Avaliação Subjetiva da Qualidade de Transcodificações Digitais de Voz

Ricardo Honório Guedes de Sousa e Abraham Alcaim

DE/3/IME-Rio, 22290-270 Rio de Janeiro - RJ, Brasil

CETUC/PUC-Rio, 22453-900 Rio de Janeiro - RJ, Brasil

e-mails: rhgsousa@epq.ime.eb.br e alcaim@cetuc.puc-rio.br

***Resumo*— Este artigo avalia a qualidade das transcodificações digitais de voz em uma rede de comunicações que utiliza codecs de tecnologia atual, oriundos de diferentes fabricantes. Estes codecs utilizam os seguintes algoritmos [1]: CELP (Code Excited Linear Prediction) a 16 kbit/s, LD-CELP (Low Delay - CELP) a 16 kbit/s e CS-ACELP (Conjugated Structure - Algebraic CELP) a 8 kbit/s, este último com capacidade, ou não, de utilizar supressão de silêncio. Um critério de avaliação subjetiva – a nota de opinião média (NOM) – é utilizado para medir o desempenho de quinze conexões desses codificadores e cinco Condições de Referência (MNRU - "Modulated Noise Reference Unit").**

I. Introdução

Nos últimos anos, a digitalização das redes de comunicações tem passado por um processo de progressiva redução da taxa de bits. A expansão das redes visando apenas a economia de recursos, sem a utilização de um critério que prime pela qualidade das comunicações, motiva a aquisição de novos equipamentos — inclusive a 8 kbit/s — e a utilização de conexões que, segundo estudos anteriores [2], podem apresentar uma qualidade de voz abaixo da razoável. Geralmente esse procedimento resulta na falta de homogeneidade na qualidade das ligações.

Em face desse panorama, este trabalho tem por objetivo identificar conexões de voz que possam causar problemas mais sérios de qualidade nas ligações empregadas em uma rede de comunicações operacional que utiliza codecs a diferentes taxas e de fabricantes distintos. Para isso, foram implementados e executados testes subjetivos formais de qualidade de voz.

São estabelecidas na Seção II, a partir de uma análise da configuração física da rede, as configuracões avaliadas e a metodologia empregada para determinar a qualidade dos sinais de voz decodificados. Os preparativos e a montagem dos testes subjetivos são descritos na Seção III. Os resultados dos testes e sua análise são apresentadas na Seção IV. Finalmente, a última Seção dedica-se às conclusões finais.

II. Metodologia, Codecs e Configurações

Para qualquer que seja o sistema de codificação de sinais de voz, o teste de qualidade definitivo é o mecanismo de percepção humana. Testes que levam em conta tal mecanismo pertencem à classe mais importante de métodos de avaliação da qualidade de voz, a dos chamados métodos subjetivos. Esses testes devem, portanto, sempre fazer parte do procedimento de projeto e avaliação de um determinado sistema.

A distorção observada em codificadores operando a taxas de bits relativamente baixas é fortemente não-linear, quando os mesmos são conectados em uma configuração de rede. Certamente não há estudo que indique uma medida objetiva que seja adequada para este fim. Por esta razão, a caracterização do desempenho das conexões a serem avaliadas neste projeto será sempre feita com base em medidas subjetivas de qualidade de voz.

Há diversas formas de testes subjetivos. Em geral, os testes subjetivos (de escuta) mais usuais são *testes informais de comparação de pares*. Nestes testes, os avaliadores externam suas opiniões através de uma escala ternária de opiniões: a qualidade do primeiro é melhor ou a qualidade do segundo é melhor ou ambos apresentam qualidades equivalentes. Entretanto, quando o número de condições a serem avaliadas é elevado e sofrem degradações de maneiras distintas, os mais indicados são os testes formais, baseados na classificação absoluta da qualidade do material de voz processado. Estes testes são denominados de *testes de categorias* ou ACR ("Assessment Category Rating"), e uma das medidas mais usadas é a *Nota de Opinião Média* ou NOM (em inglês, "Mean Opinion Score" ou MOS). Para determinação da NOM, normalmente são incluídas condições de referência que permitam comparar avaliações resultantes de outros testes. Uma classe de condições de referência muito usada é a gerada pelo MNRU ("Modulated Noise Reference Unit"), objeto da Recomendação P.81 do CCITT (atual ITU) [3]. Uma referência MNRU é obtida adicionando ao sinal de voz original um ruído branco cuja amplitude é proporcional à amplitude do sinal de voz. As referências são identificadas pelo valor da razão sinal-ruído, referida como Q. Diz-se que um sinal de voz tem *qualidade equivalente* Q quando a NOM a ele atribuída é igual ao da referência MNRU com razão sinal-ruído Q.

