Listener: A prototype system for automatic speech recognition and evaluation of Brazilian-accented English

Gustavo A Mendonça[*][†] and Sandra M Aluisio

Abstract

First part title: Text for this section.

Second part title: Text for this section.

Keywords: pronunciation training; non-native speech recognition; natural language processing

Introduction

According to the International Monetary Found (IMF) [?], in 2015, Brazil as the seventh largest economy in the world with a GDP of US$ 2.34 trillions. A survey by The Economist (2013) says that, since 2009, the growth of BRICS accounts for 55% of the entire world economy growth. The current economic scenario is extremely favourable for Brazil to increase its global influence; however with regard to the ability to communicate globally, Brazil occupies a much more modest position.

In 2015, Brazil ranked 41st out of 70 countries in the English Proficiency Index (EF-EPI) [1], classified among countries with low English proficiency, with 51.05 points. Scandinavian countries led the very high proficiency rankings, with Sweden (70.94) in the first position, Denmark (70.05) in third the spot and Norway (67.83) in fourth. Brazil performance was close to several other Latin America countries, such as Peru (52.46), Chile (51.88), Ecuador (51.67), Uruguay (50.25) and Colombia (46.54). The only exception in Latin America was Argentina that, despite the recent great depression was ranked 15th, being classified as high proficiency, with a score of 60.26.

The EF-EPI bands are aligned to the Common European Framework of Reference for Languages (CEFR) in the following way: the very high proficiency band corresponds to CEFR level B2; very low proficiency to A2; high, moderate and low proficiency bands to B1 with different punctuations. In case, Brazil's low proficiency rank is analogous to the CEFR level B1, that describes an independent language user with the intermediate communication skills:

Table 1 CEFR reference level description for B1.

#	Communication skills
1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.
2	Can deal with most situations likely to arise while traveling in an area where the language is spoken.
3	Can produce simple connected text on topics that are familiar or of personal interest.
4	Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

As one might notice, the B1 level describe someones who is usually able to understand familiar matters, deal with traveling situations, describe personal experiences and plans, and produce simple texts about subjects of personal interest. Needless to say, this is a very restricted communicative competence, which limits English usage primarily to the personal domain.

With respect of Business English proficiency, Brazil performance is even more concerning. On the Business English Index (BEI) of 2013 [2], Brazil reached the 71st position out of 77 countries analyzed. We attained a score of 3.27 points, in a scale from 1 to 10, being placed at the "Beginner" range, the lowest range considered by the index. We were close to countries such as El Salvador (3.24), Saudi Arabia (3.14) and Honduras (2.92) which up until recently had experienced civil wars or dictatorship governments. BEI describes individuals at the beginner level as those who "can read and communicate using only simple questions and statements, but can't communicate and understand basic business information during phone calls". Again,

[*] Correspondence: gustavoauma@gmail.com
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, Brazil
Full list of author information is available at the end of the article
[†] Equal contributor

we can see that this is a very limited linguistic competence, that would not allow one not even to perform the most elementary day-to-day task in a company or industry work environment.

Given this scenario, it is clear that we desperately need to improve English language proficiency among Brazilians. This project seeks to be an initial step towards this direction. We developed a prototype system for automatic speech recognition and evaluation of Brazilian-accented English, called *Listener*, which is capable of recognizing utterances in Brazilian-accented English and identifying which are the mispronunciations. The system is based on an Automatic Speech Recognition system which makes use of forced alignment, *HMM/GMM* acoustic models, context free grammars and multipronunciation dictionaries.[1]

Automatic Speech Recognition

Overview

Automatic Speech Recognition (ASR) can be defined as the task of converting spoken language into readable text by computers in real-time [3].

Speech is certainly the most natural human way of communication. Allowing people to interact with their gadgets through voice may greatly improve the user-experience, especially in a world which is becoming more and more mobile-oriented. ASR nowadays is present in many widely-used applications, such as personal assistants, speech-to-text processing, domotics, call routing, etc.

All state-of-the-art paradigms in ASR are stochastic and they basically try to solve one single equation, which is called the fundamental equation of *ASR*. It can be described as follows. Let O be a sequence of observable acoustic feature vectors and W be a word sequence, the most likely word sequence $W*$ is given by:

$$W* = \arg\max_{W} P(W|O) \tag{1}$$

To solve this equation straightforwardly, one would require a discriminative model capable of estimating the the probability of W directly from a set of observations O [4]. If we apply the Bayes' Theorem we obtain the following equivalent equation:

$$W* = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} \tag{2}$$

which is suitable for a generative model. For a single audio input, the probability of the observable acoustic feature vectors $P(O)$ is a constant and, therefore, might be discarded, in such way that we end up with:

$$W* = \arg\max_{W} P(O|W)P(W) \tag{3}$$

$P(O|W)$ is the conditional probability of an observable acoustic feature vector given a word sequence, is calculated by an acoustic model. In turn, $P(W)$, the *a priori* probability of words is reckoned by a language model or through context free grammars.

