Análise da entoação de declarativas em espanhol, português brasileiro e espanhol/LE com PentaTrainer2

Cristiane C. Silva [1], *Plínio A. Barbosa*[2]

[1] Grupo de Estudos de Prosódia da Fala, Instituto de Estudos da Linguagem, Unicamp, Brasil
[2] Grupo de Estudos de Prosódia da Fala, Instituto de Estudos da Linguagem, Unicamp, Brasil

cris.silva.unicamp@gmail.com, pabarbosa@gmail.com

Resumo

Neste estudo, analisamos duas funções comunicativas transmitidas pela entoação: a função de proeminência e fronteira em espanhol peninsular (doravante E/LM), português brasileiro (doravante PB/LM) e espanhol como língua estrangeira (doravante E/LE). Para isso, utilizamos o modelo PENTA e aplicamos a ferramenta automática PENTATrainer2. A aplicação da ferramenta permitiu avaliar o poder de síntese de contornos melódicos em duas línguas diferentes (português e espanhol (esta última falada tanto por nativos como por estrangeiros) e em dois estilos diferentes: leitura e narração. Os resultados mostraram que a precisão do modelamento é similar à encontrada por [1,2] a partir da análise de um corpus semelhante em PB e PE (português europeu).

Palavras-chave: modelamento entoacional, estilos de fala, prosódia entre línguas

1. Introdução

O modelo PENTA, diferentemente de outros modelos entoacionais, está motivado articulatoriamente, ou seja, o modelo assume que os mecanismos relacionados com a articulação guiam a geração dos contornos de F0. Este estudo tem por objetivo testar o poder explicativo do modelo ao: (1) estender o domínio prosódico da sílaba para a palavra fonológica; (2) testar o modelo em duas línguas diferentes - português e espanhol (como língua materna e como língua estrangeira), e em diferentes estilos de fala e, finalmente, (4) apresentar uma primeira proposta de esquemas de codificação para a transmissão das funções comunicativas de proeminência e fronteira analisadas.

2. Modelo PENTA

Inicialmente, Xu e Wang [3] propuseram o modelo de aproximação do alvo (TA) a partir da análise acústica de dados sobre o tom e a entoação do mandarim. Nesse modelo, definiram os **alvos de *pitch*** como as menores unidades operáveis articulatoriamente associadas com níveis de *pitch* com função linguística, ou seja, relacionados com os tons ou com os *pitch accents*. Assim, definiram dois alvos de *pitch* que podem ser **dinâmicos**, especificados como [ascendentes] ou [descendentes] ou **estáticos**, especificados como [altos] ou [baixos].

Posteriormente, Xu [4] propôs que o modelo TA não se limitasse apenas à realização de tons lexicais, mas que servisse também como mecanismo de base para a codificação de outros significados comunicativos relacionados com a F0. Em outras palavras, ele tornou explícito no modelo que outras funções além do tom lexical fossem codificadas através do processo de aproximação do alvo. Para isso, propôs o **Modelo de Codificação Paralela e Aproximação do Alvo** (modelo PENTA) definido pela manipulação dos seguintes parâmetros: **gama de *pitch*** (extensão do alvo de *pitch*), **força articulatória** (velocidade de aproximação do alvo de *pitch*), **duração** do hospedeiro (quantidade de tempo para a aproximação do alvo de ***pitch***).

Dessa forma, o modelo PENTA assume que a prosódia da fala deve transmitir funções comunicativas de forma paralela através de esquemas de codificação individuais e que, apesar de serem abstratos, tais esquemas de codificação sempre estão ligados às funções comunicativas. Assim, é por meio do processo de aproximação do alvo que se mantém uma ligação contínua entre as funções comunicativas e os contornos de F0.

Seguindo os pressupostos anteriores, o modelamento efetivo da prosódia só poderá ser alcançado se os esquemas de codificação das funções comunicativas forem simulados. Por isso, Xu e Prom-on [5,6] propuseram o modelo quantitativo de aproximação do alvo (qTA) que é a implementação dos modelos TA e PENTA.

