RESEARCH

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) [1], 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) [2], 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 its 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 [3], 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

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

Automatic Speech Recognition (ASR) can be defined as the task of converting spoken language into readable text by computers in real-time [4].

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 straightforwadly, one would require a discriminative model capable of estimating the the probability of W directly from a set of observations O [5]. 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.

Materials and Methods

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.

Pronunciation Model

Pronunciation models are lexica with words and their corresponding phonetic transcriptions, according to a given convention. In other words, pronunciation models have the role of linking phones from the acoustic model to words defined in the language model. For speech recognition purposes, phonetic gammabets like ARPAbet or SAMPA are often employed to avoid problems with data formatting or encoding. In ARPAbet, phones are converted into sequences of ASCII

[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

characters, in such a way that a word like "speech" ['spiːtʃ] becomes [s p iy1 ch] [6].

For non-native speech recognition, multipronunciation dictionaries are often employed in order to address phenomena of negative-transference from the L1 to L2. These dictionaries are a type of pronunciation model where pronunciation variants are explicitly added to the lexicon of the ASR [7]. For building the pronunciation model for Listener, the literature on pronunciation training was analyzed and transformation rules were defined based on the most common mispronunciations among Brazilians. The pronunciation model was inspired by several works for pronunciation training, focused on Brazilian-accented English [8**?** , 9]. In total, we gathered 13 mispronunciation patterns for Listener, the full list with examples can be found in Table 2.

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	Th-fronting	think	[θɪŋk]	[fɪŋk]
5	Palatalization	teen	[tʰiːn]	[tʃiːn]
6	Deaspiration in plosives	tea	[tʰiː]	[tiː]
7	Vocalization of laterals	well	[wɛl]	[wew]
8	Vocalization of nasals	beam	[biːm]	[bĩ]
9	Velar paragoge	wing	[wɪŋ]	[wɪŋg]
10	Consonantal change	think	[θɪŋk]	[fɪŋk]
11	Vowel change	put	[pʰʊt]	[pʰʌt]
12	General deletion	foot	[fʊt]	[fʊ]
13	General insertion	work	[wɜːrk]	[wɜːrks]

All linguistic contexts described by Zimmer [8] were converted into transcription rules in order to generate the variants for the pronunciation dictionary. The full list of rules can be found in the project's website. A sample of these rules can be found in Figure 1. These transcription rules are then applied to a base dictionary in order to append it with new pronunciation variants.

It is worth noticing that, in terms of context, there is often overlapping among rules. For instance, in a word like "think" [th ih ng k], there is a rule for converting [th] into [f], another one for converting it into [s], or [t], etc. There are even rules which create context for other ones to apply, for instance, in "boat" [b ow t], if epenthesis takes place, generating [b ow t ih], then [t] could undergo consonantal change/palatalization, thus producing [b ow ch ih].

Figure 1 Building the pronunciation model. Pseudocode with the rules for generating pronunciation variants (sample).

 # Initial epenthesis
if [s p] in initial position → [iy s p] # sport
if [s t] in initial position → [iy s t] # start
if [s k] in initial position → [iy s k] # skate
if [s m] in initial position → [iy s m] # small
if [s n] in initial position → [iy s n] # snake
...
 # Coda epenthesis
if [p] in final position → [p ih] # stop
if [b] in final position → [b ih] # bob
if [t] in final position → [t ih] # boat
if [d] in final position → [d ih] # and
if [k] in final position → [k ih] # book
if [g] in final position → [g ih] # dog
...
if [m] and ortho ends in <me> → [m ih] # time
if [s] and ortho ends in <ce> → [s ih] # nice
...
 # Th-fronting
if [th] → [f] # think
if [th] → [s] # think
if [th] → [t] # think
...
 # Palatalization
if [t iy] → [ch iy] # teen
if [t ih] → [ch iy] # poetic
...
 # Vocalization of nasal consonants
if [iy m] in final position → [im] # him
if [ae n] in final position → [em] # can
...

To make sure that all pronunciation variants are generated, the rules are run inside a while loop, which iterates over each word in the base dictionary generating and adding these new pronunciation to the dictionary; and the loop only stops when there are no new variants.

For the pilot system of Listener, we used as a base dictionary the CMUdict [6], which contains over 134,000 entries and their pronunciations in American English. However, in the test sets for Listener there are just 1,841 unique words, so only these were considered in this experiment. These transcription rules were run over these 1,841 unique words and 8,457 new pronunciation variants were generated (=10,298 entries in the final dictionary). In such a way, the average pronunciation per word is 5.6.

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) [10]. Markov models are very suitable for the statistical description of symbol and state sequences [11]. 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 [12]. 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

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

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:

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

For building Listener, HMM/GMM were applied to represent triphones. A triphone is a contextual phone, i.e. it is a phonetic unit of analysis which, for a given phone p, takes into account the previous phone $p-1$ and following one $p+1$. For instance, in a word like "speech" [s p iy ch], the phone [iy] would correspond to the triphone $[_{\text{p}}\text{iy}_{\text{ch}}]$, indicating that [iy] occurs after a [p] and before a [ch]. The full of transcription of "speech" in triphones would be $[_{\#}\text{s}_{\text{p}}\ _{\text{s}}\text{p}_{\text{iy}}\ _{\text{p}}\text{iy}_{\text{ch}}\ _{\text{iy}}\text{ch}_{\#}]$, it still has the same number of phone, the only difference is that the phones are now defined context.

