodos utilizam ouvintes, normalmente não treinados para simular os utilizadores típicos dos sistemas de codificação, e podem avaliar a qualidade em termos da distorção introduzida, da naturalidade, da inteligibilidade e da reconhecibilidade do orador.

Teste de inteligibilidade:

Para testar a inteligibilidade do codificador fonético, implementámos um teste equivalente ao proposto pelo projecto SAM [10], a partir dos *Logátomos* do *corpus* EUROM.1. A estrutura das palavras utilizadas é do tipo /C-V-C-V/ (C - Consoante; V- Vogal) em que se altera apenas um dos fonemas centrais. O teste é de escolha aberta, tendo os ouvintes que distinguir o fonema em teste, de entre o conjunto das consoantes ou das vogais. Avaliámos o codificador fonético com segmentação automática e em que todos os parâmetros estão

quantificados, nas versões com e sem adaptação ao orador. Para avaliar a degradação na inteligibilidade provocada por erros na classificação dos segmentos fonéticos, avaliámos também o codificador fonético com e sem adaptação ao orador, mas com alinhamento fonético automático (conhecendo a sequência fonética a alinhar). Avaliámos ainda, como referência, o codificador da norma FS-1015 [12], para a qual obtivemos o código em [1] e também a fala natural.

Produzimos dois testes a que correspondem dois conjuntos disjuntos de segmentos fonéticos. No primeiro teste (*teste1*), avaliámos a inteligibilidade de 21 consoantes e glides, apresentadas na tabela 1, recorrendo ao alfabético fonético SAMPA (SAM Phonetic Alphabet) [6], no contexto /la-*consoante (ou glide) em teste*-a/.

Tabela 1: Consoantes e glides em teste (*teste1*).

p	t	k	b	d	g	f	s	S	v	z	Z
R	r	l	L	m	n	J	j	w			

No segundo teste avaliámos a inteligibilidade de 13 vogais (*teste2*), apresentadas na tabela 2, no contexto /t-*vogal em teste*-tA/. Repare-se que a maioria das palavras geradas não têm sentido, limitando a antecipação da compreensão da palavra através do contexto.

Tabela 2: Vogais do teste (*teste2*).

a	E	i	O	u	6	e	o	6~	e~	i~	o~	u~

A avaliação foi efectuada com sinais gravados por um orador de cada sexo. Foram utilizados 10 ouvintes, não treinados, para avaliar a inteligibilidade do sinal de fala dos diversos codificadores. O teste está completamente automatizado, sendo a ordem de apresentação aleatória. As taxas de inteligibilidade para os codificadores em teste são apresentadas na tabela 3.

Tabela 3: Taxas de inteligibilidade.

Codificador	*teste1*	*teste2*
Fala natural	94	99
Codificador de referência 2400 bit/s	82	95
Adaptação e alinhamento automático	50	77
Adaptação e segmentação automática	47	74
Adaptação e alinhamento automático	52	58
Adaptação e segmentação automática	36	54

Os valores obtidos para as taxas de inteligibilidade nas consoantes são praticamente idênticas nos dois codificadores fonéticos com alinhamento automático, independentemente da adaptação ao orador, uma vez que não é nestes segmentos que esta é efectuada. Contudo, para as vogais, em que é efectuada a adaptação, a taxa de inteligibilidade sobe 18 pontos percentuais com a inclusão da adaptação, revelando uma melhor qualidade das palavras de código. De facto, sendo a adaptação ao orador baseada num critério de minimização de uma distância espectral, estas estão mais adequadas à frequência fundamental do sinal de entrada e atenuam os erros de segmentação. Em relação a estes últimos, os benefícios podem ser comprovados pelos valores próximos das taxas de inteligibilidade, dos codificadores com e sem alinhamento automático, mas com adaptação ao orador. Nas consoantes, em que não é efectuada adaptação, o codificador com adaptação ao orador e segmentação automática tem praticamente o mesmo desempenho do que com alinhamento automático, beneficiando, possivelmente, de um contexto com melhor qualidade.

O codificador sem adaptação ao orador e com segmentação automática obtém os piores resultados, quer nas consoantes quer nas vogais, mostrando mais uma vez a grande importância da adaptação ao orador, embora à custa de um aumento do débito binário.