Assim, o contorno de F0 será a resposta do processo de aproximação dos alvos de *pitch* subjacentes. Esses alvos de *pitch* estão associados ao domínio do hospedeiro (para os autores será sempre a sílaba) e especificam os alvos que se pretende alcançar quando o movimento de F0 é realizado. Na implementação matemática do modelo, para cada hospedeiro, o F0 é calculado seguindo a Equação 1:

$$f_0(t) = (m.t + b) + (c_1 + c_2.t + c_3.t^2).e^{-\lambda.t} \quad (1)$$

Os coeficientes da Equação 1 são: (m) inclinação do alvo, (b) altura do alvo e (λ) força do alvo. Já os coeficientes (c_1, c_2 e c_3) estão relacionados com as condições iniciais, ou seja, asseguram que os valores de F0 em um intervalo inicial tenham o mesmo valor, a mesma inclinação e curvatura no intervalo imediatamente precedente.

O efeito dos valores de (m, b e λ) sob a forma do contorno de F0 pode ser visto na Figura 1:

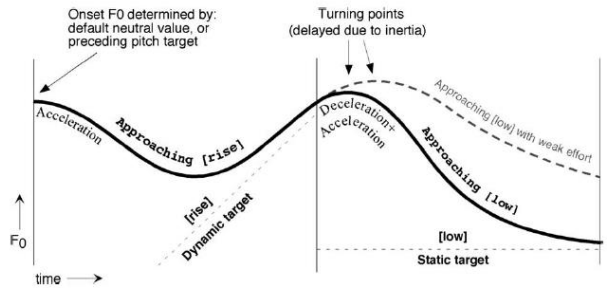

Figura 1: *Modelo TA. As linhas verticais representam as fronteiras da sílaba. As linhas pontilhadas representam os alvos subjacentes. A linha em negrito contínua representa o contorno de F0 que resultou da aproximação assintótica dos alvos. A curva pontilhada simula o efeito do menor esforço. Figura adaptada de [3]*

Assim, quando m é positivo e b é igual a zero, gera-se a aproximação do alvo dinâmico (lado esquerdo da Figura 1). Enquanto que no lado direito da mesma figura, o parâmetro m é igual a zero e b tem um valor baixo produzindo uma aproximação assintótica em direção ao mínimo (lado direito da Figura 1). A posição do ponto de inflexão é controlada pelo valor do parâmetro de força λ que está relacionado com quão rápido um alvo é alcançado: quanto mais baixo seu valor mais inclinado para a direita o ponto de inflexão estará (como se observa na linha pontilhada).

Esses três parâmetros podem ser obtidos a partir de uma técnica de aprendizagem de análise por síntese que encontra valores individuais para os três parâmetros para um conjunto inteiro de enunciados. Esse procedimento é realizado pelo algoritmo PENTATrainer2 que explicaremos a seguir.

3. Modelamento dos contornos de F0 em leitura e narração em E/LM, PB/LM e E/LE

O algoritmo PENTATrainer2, diferentemente da versão anterior, implementa uma técnica de otimização global. Essa técnica consiste em encontrar um único conjunto de parâmetros (m,b,λ) a partir de um conjunto inteiro de enunciados. Assim, partindo da Equação 1 e de um conjunto aleatório de valores para os três parâmetros associados com uma etiqueta particular, o algoritmo produz um contorno de F0 que é comparado com o contorno original. Essa comparação se dá pelo cálculo da raiz quadrada do erro quadrático médio (RMSE) entre o contorno original e o sintetizado e é usada para ajustar os parâmetros iterativamente até que um critério de parada seja alcançado.

O algoritmo de aprendizagem é precedido por uma fase de anotação e seguido por uma fase de síntese. Durante a fase de anotação, o pesquisador cria um conjunto de camadas na quais determina as fronteiras e as etiquetas de cada uma das funções comunicativas que deseja analisar. Na fase de síntese, são gerados novos contornos a partir da anotação feita previamente.

