REFLEXÕES SOBRE A CLASSIFICAÇÃO DA QUALIDADE ACÚSTICA DE DADOS DE CORPORA ORAIS

Reflections on the acoustic quality classification of spoken corpora data

FERRARI, Lucia de Almeida[1]
MELLO, Heliana Ribeiro[1]
VIEIRA, Marcelo[2]

[1]Universidade Federal de Minas Gerais
[2]McGill University

Resumo: *Corpora de fala geralmente servem como base para estudos de interface entre prosódia, pragmática e sintaxe, além daqueles fonético-fonológicos. Uma boa audibilidade é essencial, mas não sempre suficiente para estas finalidades. Neste artigo são apresentados os procedimentos metodológicos e os critérios adotados na classificação da qualidade acústica dos corpora da família C-ORAL (CRESTI; MONEGLIA, 2005; RASO; MELLO, 2012). Especificamente, serão mostrados os avanços metodológicos implementados especialmente no C-ORAL-BRASIL II (RASO; MELLO; FERRARI, em preparação). O protocolo prevê as etapas de amostragem dos trechos de áudio e a análise dos parâmetros avaliados: relação sinal-ruído, sobreposição, f0 e formantes (F1 e F2). O procedimento utilizou uma série de scripts em Praat que permitiu automatizar a extração dos dados necessários à avaliação. O papel do avaliador é determinante na conferência dos vários parâmetros, pois seu julgamento é dirigido através de critérios precisos, mas que necessitam de checagem à oitiva, por inspeção visual do espectrograma e de eventuais correções manuais. Cada parâmetro recebeu uma etiqueta que indica seu valor. Para se obter uma etiqueta final de cada áudio, que contemple os vários parâmetros, calculou-se uma média ponderada com valores arbitrários atribuídos a cada parâmetro. Os pesos maiores foram atribuídos a f0 e formantes, por se entender que eles são os mais relevantes para as análises fonético-fonológicas além daquelas pragmáticas.*
Palavras-chave*: corpus de fala; qualidade acústica; prosódia.*

Abstract: *Speech corpora usually serve as the basis for prosodic, pragmatic and syntactic interface studies, as well as phonetic and phonological ones. Good audibility is an essential feature; however, it is not always enough for research purposes. In this article, we present methodological procedures and criteria adopted in the acoustic quality classification of the C-ORAL family corpora (CRESTI; MONEGLIA, 2005; RASO; MELLO, 2012). Specifically, we will show methodological improvements implemented especially in the C-ORAL-BRASIL II (RASO; MELLO; FERRARI, forthcoming).The procedures establish the audio sampling stages and the analyzed parameters assessment: signal-noise ratio, overlapping, f0 and formants (F1 and F2). The method employed a series of Praat scripts that allows the automatic extraction of data to be analyzed. The evaluator's role is critical in checking all parameters. In fact, his judgment is guided by precise criteria, but they need hearing and spectrogram inspection and possible manual corrections. Each parameter received a tag that indicates its value. In order to obtain the final tag for each audio that could gather all the parameters, a weighted average value with arbitrary weights for each parameter was calculated. The highest weights were conferred to f0 and formants, given that they are the most relevant ones in phonetic and phonological analyses, besides pragmatic ones.*
Keywords: *speech corpus; acoustic quality; prosody.*

1 Corpora de Fala e qualidade acústica

Neste artigo apresentaremos os procedimentos metodológicos e critérios para classificação da qualidade acústica das gravações que integram o projeto C-ORAL-BRASIL, corpora de fala do português brasileiro. Para tal, mostraremos o que há registrado para a classificação de qualidade acústica da família de corpora C-ORAL, apresentaremos o C-ORAL-BRASIL, bem como discutiremos os avanços metodológicos alcançados dentro deste projeto para a classificação acústica de dados de corpora de fala em geral. A necessidade de avaliação da qualidade acústica de áudios que integram um corpus de fala justifica-se dada a natureza e propósitos de tal conjunto de dados. Corpora de fala atuais dedicam-se a estudos que frequentemente pautam-se por fenômenos de interface entre a prosódia, a pragmática e a sintaxe. Para além disso, corpora de fala também são uma importante fonte de dados para estudos fonético-fonológicos. Assim, a avaliação da qualidade acústica é fundamental para que se garanta a adequação de um corpus aos objetivos para os quais foi planejado. Embora uma boa audibilidade possa ser suficiente em muitos estudos de sintaxe e pragmática, quando se trata de interface com estudos prosódicos são necessários corpora planejados com uma atenção maior à qualidade acústica.

