INTERNATIONAL WORKSHOP
CEHUM 2014

University of Minho - Braga, Portugal
30th Janurary 2014

Listener

Listener: a Pronunciation Training System for Brazilian-Portuguese-accented English

Authors: Gustavo Mendonça, MS Candidate (ICMC/USP)
Sara Candeias, PhD (IT/UC)
Aldebaro Klautau Jr., PhD (LaPS/UFPA)
Sandra Maria Aluisio, PhD (ICMC/USP)

SCHEDULE

1. Motivation
2. Proposal
3. Gap
4. Research Hypothesis
5. A Brief Review of Automatic Speech Recognition
6. Method
7. Evaluation
8. Prototype Version

1. Motivation

Brazil is among the countries with the lowest proficiency-wise of the English language.

Education First's *English Proficiency Index* 2013[1]:
Brazil ranked 38th out of 60 countries. – Low proficiency

GlobalEnglish's *Business English Index 2013* [2]:
Brazil ranked 71st across 77 positions. – Beginner

- Business English Index 2013

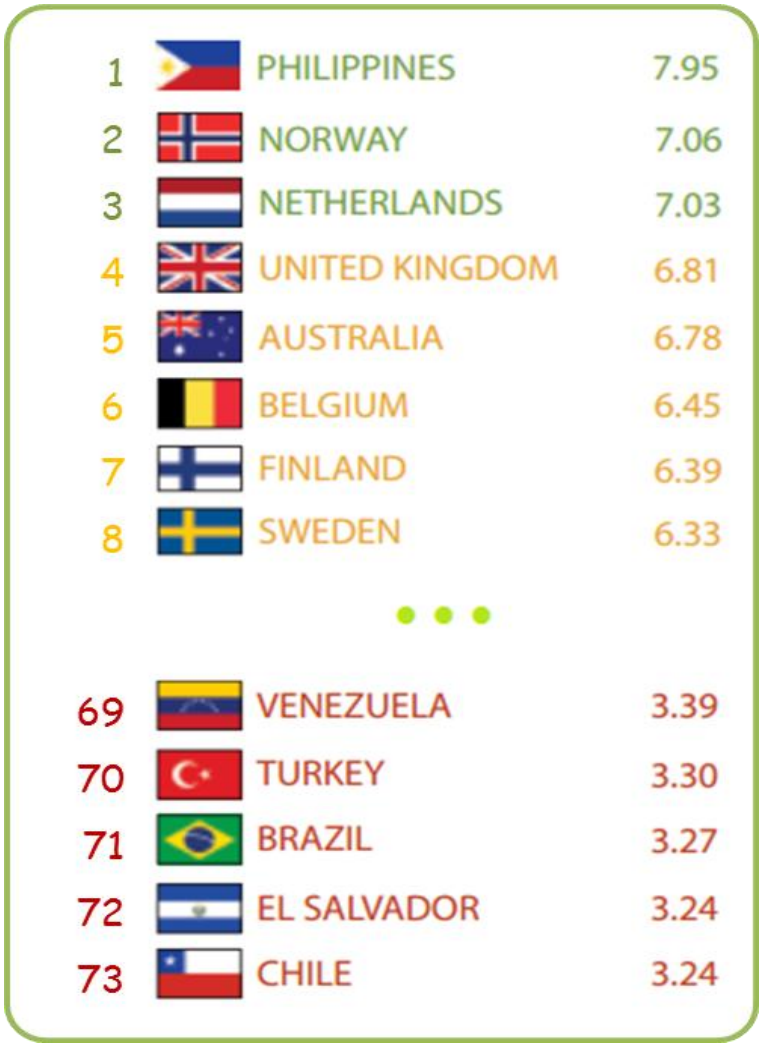


Figure 1. *Global English's Business English Index* (2013) partial ranking .

2. PROPOSAL

- The goal is to build up an ASR-based Pronunciation Training System for Brazilian-Portuguese-accented English;
- Able to provide online feedback regarding the pronunciation of the user;
- We named it *Listener*!