Para obtenção da NOM, um conjunto de locuções (frases) de pequena duração são processadas pelos vários codificadores, ou conexões de codificadores que se deseja avaliar, e pelas MNRUs. O teste ACR para

TABELA I
ESCALA DE QUALIDADE SUBJETIVA

Nota	Qualidade	Nível de Degradação
5	Excelente	Imperceptível
4	Boa	Pouco perceptível (não incômoda)
3	Razoável	Perceptível (levemente incômoda)
2	Ruim	Incômoda (não intolerável)
1	Péssima	Muito incômoda (intolerável)

TABELA II
SIGLA DOS CODIFICADORES

Sigla	Codificador
16 CELP	CELP a 16 kbit/s
16 LDCELP	LD-CELP a 16 kbit/s
8 CSACELP	CS-ACELP a 8 kbit/s
8 CSACELP (SS)	8 CSACELP c/ supressor de silêncio

determinação da NOM é aplicado a cada uma de um grupo de pessoas, que ouve as frases processadas e julga a qualidade dos sinais produzidos pelos codificadores, atribuindo uma das cinco notas da escala mostrada na Tabela I.

Os codificadores avaliados nos testes subjetivos e suas respectivas siglas estão listados na Tabela II. Foram avaliadas um total de 20 configurações, sendo 15 de diferentes conexões de codecs e 5 MNRUs, mostradas na Tabela III.

A nota de opinião média é igual à média aritmética dos números indicados pelos ouvintes. É interessante que os resultados incluam, também, o intervalo de confiança associado a cada cálculo da NOM: [NOM-δ(95%) , NOM+δ(95%)], que corresponde ao intervalo onde se situa a média populacional com 95% de confiança.

III. PREPARATIVOS E MONTAGEM DOS TESTES SUBJETIVOS

Nesta Seção serão apresentados os preparativos e os procedimentos adotados para a montagem dos testes subjetivos. Os sinais de voz usados na realização do teste foram produzidos por 4 locutores: dois do sexo masculino (M1 e M2) e dois do sexo feminino (F1 e F2). Cada locutor pronunciou em um microfone um par de frases, extraídas de [4], com uma pausa entre elas de aproximadamente 1 segundo. Os sinais foram digitalizados por uma placa de som, originando 4 locuções (L1, L2, L3 e L4). A Tabela IV apresenta a lista de frases pronunciadas por cada locutor.

Posteriormente as 4 locuções (L1, L2, L3 e L4) foram processadas por todas as 20 *configurações* (conexões de codecs e MNRU's) apresentadas na Tabela III, resultando em 80 condições a serem avaliadas. Cada condição é identificada por um código do tipo LiCj, i=1,2,3,4 e j=1,...,20, onde Li indica a locução i (ver Tabela IV) e Cj identifica a configuração j (ver Tabela III). Nota-se, então, que cada amostra ou condição representa uma locução (um par de frases pronunciados por um locutor) submetida a uma configuração.

TABELA III
CONFIGURAÇÕES DE CODECS INCLUÍDOS NO TESTE SUBJETIVO

Config.	Codecs
1	8 CSACELP → 16 CELP
2	8 CSACELP (SS) → 16 CELP
3	8 CSACELP → 16 LDCELP
4	8 CSACELP (SS) → 16 LDCELP
5	16 LDCELP → 8 CSACELP (SS)
6	8 CSACELP → 16 CELP → 16 CELP
7	8 CSACELP (SS) → 16 CELP → 16 CELP
8	16 CELP → 8 CSACELP → 16 CELP
9	16 CELP → 8 CSACELP (SS) → 16 CELP
10	16 CELP → 8 CSACELP → 16 LDCELP
11	16 CELP → 8 CSACELP (SS) → 16 LDCELP
12	16 LDCELP → 8 CSACELP → 16 CELP
13	16 LDCELP → 8 CSACELP → 16 LDCELP
14	16 LDCELP → 8 CSACELP (SS) → 16 LDCELP
15	16 CELP → 8 CSACELP → 16 CELP → 16 LDCELP
16	MNRU: Q = 10 dB
17	MNRU: Q = 15 dB
18	MNRU: Q = 20 dB
19	MNRU: Q = 25 dB
20	MNRU: Q = 30 dB