2 Corpora de fala e prosódia: a família C-ORAL

O C-ORAL-ROM (CRESTI; MONEGLIA, 2005) é um corpus de referência de fala espontânea das principais línguas românicas (italiano, francês, espanhol e português europeu). Sua compilação foi realizada por um grupo de laboratórios linguísticos de universidades europeias: (a) LABLITA, *Laboratorio linguistico del Dipartimento di italianistica* (Università di Firenze); (b) DELIC, *Description Linguistique Informatisée sur Corpus* (Université de Provence); (c) CLUL, Centro de Linguística da Universidade de Lisboa; (d) *Departamento de Lingüística, Laboratorio de Lingüística Informática* (Universidad Autónoma de Madrid). Os diferentes grupos, seguindo uma arquitetura comum, puderam reutilizar material previamente coletado para outros projetos (de 1978 até 2002) ou realizar novas gravações para o C-ORAL-ROM.

Como nos vários corpora uma boa parte do material havia sido gravado de forma analógica, a avaliação da qualidade acústica pautou-se principalmente em diferenciar as gravações realizadas em formato digital. A qualidade acústica foi, portanto, dividida em três faixas: a qualidade A se refere a todas as gravações digitais, a B a gravações analógicas com boa resposta dos microfones, baixo ruído de fundo, baixa porcentagem de sobreposições e possibilidade de computação da curva de f0 em quase todos os arquivos. A qualidade C, a mais baixa, se refere a gravações analógicas que apresentam baixa resposta dos microfones, ruído de fundo, uma porcentagem média de enunciados com sobreposição e possibilidade de computação da curva de f0 em boa parte dos arquivos (MONEGLIA, 2005, p. 32).

O C-ORAL-BRASIL I (RASO; MELLO, 2012), corpus de referência do português brasileiro falado informal, segue a mesma arquitetura do C-ORAL-ROM. A coleta de dados, realizada entre 2006 e 2011, mas principalmente entre 2008 e 2010, foi feita inteiramente para a compilação do corpus. Uma das preocupações foi a de obter a melhor qualidade acústica possível em contexto natural: para tal foram utilizados gravadores digitais (Marantz PMD660 Professional Solid State Recorder) e microfones omnidirecionais (Sennheiser MD 421-II 4 e Shure PG58-XLR) e de lapela (Sennheiser ME 4 clip-on), além de um *mixer* (Behringer XEXYX 1222 FX) quando as interações eram de mais de três pessoas. A avaliação dos áudios selecionados foi feita com base nos seguintes critérios: (a) resposta dos microfones; (b) possibilidade de análise fonética; (c) sobreposições; (d) ruído de fundo; (e) possibilidade de computação da curva de F0; (f) para as gravações com menor qualidade acústica, nível de clareza da escuta. Estes critérios foram avaliados por um time de pesquisadores mais familiarizados com pesquisas prosódicas e fonéticas que classificaram os áudios em cinco faixas: A, AB, B, BC e C, sendo A classificada como qualidade ótima e C como baixa (Raso, 2012, p. 74). A avaliação, de cunho perceptual, foi efetuada através da inspeção visual do espectrograma e da oitiva.

2.1 A avaliação acústica do C-ORAL-BRASIL-II

Para o C-ORAL BRASIL II (Raso, Mello e Ferrari, em preparação) desenvolveu-se um protocolo de avaliação da qualidade acústica dos áudios que o compõem. Esse protocolo é aplicado através de dois *scripts* do Praat (Boersma e Weenink, 1992-2019), *Audio_quality* e *Overlapping*, implementados por Marcelo Vieira. Como softwares livres, esses *scripts* foram desenvolvidos como ferramentas flexíveis e serão disponibilizados em breve em <https://vieiramarcelo.wordpress.com/praat-scripts/>. Os parâmetros gerais de avaliação são: relação sinal-ruído, sobreposição, f0 e formantes (F1 e F2).