3. GAP

Similar tools are available for other languages, such as Japanese [3], French [4], Spanish [5] and Dutch [6], however, for BP, there is still a gap to be explored.

4. Research Hypothesis

The research hypothesis states that it is possible to build up an effective Pronunciation Training System through:

(i) an error classification list that takes into account grapho-phonic-phonological transfer from L1 to L2;
(ii) an acoustic model that contains speech data from both native and L2 English speakers;
(iii) a pronunciation dictionary which includes the transcription of the mispronunciations that learners are likely to make;
(iv) and a language model befitting the syntax of the learner.

5. A Brief Review of Automatic Speech Recognition through Hidden Markov Models (HMM)

- Noisy-channel Metaphor [7]: the recognition system tries to estimate, for a language $\mathcal{L}$, given a certain acoustic input *O*, what the most likely sentence $\hat{W}$ out of all sentences *W* is:

$$\hat{W} = \underset{W \in \mathcal{L}}{argmax}\ P(W|O) \quad \therefore \quad \hat{W} = \underset{W \in \mathcal{L}}{argmax}\ \underbrace{P(O|W)}_{\text{AM}}\ \underbrace{P(W)}_{\text{LM}}$$

ESTIMATED BY: AM LM

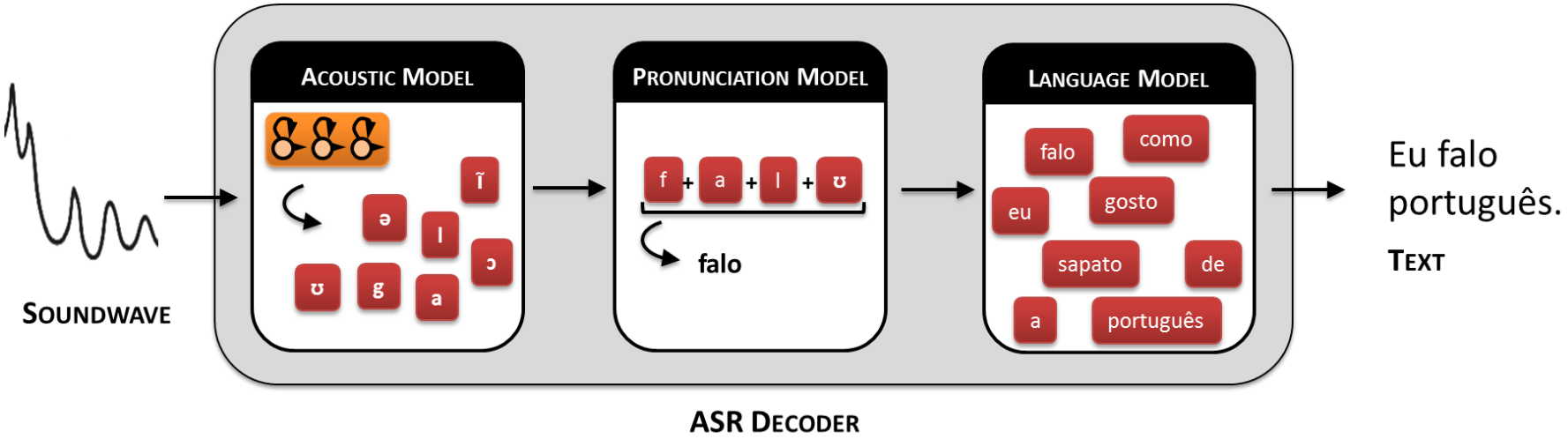


Figure 2. *Architecture of an HMM Automatic Speech Recognition System.*

6. METHOD

- The General American (GA) is considered the standard accent for the recognition system.

- The engine *Julius* [8] is used as the basis of the speech recogniser.

- Nine types of mispronunciations were selected, according to Zimmer et al. (2009) and Zimmer (2004).

Table 1. Selected mispronunciations for the Pronunciation System.