3.1 Corpus

O corpus consiste de produções paralelas em E/LE, PB/LM de 10 brasileiros (5 mulheres e 5 homens) com nível superior. Os brasileiros são do estado de São Paulo e têm idades entre 27 e 48 anos. Todos aprenderam espanhol depois dos 18 anos e moravam em Madri no momento das gravações há no mínimo 6 meses e no máximo 16 anos. Já as produções em E/LM são de 5 espanhóis (3 mulheres e 2 homens), sendo 3 informantes de Madri e dois de cidades próximas a Madri. Todos têm nível superior.

A tarefa consistiu, primeiramente, na leitura de um excerto da história de Don Quixote intitulado "*Gigantes con aspas*" [7]. Em seguida, na leitura de 39 enunciados retirados do texto e lidos isoladamente em ordem aleatória e, finalmente, na narração da história lida. A análise se baseia em excertos de **368** palavras lidas por estilo (frases selecionadas da leitura do texto e frases lidas isoladamente) em espanhol e de **354** palavras (nos dois estilos) lidas em português. O que corresponde a 39 enunciados lidos em cada estilo. Os excertos analisados das narrações em espanhol e português têm entre 106 a 368 palavras. Isso corresponde a no mínimo 8 enunciados e no máximo 32 enunciados por informante no estilo de narração. A delimitação dos enunciados no estilo de narração foi definida de acordo com critério pragmático [8].

3.2 Teste de percepção de proeminência

Foi realizado um teste de identificação com 195 ouvintes espanhóis divididos em 15 grupos de 13 ouvintes. Dessa maneira, cada grupo de 13 ouvintes avaliou os dados em E/LM e E/LE de cada um dos informantes (10 brasileiros e 5 espanhóis). O mesmo procedimento foi adotado para avaliação da proeminência em PB. Assim, o teste foi realizado com 130 brasileiros também divididos em grupos de 13 ouvintes. O corpus que serviu de estímulo para a realização dos dois testes de percepção foi o mesmo usado para análise utilizando o PENTATrainer.

A tarefa consistiu em escutar cada uma das frases e marcar as palavras que o ouvinte considerasse proeminentes, ou seja, as palavras que o locutor pronunciou com a intenção de chamar a atenção de quem o ouvia. O teste de percepção foi realizado pela internet através do site (www.surveygizmo.com).

Depois de aplicado o teste, a porcentagem de ouvintes que marcaram uma palavra como proeminente foi usada para determinar a proeminência. Para isso, utilizamos um teste z de proporção. Dado que o número total de juízes era 13 (N=13) e o nível de significância de 0.05, palavras que foram marcadas como proeminentes por até 15% dos ouvintes **não** foram consideradas proeminentes. Já as palavras consideradas proeminentes por 16% até 100% dos ouvintes foram consideradas proeminentes.

3.3 Fronteira prosódica

Consideramos fronteira prosódica toda quebra no enunciado seguido por pausa silenciosa e para cada uma dessas interrupções do enunciado atribuímos o significado de continuidade (a pessoa ia continuar falando) ou terminalidade (o falante já terminou de falar).

3.4 Análise por síntese

Utilizamos a versão do PENTATrainer2 [9] que funciona como um *plugin* do Praat [10] para modelar os contornos de F0.

Primeiramente, anotamos em dois níveis tanto a função de proeminência quanto a de fronteira. Consideramos como domínio para a anotação das funções a palavra fonológica [1,2]. Realizamos alguns testes prévios considerando o domínio da sílaba, mas verificamos que tanto para os dados em português

como em espanhol, a aprendizagem se mostrou sempre melhor quando utilizávamos a palavra fonológica como domínio em comparação com a sílaba. Cada função comunicativa foi anotada em uma camada separada do Praat. A primeira função foi a de proeminência sendo (p) proeminente *versus* (n) não proeminente. A segunda de fronteira sendo (t) fronteira terminal *versus* (c) fronteira não terminal, ou seja, continuativa. Analisamos também a função comunicativa de modalidade, porém, neste estudo apresentamos somente os resultados com relação às declarativas.