Acoustic Model

Acoustic Models (AM) are used within speech recognition to map the acoustic parameters of into phonemes. AMs are estimated through supervised training over a transcribed speech corpus – often with the Forward-Backward algorithm by modeling phones via Hidden Markov Models (HMM) [5]. Markov models are very suitable for the statistical description of symbol and state sequences [6]. Within Markov processes, systems are assumed to be memoryless, that is, the conditional probability of future states is only dependent on the present state. To put it another way, the current state does not depend upon the sequence of events that preceded it. Hidden Markov Models (HMM) are just a special type of Markov processes which contain hidden states.

HMMs are the most widespread models used in ASR [7]. They can be formally described as a 5-tuple $\lambda = (Q, O, \Pi, A, B)$. $Q = \{q_1, q_2, q_3, ..., q_N\}$ represents a set of hidden N states. $O = \{o_1, o_2, o_3, ..., o_T\}$ is a set of T observations taken from time $t = 1$ to $t = T$. At each time t it is assumed that the system will be at a specific state q, which is hidden, and only the observations o are directly visible. $\Pi = \{\pi_i\}$ is a vector with the initial state probabilities, such that

$$\pi_i = Pr(q_i), t = 0 \tag{4}$$

In addition, $A = [a_{ij}]$ is matrix with the state transition probabilities so that

$$a_{ij} = P(q_t = j | q_{t-1} = i), 1 \leq, i, j \leq N \tag{5}$$

and $B = [b_{jt}]$ is a matrix with the emission probability of each state. Assuming a *GMM* to model the state emission probabilities – the so-called GMM/HMM model in ASR; we can define that, for a state j, the probability $b_j(o_t)$ of generating o_t is given by

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_{js}} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \tag{6}$$

[1] All files, resources and scripts developed are available at the project website. Due to copyright reasons, the corpora used for training the acoustic models cannot be made available: http://nilc.icmc.usp.br/listener

where γs is a stream weight, with default value is one, M_{js} is the number of mixture components in state j for stream s, c_{jsm} is the weight of the m^{th} component and $\mathcal{N}(\cdot; \mu_{jsm}, \Sigma_{jsm})$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

$$\mathcal{N}(o; \mu, \Sigma) = (\sqrt{(2\pi)^n \, |\Sigma|})^{-e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)}} \tag{7}$$

where n is the dimensionality of o. The following constraints apply to the model:

$$a_{ij} \geq 0 \tag{8}$$

that is, the probability of moving from state from any state i to j is not null, and the sum of all state transitions add up to unity:

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i \tag{9}$$

Context Free Grammars

Context Free grammarsPronunciation models have the role of linking phones from the AM to words in the LM. To sum up, pronunciation models are basically lexica with words and their corresponding phonetic transcriptions, according to a given transription.

Pronunciation Model

Pronunciation models have the role of linking phones from the AM to words in the LM. To sum up, pronunciation models are basically lexica with words and their corresponding phonetic transcriptions, according to a given transription.

Multipronunciation dictionaries are a type of pronunciation model where pronunciation variants are explicitly added to the lexicon of the ASR [8].

1 Brazilian-Accented English

Table 2 Mispronunciation types selected for the prototype system with examples of the expected pronunciation and the one with negative transfer from L1 to L2.

#	Description	Example	Expect.	Mispron.
1	Initial epenthesis	school	[skuːl]	[iskuːl]
2	Coda epenthesis	dog	[dɑːg]	[dɑːgi]
3	Terminal devoicing	does	[dʌz]	[dʌs]
4	Consonantal change	think	[θıŋk]	[fıŋk]
5	Deaspiration in plosives	tea	[tʰiː]	[tiː]
6	Vocalization of laterals	well	[wɛl]	[wew]
7	Vocalization of nasals	beam	[biːm]	[bĩ]
8	Vowel change	put	[pʰʊt]	[pʰʌt]
9	Velar paragoge	wing	[wıŋ]	[wıŋg]

2 Materials and Methods

2.1 Architecture of Listener

For speech recognition – or, in fact, any supervised machine learning task – the best scenario for training a model is when you have a huge amount of data which is large and diverse enough so that it fully represents population. However, this is usually not the case. There is not much data available for training acoustic models for many languages.