O primeiro passo é a amostragem dos áudios. Dividiram-se os áudios em vários trechos equidistantes (*stretches*), dentro dos quais uma janela de 10 segundos de áudios é analisada. O número de trechos depende da duração dos áudios: a) duração < 1 min: 3 trechos; b) 1 min ≤ duração < 5 min: 5 trechos; c) 5 min ≤ duração < 10 min: 10 trechos; d) duração ≥ 10 min: 15 trechos. Para áudios com menos de 30 s – em que não se pode obter 3 janelas de análise de 10 s –, o tamanho das janelas é igual a 1/3 da duração do áudio. Dessa forma, o maior áudio (22' 12'') teve mais de 11% de sua duração analisada, sendo que cada janela de análise iniciava-se em frações diferentes do áudio – o que contribui para melhor representatividade. Em cada janela, o avaliador verifica f0 e formantes. A análise de sobreposição e ruído são procedimentos semi-automatizados.

Para f0, em cada janela de análise, são disponibilizados o objeto *Pitch*, o áudio original, o áudio do objeto *Pitch* (*Hum sound*) e o áudio original com a f0 substituída por aquela calculada no objeto *Pitch* (esse processo permite que, havendo erro no cálculo de f0, sua consequência possa ser ouvida no áudio original). Além disso, há uma imagem do espectrograma sobreposto pela curva de f0 em duas versões: bruta e suavizada (10 Hz de largura de banda). A seguir, os seguintes parâmetros binários são avaliados:

1) Se, por oitiva, os três áudios (original, *hummed* e o original manipulado) são compatíveis; isto é, se não há nenhuma alteração substancial na impressão auditiva do contorno melódico causada por erros de cálculo de f0;
2) Se a curva de f0 segue o contorno dos harmônicos no espectrograma.
3) Se há erros severos de cálculo de f0 (*having*, *doubling*, ou erros de outra natureza);
4) Se é possível corrigir manualmente os contorno de f0 através do objeto *Pitch*.

A faixa de *pitch* (*pitch range*) é automaticamente pré-ajustada seguindo o procedimento de Hirst (2007), sendo possível reajustá-la se necessário. Isso permite que erros de *having* e *doubling* devidos a um mau ajuste de faixa de *pitch* sejam corrigidos antes da avaliação.

Para formantes, o *script* seleciona aleatoriamente três vogais para cada trecho. A identificação prévia das vogais é feita pelo *script Syllable Nuclei* (JONG; WEMPE, 2009) adaptado por Hugo Quené, Ingrid Persoon e Nivja de Jong, versão 2 (<https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2>). Caso o segmento identificado não seja uma vogal, o avaliador assinala isso e, então, o *script Audio_quality* fornece um novo segmento. Para cada vogal, disponibiliza-se o áudio original, o objeto *Formant* e seu áudio, e o áudio sintetizado com base em F1 e F2 calculados no objeto *Formant* e na média de f0 da vogal do áudio original. Por fim, o avaliador julga os seguintes parâmetros binários:

1) Se, por oitiva, os três áudios (original, objeto *Formant* e áudio sintetizado) são compatíveis; isto é, se não há nenhuma alteração substancial na impressão auditiva da qualidade vocálica (tendo o áudio original como referência);
2) Se, por base na visualização, é possível distinguir bem os dois primeiros formantes.

A análise de sinal-ruído, por sua vez, é feita através de um procedimento que calcula a diferença entre a intensidade do ruído e a do sinal de fala. Para isso, o avaliador identifica os intervalos de fala e os intervalos de silêncio/ruído na janela de análise. Daí, calcula-se a diferença entre as médias ponderadas de intensidade entre os intervalos de fala e de ruído. O peso inserido no cálculo das médias é a duração de cada intervalo.

Finalmente, para análise de sobreposição, um *script* calcula a proporção de palavras sobrepostas em relação ao número total de palavras do corpus. Para tanto, é necessário uma transcrição prévia que contenha etiquetas próprias para casos de sobreposição.