Todas as camadas etiquetadas e os contornos de F0 originais serviram de entrada para o algoritmo de aprendizagem do PENTATrainer2. Para obter uma melhor estimativa dos valores dos contornos de F0, restringimos os valores mínimos e máximos de F0 entre 75 e 250 Hz para os homens e 120 e 400 Hz para as mulheres [2].

Na fase de aprendizagem é feita a inferência dos valores dos três parâmetros (m, b, λ) para o mesmo conjunto de enunciados. Por essa razão o corpus foi dividido em 75 subconjuntos: 10 brasileiros x 2 línguas x 3 estilos = 60 subconjuntos e 5 espanhóis x 3 estilos =15 subconjuntos. Dessa maneira, cada subconjunto contém os enunciados de um sujeito e de um estilo específico em E/LE, E/LM ou PB/LM.

Para o algoritmo de aprendizagem mantivemos os valores *default* sugeridos pelo programa [6].

Na fase final de síntese, é possível observar e escutar o contorno original e o sintetizado além de serem apresentados também os alvos de *pitch* (Figura 2):

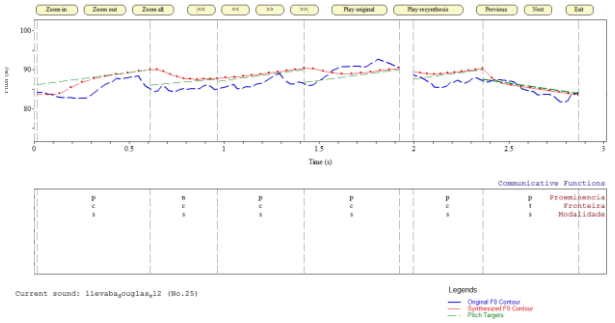

Figura 2: *Contorno de F0 original (azul) contorno de F0 sintetizado (vermelho) e alvos de pitch (verde) na parte superior. Na parte inferior, Funções comunicativas do enunciado. Don Quijote llevaba camisas limpias y algún dinero. Pronunciado pelo informante brasileiro Dou.*

Os resultados apresentados na próxima seção se referem à análise dos arquivos de saída *parameters* que contém os valores dos três parâmetros por etiqueta para cada subconjunto e o *accuracy_synthesis* que apresenta a correlação e a raiz quadrada do erro quadrático médio (RMSE) em semitons re 1Hz entre os enunciados originais e seus respectivos enunciados sintetizados.

4. Resultados

A Tabela 1 mostra os valores da mediana e desvio-padrão para o RMSE (em semitons) e mediana e desvio-padrão da correlação (em porcentagem) entre os contornos original e sintetizado por língua e estilo correspondentes aos 39 enunciados lidos nos dois estilos de leitura e aos 8 a 32 enunciados da narração. Podemos observar que o desempenho do modelo é semelhante para os dados em PB/LM e E/LE e que é ligeiramente superior ao desempenho com os dados em E/LM no estilo de leitura.

Tabela 1: *Mediana de RMSE (e desvio-padrão) e mediana de correlação (e desvio-padrão) por língua e por estilo.*

Língua	Estilo	RMSE	Correlação
E/LM	Leitura Frases	2.3(0.8)	57(23)
E/LE		2.4(0.8)	62(24)
P/B		2.5(0.9)	62(25)
E/LM	Leitura Texto	2.3(0.7)	59(22)
E/LE		2.3(0.8)	61(21)
P/B		2.5(0.8)	65(24)
E/LM	Narração	1.8(0.7)	50(23)
E/LE		1.9(0.7)	49(23)
P/B		2.1(0.7)	49(21)

Uma ANOVA de dois fatores com variável dependente 'correlação' e variáveis independentes: 'língua (E/LM, E/LE e PB/LM)', 'estilo da gravação' (leitura das frases, leitura do texto e narração), e 'sexo' (masculino e feminino) revelou diferença estatisticamente significativa para 'língua' ($F(2,1931) = 6.53$, $p < 0.002$) e 'estilo' ($F(1,1931) = 24.58$; $p<0.0000001$).