To build a speech corpora, one must first carry out an analysis of the phones in a language, in order to examine how sounds are distributed and which phonological phenomena might be involved. Then define a corpus to be read by subject which is representative of the. Contact the subjects and coordinate the recordings, making sure that the corpora will be sociolinguistcally representative in terms of sex, age, dialect and social strata, etc. Postprocess the audio files, by splitting, organizing, checking the audio quality.

As one might notice, compiling speech corpora is something that is not only complex, but also quite time consuming – and therefore costly. Obviously, the scenario is even worse for non-native speech recognition. Due to this data scarcity, one can find in the literature for *CAPT* several approaches which make use of data from acoustic models with data from different sources.

We can group these models into four types:

Acoustic models for pronunciation training can be divided into three groups, according to the source of the data.

Materials and Methods

Multipronunciation Dictionary

Multipronunciation dictionaries are a type of pronunciation model where pronunciation variants are explicitly added to the lexicon of the ASR [8].

Content

Text and results for this section, as per the individual journal's instructions for authors.

Section title

Text for this section . . .

Sub-heading for section

Text for this sub-heading . . .

Sub-sub heading for section

Text for this sub-sub-heading . . .

Sub-sub-sub heading for section Text for this sub-sub-sub-heading …In this section we examine the growth rate of the mean of Z_0, Z_1 and Z_2. In addition, we examine a common modeling assumption and note the importance of considering the tails of the extinction time T_x in studies of escape dynamics. We will first consider the expected resistant population at vT_x for some $v > 0$, (and temporarily assume $\alpha = 0$)

$$E\big[Z_1(vT_x)\big] = E\bigg[\mu T_x \int_0^{v\wedge 1} Z_0(uT_x)\exp\big(\lambda_1 T_x(v-u$$

If we assume that sensitive cells follow a deterministic decay $Z_0(t) = xe^{\lambda_0 t}$ and approximate their extinction time as $T_x \approx -\frac{1}{\lambda_0}\log x$, then we can heuristically estimate the expected value as

$$\begin{aligned} E\big[Z_1(vT_x)\big] &= \frac{\mu}{r}\log x \int_0^{v\wedge 1} x^{1-u} x^{(\lambda_1/r)(v-u)}\, du \\ &= \frac{\mu}{r} x^{1-\lambda_1/\lambda_0 v}\log x \int_0^{v\wedge 1} x^{-u(1+\lambda_1/r)} du \\ &= \frac{\mu}{\lambda_1-\lambda_0} x^{1+\lambda_1/rv}\left(1-\exp\left[-(v\wedge 1)\left(1+\frac{\lambda_1}{r}\right)\log x\right]\right) \qquad (10) \end{aligned}$$

Thus we observe that this expected value is finite for all $v > 0$ (also see [9–13]).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Text for this section …

Acknowledgements

Text for this section …

References

1. EducationFirst: EF English Proficiency Index 2015. Education First Ltd., Lucerne (2015)
2. GlobalEnglish: The 2013 Business English Index & Globalization of English Report, p. 15. Pearson Always Learning, Pearson (2013)
3. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, 1st edn., p. 980. Prentice Hall PTR, Upper Saddle River, NJ, USA (2001)
4. Gales, M., Young, S.: The application of hidden markov models in speech recognition. Foundations and trends in signal processing **1**(3), 195–304 (2008)
5. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
6. Fink, G.A.: Markov Models for Pattern Recognition: from Theory to Applications. Springer, London (2014)
7. Juang, B., Rabiner, L.: In: Brown, K. (ed.) Automatic Speech Recognition - A Brief History of the Technology, p. 24. Elsevier, Amsterdam (2005)
8. Strik, H.: Pronunciation adaptation at the lexical level. In: ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition (2001)
9. Koonin, E.V., Altschul, S.F., Bork, P.: Brca1 protein products: functional motifs. Nat Genet **13**, 266–267 (1996)
10. Kharitonov, S.A., Barnes, P.J.: Clinical Aspects of Exhaled Nitric Oxide. in press
11. Zvaifler, N.J., Burger, J.A., Marinova-Mutafchieva, L., Taylor, P., Maini, R.N.: Mesenchymal cells, stromal derived factor-1 and rheumatoid arthritis [abstract]. Arthritis Rheum **42**, 250 (1999)
12. Jones, X.: Zeolites and synthetic mechanisms. In: Smith, Y. (ed.) Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore, pp. 16–27 (1996). Stoneham: Butterworth-Heinemann
13. Margulis, L.: Origin of Eukaryotic Cells. Yale University Press, New Haven (1970)

Figures

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Figure legend text.

Tables

Table 3 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.