Teste de reconhecibilidade do orador:

De modo a validar a metodologia de adaptação ao orador, implementámos um teste de reconhecibilidade do orador adaptado da proposta de Schmidt-Nielson e Brock [9]. O teste baseia-se no julgamento de pares de frases como sendo do mesmo orador ou de oradores diferentes, utilizando 10 oradores de cada sexo. Foram efectuadas duas experiências. Na primeira experiência, que designaremos por NP-P, a primeira frase do par não é processada, enquanto que a segunda é processada pelo codificador em teste. Esta experiência testa a capacidade do codificador para preservar as características originais do orador. Na segunda experiência, que designaremos por P-P, ambas as frases do par são processadas pelo codificador em teste. Esta experiência testa a capacidade do codificador para captar características capazes de se distinguirem entre diferentes oradores, mesmo que a voz produzida seja alterada em relação à voz original.

Em ambas as experiências, um conjunto de ouvintes testam os pares de frases produzidas pelo mesmo orador ou por oradores diferentes. Após ouvir cada par, o ouvinte é inquirido sobre a sua opinião, quanto a ser ou não o mesmo orador.

Utilizámos 20 oradores, sendo 10 de cada sexo, divididos em grupos de 5. Os oradores têm predominantemente uma pronúncia de Lisboa e idades compreendidas entre os 20 e os 43 anos. Para a experiência P-P, foram garantidas todas as combinações entre oradores do mesmo bloco e o mesmo número de frases com o mesmo orador e com oradores diferentes (4x10). Designámos o conjunto dos pares com o mesmo orador por *MESMOS* e o conjunto dos pares com oradores diferentes por *DIFERENTES*. Para a experiência NP-P, foram garantidas todas as combinações entre oradores do mesmo bloco, o que duplica (4x20) o número de pares no conjunto dos *DIFERENTES*, uma vez que não existe comutatividade entre a ordem de apresentação dos oradores. Foi, no entanto, mantido o número de pares no conjunto dos *MESMOS* (4x10).

Foram utilizados 8 ouvintes, não treinados, para efectuar o teste de reconhecibilidade do orador. Os resultados do teste são mostrados na tabela 4. Existe uma maior facilidade de distinguir entre locuções de oradores diferentes do que do mesmo orador. Consequentemente, a percentagem global de respostas correctas é adulterada para o teste NP-P, pois existem mais locuções de oradores diferentes do que do mesmo orador, provocando um aumento global nas respostas correctas. Esta adulteração pode ser compensada, utilizando, como reportado em [9],

o valor da área debaixo da curva ROC *(Receiver Operating Characteristic*). Neste caso particular, a compensação resume-se a calcular separadamente as percentagens de respostas correctas para o conjunto dos *MESMOS* e para o conjunto dos *DIFERENTES*, sendo a percentagem global calculada como a média das percentagens individuais.

Tabela 4: Taxas de reconhecibilidade do orador, nos conjuntos dos *MES*(*MOS*), dos *DIF*(*ERENTES*) e respectiva (*MÉD*)*IA*.

	NP-P			PP		
Codificador	MÉD	MES	DIF	MÉD	MES	DIF
Fala natural	86	81	91	----	----	----
Referência	69	52	85	80	73	88
c/ adaptação	66	54	78	74	70	78
s/ adaptação	59	35	83	66	67	65

De realçar o valor extremamente baixo da taxa de reconhecibilidade, definida como a percentagem de respostas correctas, na experiência NP-P para o conjunto dos *MESMOS*, para o codificador fonético sem adaptação ao orador. Este resultado (35%), esperado, deve-se ao facto de que o codificador tem sempre uma voz típica, baseada nos oradores que contribuíram para a criação do livro de código, diferenciando-se os oradores exclusivamente por aspectos prosódicos, tais como o valor médio da frequência fundamental. Já para o conjunto dos *DIFERENTES*, o número de respostas correctas é elevado, pois a voz típica do codificador não é parecida com a voz de nenhum dos oradores de entrada. Para a experiência P-P, as características do sinal sintetizado, no que respeita ao orador, são sempre muito parecidas, fazendo aumentar o número de respostas classificando como o mesmo orador, quer realmente o seja ou não. A melhoria introduzida pelo bloco de adaptação ao orador pode ser verificada, através do aumento da taxa de reconhecibilidade para o conjunto dos *MESMOS*, na experiência NP-P, em relação ao codificador sem adaptação ao orador. Este valor, que passa de 35% para 54%, situa-se mesmo 2% acima do conseguido pelo codificador de referência, embora este tenha um débito binário 4 vezes mais elevado.