SELECTED MISPRONUNCIATIONS		
NO.	**DESCRIPTION**	**EXAMPLE**
1	Syllable simplification	[st] → [ist] ['istart]
2	Consonant change (substitutions)	[θ] → [s], [f] or [t] "think" → ['fɪŋk]
3	Deaspiration of voiceless plosives in initial or stressed position	[kh] → [k] "cup" → ['kʌp]
4	Terminal devoicing in word-final obstruents	[z] → [s] "does" → ['dʌs]
5	Delateralization and rounding of lateral liquids in final position	[l] or [ɬ] → [ʊ] "feel" → ['fiʊ]
6	Vocalization of final nasals	[ɪm] → [ĩ] "him" → ['hĩ]
7	Velar consonantal paragoge	[ŋ] → [ŋg] "sing" → ['sɪŋg]
8	Vowel assimilation	[æ] → [ɛ] "bad" → ['bɛd]
9	Interconsonantal epenthesis (-ed morpheme)	[d] or [t] → [id] or [ed] "danced" → ['dænsed]

- The Acoustic Model (AM) will be built up, in a pooled fashion, based on three corpora:
  - English native speakers: *TIMIT Acoustic-Phonetic Continuous Speech Corpus* [11];
  - English L2 learners: *COBAI - Corpus Oral Brasileiro de Aprendizes de Inglês* [12] and Listener corpus of phonetically balanced sentences (to be compiled).

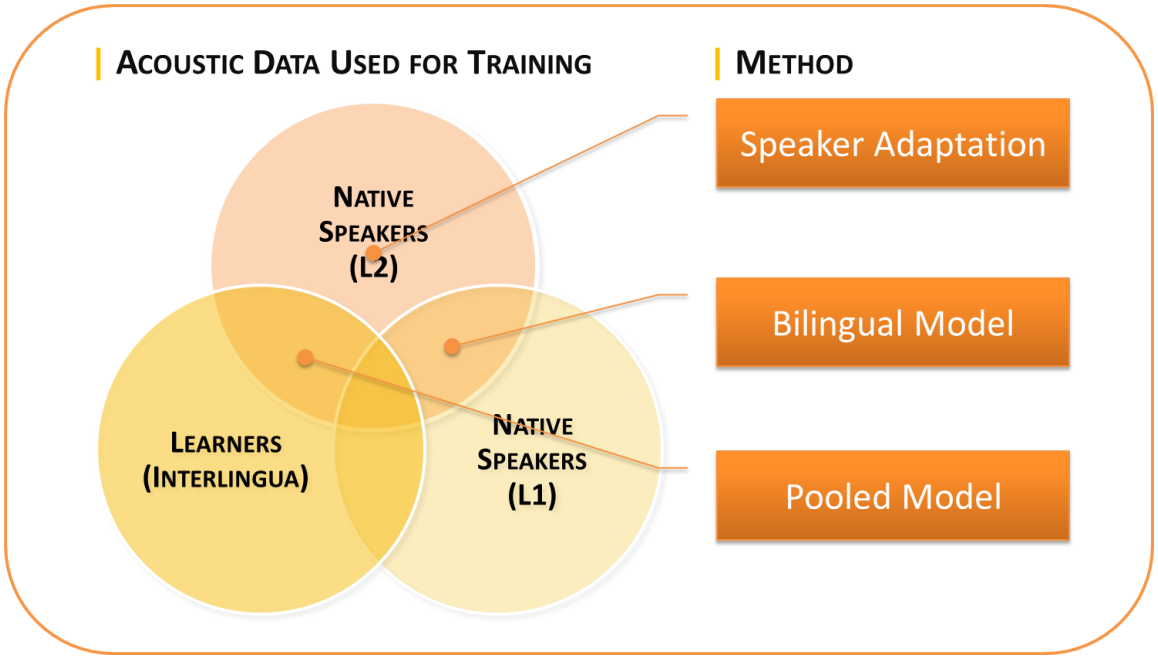


Figure 3. Methods for adapting the Acoustic Model (AM) to non-native speakers.