TABELA IV
LISTA DE FRASES USADAS NO TESTE SUBJETIVO

Locução	Frases
L1/M1	O jogo será transmitido bem tarde. É possível que ele já esteja fora de perigo.
L2/M2	Esse empreendimento será de enorme sucesso. As feiras livres não funcionam amanhã.
L3/F1	O vão entre o trem e a plataforma é muito grande. Infelizmente não comparecí ao encontro.
L4/F2	O sinal emitido é captado por receptores. A mensalidade aumentou mais que a inflação.

O resultado de um teste subjetivo de opinião é influenciado pela ordem de apresentação das condições e pelo estado do avaliador. É certamente importante que o teste não cause fadiga ao avaliador. É também importante que a ordem de apresentação seja tal que as avaliações possam ser consideradas independentes umas das outras. Isso pode ser alcançado realizando uma partição do conjunto de amostras com base em um *quadrado greco-latino* (uma matriz cujos elementos são pares ordenados com componentes que aparecem uma vez em cada coluna e uma vez em cada linha), onde cada avaliador escuta um único sub-conjunto dessa partição (uma fila da matriz) [5]. Uma variação dessa estratégia [6], adotada no presente trabalho, será descrita a seguir.

O conjunto de 80 amostras de voz foi dividido em 4 sub-conjuntos de 20 amostras, denominados: A, B, C e D. Esta divisão obedeceu os seguintes critérios:

- cada sub-conjunto contém todas as 20 configurações a serem avaliadas, em ordem de apresentação aleatória;
- uma dada configuração aparece em posições diferentes nos quatro sub-conjuntos;
- cada locução aparece 20 vezes, segundo uma distribuição aleatória e balanceada, em que não foi

TABELA V
DIVISÃO DAS AMOSTRAS DE VOZ EM SUB-CONJUNTOS
Li: Locução i (i=1,...,4),
Cj: Configuração j (j=1,...,20)

A	B	C	D
L4C16	L2C20	L1C6	L3C19
L2C6	L1C10	L3C16	L4C18
L4C17	L3C6	L1C18	L2C14
L1C15	L4C5	L2C8	L3C11
L2C18	L3C17	L4C3	L1C12
L1C1	L2C12	L3C20	L4C4
L3C13	L1C4	L4C19	L2C16
L2C19	L4C1	L3C4	L1C13
L4C10	L1C16	L2C1	L3C9
L3C8	L4C15	L1C11	L2C3
L1C20	L2C9	L3C10	L4C6
L4C7	L3C3	L1C2	L2C10
L1C5	L4C11	L2C15	L3C1
L2C4	L3C18	L4C14	L1C8
L3C12	L1C19	L2C17	L4C2
L1C3	L3C7	L4C12	L2C5
L4C9	L2C13	L3C5	L1C17
L3C2	L1C14	L2C7	L4C20
L2C11	L4C8	L1C9	L3C15
L3C14	L2C2	L4C13	L1C7

permitido que duas posições vizinhas em quaisquer dos sub-conjuntos fossem ocupadas pela mesma locução;

- os 4 sub-conjuntos contêm 4 diferentes locuções de uma mesma configuração em ordem aleatória.

Os quatro sub-conjuntos estão listados na Tabela V. Para tornar as avaliações mais independentes umas das outras seria interessante que cada avaliador avaliasse somente um dos subconjuntos. Contudo, esta tarefa requer um grande número de avaliadores e, por isso, optou-se por submeter um dos seguintes pares de sub-conjuntos a cada avaliador: AB, AC, AD, BC, BD, CD, BA, CA, DA, CB, DB e DC. Como foram utilizados 24 avaliadores (2 por cada par de sub-conjuntos), cada condição recebeu 48 notas. O número total de notas obtidas após a aplicação do teste foi, então, de: (20 condições) $\times$ (48 notas/condição) = 960 notas.

Após cada amostra de voz, foi inserida uma pausa de aproximadamente 5 segundos para que o avaliador pudesse dar a nota. Antes da aplicação do teste, cada avaliador foi submetido a um treinamento de modo a acostumá-lo com as diferentes qualidades de voz que ele encontraria e, também, ter uma experiência prática com o teste a que seria submetido.