Após essas etapas, tabelas com dados de cada janela de análise são fornecidas. Contudo, no fim, deve-se reduzir todos esses dados para uma única informação por áudio. Para tanto, calcula-se a média geral do áudio para todos os itens de cada parâmetro, através da promediação dos valores de cada janela. Na tabela dos formantes, porém, o resultado é apresentado para cada vogal. Assim, é feita, antes, uma promediação por vogais dentro de cada janela de análise, para então se calcular a média para todo o áudio. Ressalta-se que esses processos não se aplicam à análise de sobreposição, a qual é feita de uma só vez com base na transcrição completa do áudio. Além disso, é importante dizer que, enquanto a média geral para formantes e f0 são valores de 0 a 1 (pelo fato de os parâmetros serem binários), os valores para análise de sobreposição e ruído são dados por porcentagem e dB, respectivamente.

O resultado das promediações, bem como o da análise de sobreposição, é transformado em etiquetas para cada parâmetro: A, AB, B, BC e C – sendo A, a melhor qualidade e C, a pior. Para isso, nos dados sobre f0 e formantes, faz-se a média ponderada de cada um dos itens de análise descritos acima. Os pesos são arbitrários e definidos *a priori* pelos avaliadores. No caso do CORAL-BRASIL II, adotaram-se peso 0 (nulo) para os itens (1) e (2) de f0, já que todos os áudios satisfaziam esses parâmetros. Para o item (3), o peso foi 0.5, enquanto para o item (4), foi 1. Isso se justifica uma vez que um áudio com erros incorrigíveis é pior. A presença de erros corrigíveis demanda mais trabalho e atenção dos pesquisadores e, portanto, decidiu-se contabilizar esse item à parte, ainda que com um peso menor. Para os formantes, o item (1) tem peso 1, enquanto o item (2) tem peso 3. Isso se deve ao fato de a inspeção visual ser, de fato, o que detecta erros nos formantes. É possível que haja erros que comprometam toda a média de F1 e F2, sem que a impressão acústica seja alterada substancialmente.

Finalmente, para atribuir uma etiqueta para o áudio como um todo, foi feita uma média ponderada de todos os parâmetros (f0, formante, sobreposição e ruído). Para esse fim, as etiquetas de cada parâmetro foram transformadas em números de 1 a 5, sendo 1 equivalente a C, e 5, a A. A seguir, esses valores foram inseridos no cálculo da média junto a seus respectivos pesos: 1 para sobreposição e também para ruído; 2 para formantes e 3 para f0. Os valores foram escolhidos por se entender que sobreposição e ruído, embora influentes por si, são potencialmente mais prejudiciais quando acarretam prejuízos a f0 e aos formantes. Portanto, preferiu-se valorizar mais esses dois últimos parâmetros, dada a importância que possuem em estudos fonéticos e fonológicos.

REFERÊNCIAS

1. BOERSMA, P.; WEENINK, D. *Praat: doing phonetics by computer.* 1992-2019. Disponível em: www.praat.org.
2. CRESTI, E.; MONEGLIA, M. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages.* John Benjamins Publishing Company, 2005
3. HIRST, D. A. Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. In: *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, p.1233-1236, 2007.
4. JONG, N.; WEMPE, T. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41, 2, p. 385-390, 2009.
5. MONEGLIA, M. The C-ORAL-ROM resource. In: CRESTI, E.; MONEGLIA, M. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. John Benjamins Publishing Company, p. 1-40, 2005
6. RASO, T.; MELLO, H. (Org.). *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal.* Belo Horizonte: Editora UFMG, 2012.
7. RASO, T. Specifications. In: RASO T.; MELLO, H. (Org.). *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal.* Belo Horizonte: Editora UFMG, p. 1-130, 2012

8. RASO, T. O corpus C-ORAL-BRASIL. In: RASO, T.; MELLO, H. (Org.). *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG, p. 55-90, 2012
9. RASO, T.; MELLO, H.; FERRARI, L. A. *C-ORAL-BRASIL II* (em preparação)