- *CMU Pronouncing Dictionary* ' [13] will serve as the basis for the Pronunciation Model (MP).
- Mispronunciations of the learners will be added to the dictionary, manually and also automatically, by machine learning algorithms, such as *Transformation-Based Learning* (Brill, 1995).
- The Language Model (LM) will be compiled over 108,943 articles from Simple English Wikipedia [15].

7. Evaluation

Word Error Rate (WER), Character Error Rate (CER) and confusion matrices will be the measurements used to evaluate the performance of the recogniser. These metrics will be applied to both the corpora used through a ten-fold stratified cross-validation technique.

8. Prototype Version

- Only syllable simplification cases were considered.

- Acoustic Model – ~3h30min of speech data:
  - An excerpt of COBAI - *Corpus Oral Brasileiro de Aprendizes de Inglês* [13];
  - A corpus of induced errors, designed exclusively for this prototype
- Pronunciation Model – Extracted from COCA 5,000 word list:
  - 1,855 words showed context prone to syllable simplification, therefore were selected.
  - Variants were produced through a set of 20 rules in an iterated fashion.

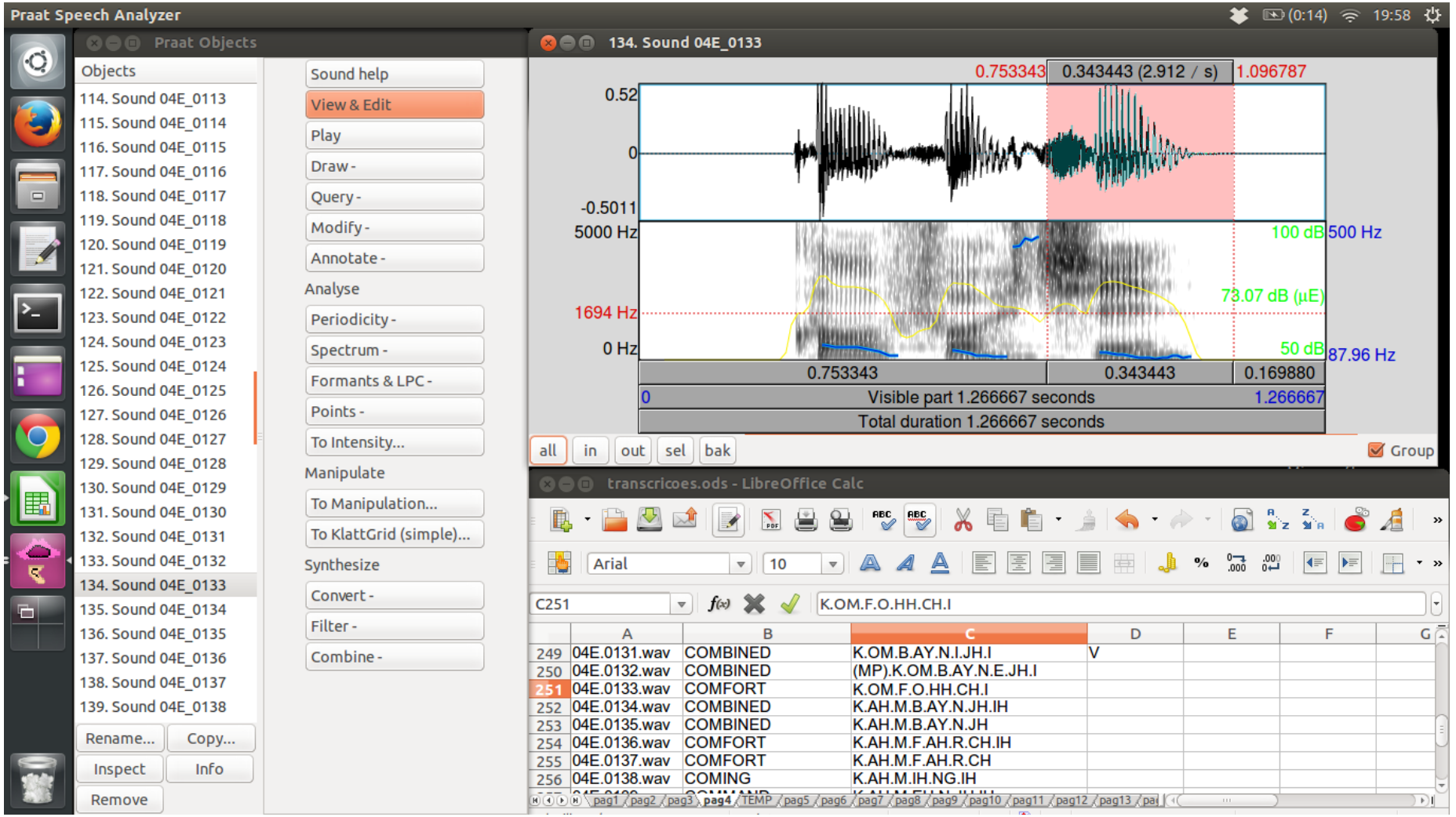
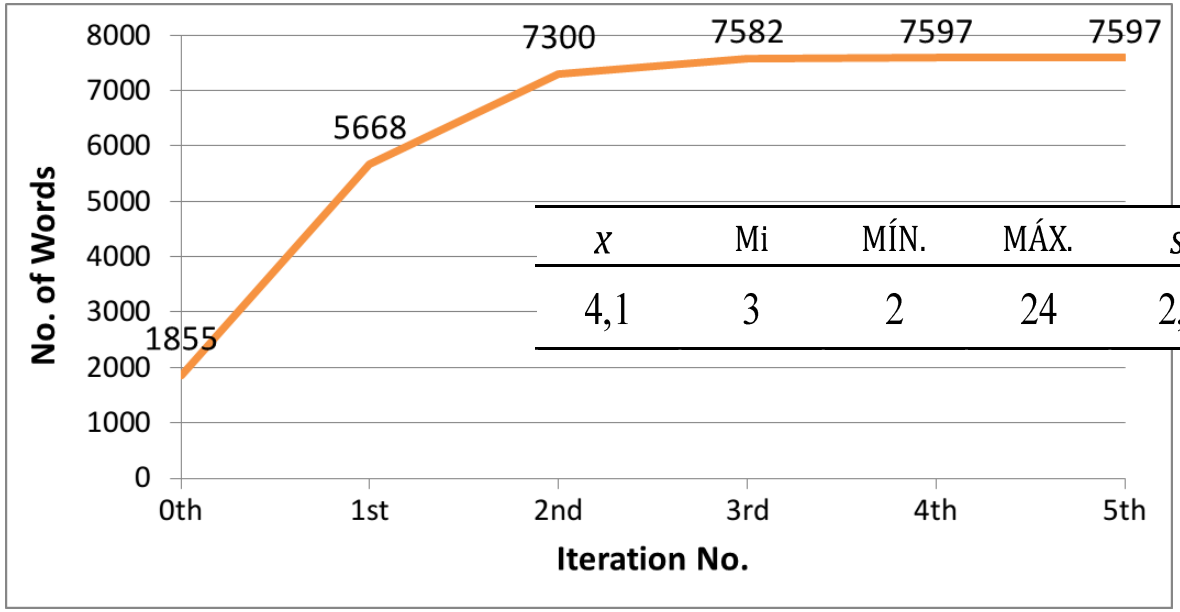


Figure 4. Workspace used for data transcription.

- We achieved 7,957 variants applying those rules!

Graph 1. Number of words in the pronunciation model per iteration.



x	Mi	MÍN.	MÁX.	s	σ
4,1	3	2	24	2,5	6,4

- Not quite good! The dictionary might overgrow in the complete version of the pronunciation system.

- Concerning recognition:
  - we achieved 61% word correctness, in a isolated word task, and 78% in the same task with forced alignment (via Viterbi algorithm).