Um teste post hoc Tukey de comparações de médias revelou diferença significativa entre as médias das correlações do E/LM em comparação com as do E/LE e entre E/LM e PB/LM, porém não houve diferença significativa entre as médias em E/LE e PB/LM. Isso significa que a aprendizagem foi estatisticamente melhor com os dados E/LE e PB/LM.

Em seguida, analisamos os parâmetros do modelo associados com as funções comunicativas de proeminência e fronteira para propor um esquema de codificação para cada função.

A Figura 3 apresenta os histogramas para a inclinação em semitons/s e altura em semitons dos alvos de *pitch* das declarativas em E/LM, E/LE, e PB/LM nos três estilos (leitura das frases, do texto, narração) para a etiqueta p (palavra proeminente) e t (fronteira terminal).

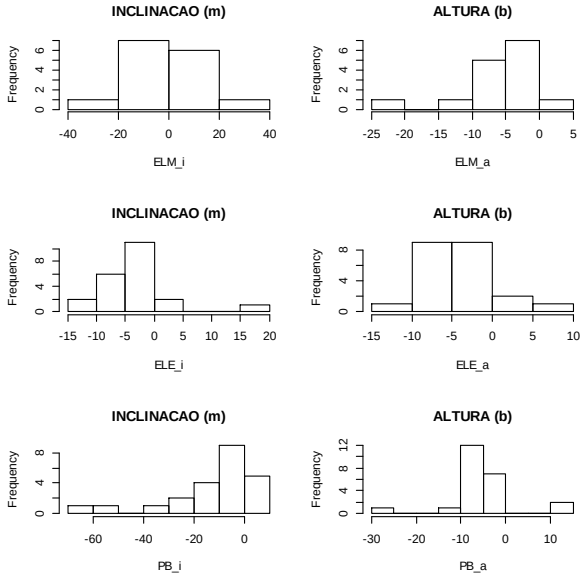

Figura 3: *Histogramas da inclinação do alvo (st/s) e altura (st) em E/LM (topo), E/LE (centro) e PB/LM (embaixo) para as palavras proeminentes em fronteira terminal.*

A observação dos histogramas revela que, com exceção dos valores de inclinação (m) em E/LM, todos os demais histogramas apresentam principalmente a combinação de valores de inclinação (m) e altura (b) negativos nas palavras proeminentes em fronteira terminal tanto em leitura como na narração. Essa combinação de inclinação e altura negativas tem como resultado alvos de *pitch* descendentes. Já em E/LM, observamos que o histograma de inclinação é bimodal. A ocorrência de valores positivos de (m) se concentra no estilo de narração. Segundo [2] o efeito da combinação de inclinação do alvo positiva e altura negativa é a atenuação da descida do contorno de superfície.

A comparação dos parâmetros (m) e (b) sugere que os valores de inclinação e altura são negativos e mais elevados em PB/LM e E/LE que em E/LM, essa diferença, porém, não foi estatisticamente significativa.

A Figura 4 apresenta os histogramas para a inclinação em semitons/s e altura em semitons dos alvos de *pitch* das declarativas em E/LM, E/LE, e PB/LM nos três estilos para a etiqueta p (palavra proeminente) e c (fronteira continuativa).

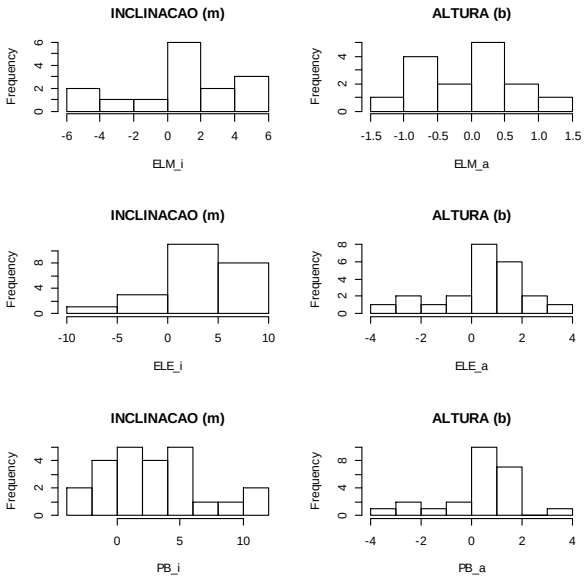

Figura 4: *Histogramas da inclinação do alvo (st/s) e altura (st) em E/LM (topo), E/LE (centro) e PB/LM (embaixo) para as palavras proeminentes em fronteira continuativa.*