IV. AVALIAÇÃO DOS RESULTADOS

Esta Seção apresenta inicialmente os resultados dos testes subjetivos de medida de qualidade de voz resultantes de transcodificações digitais na rede de comunicações, utilizando as configurações descritas na Seção anterior. Após a apresentação dos resultados é feita uma análise criteriosa da qualidade de voz resultante. Esta análise aponta as diretrizes e recomendações técnicas, relativas ao emprego dos codecs em redes de comunicações, e as principais limitações (ou benefícios) de determinadas conexões.

A Tabela VI apresenta a nota de opinião média (NOM), assim como seu intervalo de confiança, através dos valores NOM-δ(95%) e NOM+δ(95%), que corresponde ao intervalo de confiança dentro do qual está a média das notas dos avaliadores com probabilidade de 0,95.

TABELA VI
RESULTADOS DO TESTE SUBJETIVO

Config.	NOM-δ(95%)	NOM	NOM+δ(95%)
1	2,86	3,06	3,27
2	2,43	2,71	2,98
3	3,01	3,27	3,53
4	2,42	2,67	2,91
5	2,71	2,96	3,21
6	2,70	2,96	3,22
7	2,30	2,56	2,86
8	2,60	2,86	3,19
9	2,24	2,50	2,76
10	2,66	2,90	3,13
11	2,12	2,37	2,63
12	2,43	2,67	2,90
13	2,76	3,02	3,28
14	2,30	2,58	2,86
15	2,26	2,52	2,78
16	1,36	1,58	1,81
17	1,81	2,10	2,40
18	3,54	3,79	4,04
19	4,15	4,35	4,56
20	4,61	4,75	4,89

Os melhores resultados (NOM$>$ 3) — que correspondem a uma qualidade de voz acima de razoável — foram obtidos com duas configurações de dois codecs (1 e 3) e uma configuração de três codecs (13). É importante, entretanto, que se considere as conversações nos dois sentidos. Como a conexão 13 é simétrica, o resultado é obviamente válido nos dois sentidos.

Resultados experimentais anteriores [2] mostram que quanto mais próximo do início da conexão estiver o codificador de mais baixa taxa, maior será a degradação da qualidade de voz. Isso nos permite pressupor que os caminhos inversos das ligações 1 e 3, embora não avaliados, deveriam fornecer resultados superiores. Essa afirmativa é corroborada pela comparação dos resultados obtidos com as conexões simétricas 4 e 5, que usam os mesmos codecs em dois sentidos distintos. Pode-se observar que o codec de 8 kbit/s seguido do de 16 kbit/s fornece uma NOM=2,67 enquanto que para a ligação inversa a NOM é de 2,96. A exceção a esse tipo de comportamento foi observada quando um codec empregado utiliza supressor de silêncio. Isso pode ser verificado comparando os desempenhos obtidos com o par 3 $\times$ 5.

Se levarmos em conta o intervalo de confiança, outras configurações que também estão dentro de uma faixa de NOM aceitável são a 5 e a 6. Contudo, a 5 deve ser descartada tendo em vista que sua conexão no sentido oposto (a 4) apresenta um desempenho sig-

nificativamente inferior (entre ruim e razoável). Por outro lado, como a configuração 6 começa com um codec de 8 kbit/s e termina com um de 16 kbit/s, sua conexão inversa deverá fornecer resultados melhores ainda.

Como esperado, o aumento do número de codecs piora a qualidade de voz, independentemente se o codec adicional for posicionado no início ou no fim da conexão. Isso pode ser verificado pelos resultados obtidos com os pares de configurações listadas abaixo, onde a segunda, com um codec adicional, fornece uma NOM inferior à primeira (a exceção é quando se usa codec com supressor de silêncio, como será discutido mais adiante): 1×6, 2×7, 3×10, 4×11, 5×14, 1×8, 2×9, 3×13, 4×14 e 8×15.

Os resultados obtidos para os seguintes pares de configurações sugerem a não utilização de supressor de silêncio, independentemente do número de codecs da conexão: 1×2, 3×5, 8×9, 3×4, 6×7, 10×11 e 13×14. Em cada um desses pares os codificadores utilizados são idênticos, com a diferença que na segunda configuração de cada par um dos codificadores — o CS-ACELP a 8 kbit/s — usa supressor de silêncio. Em todos os casos o desempenho da segunda configuração é bem inferior ao da primeira, podendo chegar a uma diferença de NOM da ordem de 0,5.