- We are developing the website interface:

http://nilc.icmc.usp.br/listener

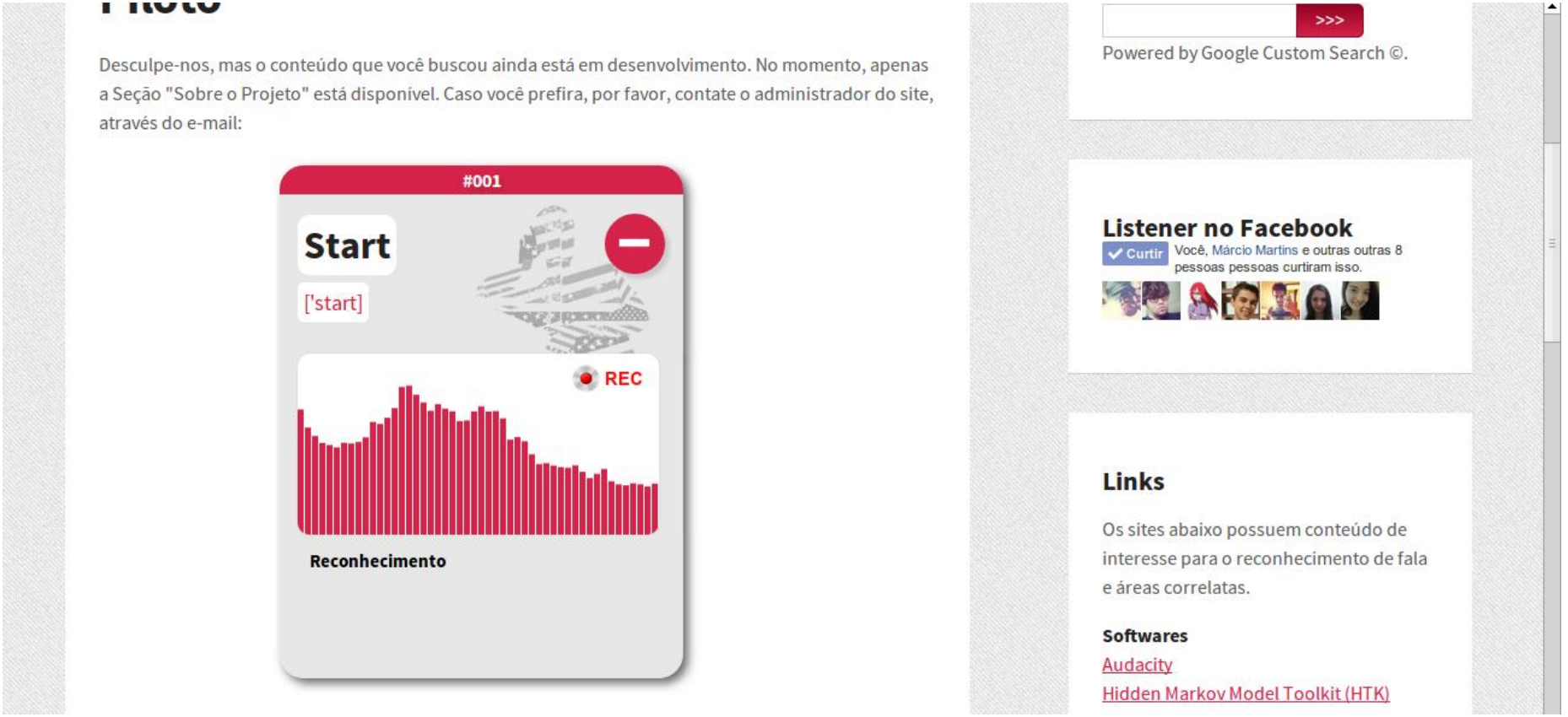


Figure 5. Preview of the web interface -- Site overview, audio recording with frequency spectrum.