Tanto em E/LM como em E/LE e PB/LM ocorre a combinação de valores positivos e baixos de inclinação (m) e altura (b) o que resulta em alvos de *pitch* ascendentes. Em E/LM, porém, houve diferença de gênero na narração, pois os homens apresentaram alvos de *pitch* ascendentes e as mulheres alvos de *pitch* descendentes. Há uma tendência a que os valores de inclinação e altura sejam mais elevados em E/LE e PB, essa diferença, porém, não foi estatisticamente significativa.

5. Discussão e Conclusão

Os resultados obtidos para as correlações em PB e E/LE são semelhantes aos encontrados por [2] para os dados em PB e PE para estilos e funções comunicativas semelhantes, porém, são inferiores aos obtidos por [6] que analisaram três *corpora* distintos em tailandês, mandarim e inglês. Essa diferença, porém, não podemos atribuir ao domínio utilizado para análise, pois realizamos testes com os dados em E/LM, E/LE e PB/LM e a aprendizagem foi sistematicamente pior quando utilizávamos o domínio da sílaba em detrimento da palavra fonológica. Acreditamos que as diferenças na aprendizagem entre o nosso estudo e o estudo de [2] com relação ao estudo de [6] não se devam ao domínio utilizado para a determinação das funções comunicativas (palavra fonológica vs sílaba), mas principalmente ao tipo de *corpus* analisado, ou seja, mais próximo à fala natural vs *corpus* muito controlado.

Dessa forma, apesar de termos obtido correlações médias de 60% e portanto inferiores às obtidas por [6], os resultados nos encorajam a seguir utilizando uma análise por síntese com o modelo qTA, dado que analisamos apenas duas funções comunicativas e utilizamos um corpus bem próximo ao da fala natural.

Na etapa seguinte do estudo, será discutida também a função comunicativa de modalidade por meio da análise das interrogativas totais e parciais do corpus da pesquisa. Além disso, discutiremos com maior profundidade a questão da variabilidade entre as línguas analisadas e entre os sujeitos a fim de verificar em que medida os parâmetros do modelo são realmente capazes de explicar os padrões gerais da dinâmica de F0 associados com as funções comunicativas de proeminência e fronteira em E/LM, PB/LM e E/LE.

6. References

[1] Barbosa, P. A.; Mixdorff, H.; Madureira, S. Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese. In: Interspeech, 2011, Florença. Proceedings… Florença, 2011.

[2] Barbosa, P. A. "Intonation modeling in cross-linguistic research". Benjamins. No prelo.

[3] Xu, Y., Wang, Q.E., "Pitch targets and their realization: Evidence from Mandarin Chinese", Speech Communication, 33: 319–337, 2001.

[4] Xu, Y., "Speech melody as articulatory implemented communicative functions" Speech Communication, 46: 220-251, 2005.

[5] Prom-on, S.; Xu, Y. Thipakorbn, B. "Modeling tone and intonation in Mandarin and English as a process of target approximations". The Journal of the Acoustical Society of America**.** v. 125(1), p. 405-424, 2009.

[6] XU, Y.; Prom-on, S. "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning". Journal of Phonetics, v. 57, p. 181-208, 2014.

[7] Aguilar, S. A. Quijote: Adaptación, Nota y Actividad. Barcelona: Vicens Vives, 2004.

[8] Cresti, E. Corpus di italiano parlato. v. 1. Florence: Accademia della Crusca, 2000.

[9] Xu, Y., Prom-on, S., "PENTAtrainer.praat", Online: http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer2/

[10] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" (Version 5.4.17) [Computer program], Online: http://www.praat.org.