Uma outra observação interessante é que uma conexão com um codec a mais, porém que não empregue supressor de silêncio, fornece um desempenho superior à conexão com um número menor de codecs, se esta última utilizar supressor de silêncio. Essa foi a exceção encontrada para a regra de redução de desempenho com o aumento do número de codificadores. Esses resultados podem ser verificados analisando os pares de configurações a seguir: 6×2, 10×4, 13×5, 8×2, 13×4, 15×9 e 15×11.

Os resultados dos testes mostram ainda que a substituição do codificador 16 LDCELP em uma conexão pelo codificador 16 CELP acarreta em uma queda de qualidade de voz. Essa queda, em termos de NOM, pode chegar a 0,3 que é bastante significativa. Isso pode ser observado analisando os seguintes pares de configurações: 3×1, 13×12, 14×11, 10×8, 13×8 e 14×9. Entretanto, se um dos codecs da conexão empregar supressor de silêncio, o uso do CELP no lugar do LD-CELP produz uma NOM superior, como mostram os resultados das configurações 2×4 e 9×11. Embora nesses casos as NOMs tenham sido maiores, observou-se durante a fase de montagem dos testes que as conexões que envolviam codecs com supressor de silêncio e codecs CELP, produziam distorções inaceitáveis quando os interlocutores falavam simultaneamente. Tais distorções se apresentavam na forma de eco, cortes de ambas as vozes dos interlocutores e cliques. Esses efeitos, altamente indesejáveis, não eram percebidos quando o CELP era substituído pelo LD-CELP.

Dos resultados obtidos nesse trabalho, pôde-se ainda verificar que o pior resultado, com uma NOM=2,37 foi para a configuração 11 que emprega três codecs, onde dois deles já mostraram deterioração de desempenho em outras conexões: o CELP e o CS-ACELP com supressor de silêncio.

V. Conclusões Finais

Levando em conta as ligações nos dois sentidos, conclui-se que as configurações que fornecem melhores resultados são: 1, 3, 6 e 13. De acordo com todas as observações e discussões anteriores, as configurações mais recomendáveis seriam, incluindo suas inversas:

- 8 CSACELP $\rightarrow$ 16 LDCELP
- 8 CSACELP $\rightarrow$ 16 LDCELP $\rightarrow$16 LDCELP
- 16 LDCELP $\rightarrow$ 8 CSACELP $\rightarrow$ 16 LDCELP

Finalmente, é importante ressaltar que a conversação simultânea de dois interlocutores não foi exposta para avaliação dos ouvintes, por sair da formalidade do tipo de teste aplicado. Contudo, os efeitos da conversação simultânea devem ser considerados, mesmo que de modo informal e fora dos testes de avaliação subjetiva como os aqui realizados. Uma possibilidade interessante seria avaliar apenas seus efeitos através de testes de comparação do tipo A versus B.

Referências

[1] A. Alcaim, *Processamento de Voz e Imagem*, Publicação CCE-PUC/Rio, 1a. Edição, Setembro 1999.

[2] J.R. Boisson, A. Alcaim, J.A. Solewicz, *Efeito de Conexões em Tandem Sobre o Desempenho de Codificadores Digitais de Voz*, Anais do 13o Simpósio Brasileiro de Telecomunicações, Águas de Lindóia, SP, Setembro, 1995, vol I, pp. 89-92.

[3] CCITT, *Recommendation P.80: Methods for the Subjective Determination of Transmission Quality, Blue Book*, ITU, Genebra, 1989, vol. V, pp. 197-198.

[4] A. Alcaim, J.A. Solewicz e J.A. Moraes, *Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro*, Revista da Sociedade Brasileira de Telecomunicações, vol. 7, no. 1, Dezembro 1992, pp. 23-41.

[5] Campos Neto, S.F., *Metodologias de Avaliação de Algoritmos de Codificação de Voz*, Dissertação de Mestrado, UNICAMP, Abril 1993.

[6] Silva, L.M., *Contribuições para a Melhoria da Codificação CELP a Baixas Taxas de Bits*, Tese de Doutorado, PUC-Rio, Fevereiro 1996.