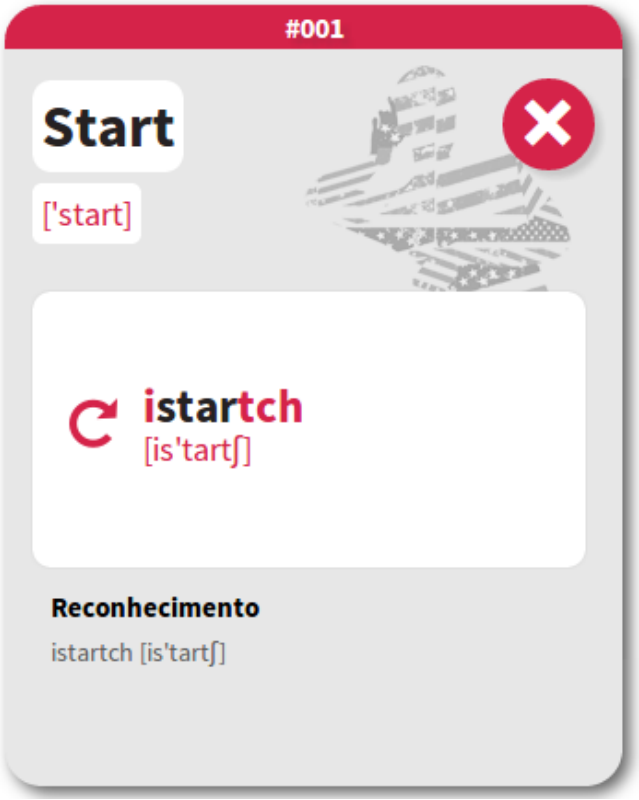


Figure 6. Preview of the web interface -- Word recognized and transcription.

Figure 7. Preview of the web interface -- Feedback of the pronunciation quality.

REFERENCES

[1] Education First (2012) English Proficiency Index 2012. Retrieved from: http://www.ef.com.br/epi/downloads. May, 2013.

[2] GlobalEnglish (2012) The Business English Index 2012 Report: Analyzing the Trends of Global Readiness for Effective 21st-Century Communication. Brisbane: GlobalEnglish.

[3] Tsubota, Y., Dantsuji, M., & Kawahara, T. (2004). An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors. ReCALL, 16, 173-188.

[4] Genevalogic. (2006). SpeedLingua - Language Learning Accelerator. Retrieved from: http://www.speedlingua.com/
Speedlingua. Sept, 2013.

[5] Reis, J., & Hazan, V. (2011). Speechant: a vowel notation system to teach English pronunciation. ELTJournal, 66/2, pp. 156-165.

[6] Strik, H., Doremalen, J., & Cucchiarini, C. (2008). A CALL system for practicing speaking . Proceedings of the XIIIth Int. CALL Conference, (pp. 123-125). Antwerp.

[7] Jurafsky, D.; Martin, J. (2009) Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. London, Pearson Education Ltd.

[8] Lee, A.; Kawahara, T. (2009) Recent Development of Open-Source Speech Recognition Engine Julius. In: Proceedings APSIPA ASC 2009 - Asia-Pacific Signal and Information Processing Association, pp. 131-137.

REFERENCES

[9] Zimmer, A., Silveira, R., & Alves, U. (2009). Pronunciation Instruction for Brazilians: Bringing Theory and Practice Together. Newcastle: Cambridge Scholars.[10]

[10] Zimmer, M. (2004). A Transferência do Conhecimento Fonético-Fonológico do Português Brasileiro (L1) para o Inglês (L2) na Recodificação Leitora: Uma Abordagem Conexionista. Dissertação de Doutorado. Porto Alegre: Pontifícia Universidade Católica do Rio Grande do Sul.

[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren (1990) DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology.

[12] Mello, H., Avila, L., Neder-Neto, T., & Orfano, B. (2012). LINDSEI-BR: an oral English interlanguage corpus. Proceedings of the VII GSCP International Conference: Speech and Corpora. 1, pp. 85-86. Florença, Itália: Firenze University Press.

[13] Weide, H. (1998). The CMU Pronouncing Dictionary. Acesso em 10 de Setembro de 2013, disponível em http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[14] Brill, E. (1992). A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.

[15] Simple English Wikipedia contributors (2014). Simple English Wikipedia [Internet]. Acesso em 20 de Janeiro de 2014. Disponível em: http://simple.wikipedia.org/w/index.php?title=Simple_English_Wikipedia&oldid=4705721.

[16] Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer, 5.3.62. Acesso em 4 de Janeiro de 2014, disponível em http://www.praat.org/