Na experiência P-P, o codificador com adaptação ao orador tem uma percentagem de respostas correctas que se situa, como esperado, entre o codificador sem adaptação e o de referência. Este resultado, realça mais uma vez as melhorias conseguidas pelo bloco de adaptação ao orador, mostrando.

Após o teste, os ouvintes são informalmente inquiridos sobre alguns aspectos genéricos da qualidade dos sinais de fala. Das respostas obtidas, destaca-se que a identificação do género do orador nunca foi um problema, nem mesmo para o codificador sem adaptação ao orador, validando desta forma a estratégia de utilização de livros de código, diferenciados por género, e a introdução do respectivo módulo de identificação automática.

CONCLUSÕES E DESENVOLVIMENTOS FUTUROS

Este artigo apresentou um codificador fonético com capacidade de adaptação ao orador e um conjunto de estratégias de codificação dos parâmetros a transmitir que possibilita compromissos entre o débito binário e a qualidade dos sinais sintetizados. Da análise dos resultados dos testes de reconhecibilidade do orador, verifica-se uma melhoria da qualidade com a inclusão da adaptação ao orador e, dado que esta se baseia na minimização de uma distância espectral é melhorada a qualidade geral e a inteligibilidade dos sinais sintetizados.

A geração de novos livros de códigos que nomeadamente minimizem as distorções prosódicas, será um passo importante a ser dado no futuro, com o objectivo de melhorar a inteligibilidade do codificador fonético. Uma das possibilidades a testar é a geração de livros de código dependentes de gamas da frequência fundamental e da dimensão dos segmentos fonéticos.

REFERÊNCIAS

[1] J. A. Figerhut,http://www.arl.wustl.edu/~jaf/lpc/lpc10-1.5.tar.gz

[2] W. B. Kleijn, J. Haagan, “Waveform Interpolation", "Speech Coding and Synthesis”, *Cap. 5, Elsevier, W. B. Kleijn, K. K. Paliwal Editors*, 1995.

[3] R. J. McAulay, T. Champion, “Improved Interoperable 2.4 kb/s LPC Using Sinusoidal Transform Code Techniques”, *Proc. of the Int. Conf. Acoust., Speech and Signal Processing*, pp.641-643, 1990.

[4] A. V. McCree, K. Truong, E. B. George, T. P. Barnwell, V. Viswanathan, “A 2.4 KBIT/S MELP Coder Candidate for the New U.S. Federal Standard”, *Proc. of the Int. Conf. Acoust., Speech and Signal Processing*, pp.200-203, 1996.

[5] B. Mouy, P. de La Noue, G. Goudezeune, “NATO STANAG 4479: A Standard for an 800 BPS Vocoder and Channel Coding in HF-ECCM System”, *Proc. of the Int. Conf. Acoust., Speech and Signal Processing*, pp.480-483, 1995.

[6] C. M. Ribeiro, I. M. Trancoso, M. C. Viana, “EUROM.1 Portuguese Database”, ESPRIT 6819 SAM-A - Speech Technology Assessment in Multilingual Applications, Report D6, 1993.

[7] C. M. Ribeiro, I. M. Trancoso, “Phonetic Vocoding”, *Actas da II Conf. de Telecomunicações*, 1999.

[8] C. M. Ribeiro, I. M. Trancoso, “Speaker Adaptation in a Phonetic Vocoding Environment”, *Proc. of the 1999 IEEE Workshop on Speech Coding*, pp.64-66, 1999.

[9] A. Schmidt-Nielsen, D. P. Brock, “Speaker Recognizability Testing For Voice Coders”, *Proc. of the Int. Conf. Acoust., Speech and Signal Processing*, pp.1149-1152, 1996.

[10] “SAM Final Report", *Section II - Speech Output Assessment*, 1992.

[11] J. Slifka, T. R. Andersom, “Speaker Modification with LPC pole analysis”, *Proc. of the Int. Conf. Acoust., Speech and Signal Processing,* pp.645-647, 1995.

[12] T. E. Tremain, “The Government Standard Linear Predictive Coding Algorithm: LPC-10”, *Speech Technology,* Vol.1 nº2, pp.40, Abril de 1982.