For estimating the values and probabilities of the HMM/GMM the CMU Sphinx Toolkit was used [13]. Particularly, the acoustic model was trained over several different corpora, which contained, in total, 40 hours of audio from native speakers of English or Brazilian Portuguese, as well as non-native data in Brazilian-accented English. The acoustic model was estimated considering a phonetic inventory of XX phones, containing 4,000 tied states and 16 gaussian densities per state. The last two values were defined based on a pilot experiment over a sample from the available corpora.

Context Free Grammars

Context Free Grammars are formal grammars in which every rule takes the form:

$$A \rightarrow \gamma \tag{10}$$

where A is a nonterminal and γ corresponds to a single or sequence of nonterminal or terminal symbol [14]. In speech recognition, Context Free Grammars were the first attempt to broaden speech recognition to a context larger than digits, letters and menu commands; but they were rapidly replaced by statistical Language Models, as the latter scale better and require much less manual work [5]. However, CFGs are still used when the user input is part of a limited set and the accuracy is more relevant than coverage. For instance, a CFG for defining a simple grammar for calling or contacting a friend can be represented as follows:

Results and Discussion

The system was evaluated on three different test sets. The first consists in a corpus of induced errors in isolated words (∼2 hours), which we will call *Induced*.

This corpus was recorded by a single male speaker with good proficiency of English, who induced pronunciation errors while reading prompts with isolated words in English. The recordings were made with a high-fidelity microphone in quiet room, in order to reduce background noise. The second test set is a corpus compiled through crowdsourcing specially for this project. It is called *Listener Corpus* and it contains native-speakers of Brazilian Portuguese reading pre-defined sentences in English. There corpus was recorded by 67 individuals (∼13 hours) through a website and the environment was not controled, the subjects used their own laptops and personal computers to do the recordings. There is usually a lot of noise in the channel as well as background noise (music, traffic, fan, animal sounds and so on). The third test set is a subset of the Listener Corpus which contains only prompts with isolated words, we will refer to it as *Listener (Isolated-Words)*.

The recognition results for the Induced test set can be found in Table 3.

Table 3 Recognition results for each phone in the Induced test set. Results are grouped by mispronunciation pattern, and the percentages for True Positives (TP) and Type-I/Type-II errors are shown.

#	Category	Counts	TP	Type-I	Type-II
0	Expected phone	2265	0.96	0.03	0.02
1	Initial epenthesis	38	1.00	0.00	0.00
2	Coda epenthesis	14	0.79	0.00	0.21
3	Terminal devoicing	0	-	-	-
4	Th-fronting	6	0.67	0.00	0.33
5	Palatalization	197	0.90	0.03	0.07
6	Deaspirtation in plosives	103	0.89	0.05	0.06
7	Vocalization of laterals	11	0.45	0.00	0.55
8	Vocalization of nasals	301	0.86	0.04	0.10
9	Velar paragoge	25	0.96	0.04	0.00
10	Consonantal change	304	0.90	0.06	0.04
11	Vowel change	300	0.92	0.04	0.04
12	General deletion	0	-	-	-
13	General insertion	260	0.97	0.03	0.00
Total/W Avg		3553	0.93	0.03	0.03
Total/Avg (wo Exp.)		1288	0.90	0.04	0.06

As one might observe, the overall phone recognition for the Induced corpus was 0.93. Considering only phones with pronunciation errors, the ratio of true positives was 0.90. The expected phones showed a true positives ratio of 0.96. Initial epenthesis was the mispronunciation which was recognized with the highest accuracy, all 38 cases in the corpus were correctly identified. The system was able to detect initial sequences of [iy s C], as in "stop" [iy s t ao p], with no losses. Following, the best recognition performance was found in cases of general insertion, for instance, when one adds an extraneous phone to the end of a word, e.g. "work" [w ah r k s]. The true positive rate for generation insertion was 0.97. were the ones which were identified with the velar paragoge was the mispronunciation pattern which was recognized in most. Cases of velar paragoge, as in "king" [k ih ng g] or [k ih ng g ih] were accurately inferred in 0.97 cases. The worst results were found for vocalization of laterals (0.45), followed by th-fronting (0.67). The lower performance for these mispronunciations patterns types might be due to the fact that these errors involve phones which are acoustically similar. It could also be due to the fact that there is less data for these cases, but further investigation is needed. The rate of Type-I error, or false positives, was very small. The highest value was found in cases of consonantal change, which had a Type-I error rate of 0.06. Type-II errors occurred more often, vocalization of laterals had a ratio of 0.55, th-froning of 0.33 and coda epenthesis had 0.21.

Results for the Listener test set, considering the entire corpus (sentences + isolated-words), are described in Table 4.

Table 4 Recognition results for each phone in the Listener Corpus (all). Results are grouped by mispronunciation pattern; the values for True Positives (TP) and Type-I/Type-II errors are presented.

#	Category	Counts	TP	Type-I	Type-II
0	Expected phone	9245	0.76	0.17	0.07
1	Initial epenthesis	34	0.12	0.26	0.62
2	Coda epenthesis	49	0.24	0.49	0.27
3	Terminal devoicing	0	-	-	-
4	Th-fronting	179	0.23	0.20	0.57
5	Palatalization	93	0.03	0.13	0.84
6	Deaspirtation in plosives	60	0.03	0.20	0.77
7	Vocalization of laterals	179	0.31	0.35	0.34
8	Vocalization of nasals	1017	0.37	0.29	0.35
9	Velar paragoge	1	1.00	0.00	0.00
10	Consonantal change	260	0.23	0.22	0.55
11	Vowel change	1366	0.37	0.18	0.45
12	General deletion	421	-	-	-
13	General insertion	131	0.30	0.27	0.44
Total/Avg		11669	0.63	0.19	0.19
Total/Avg (wo Exp.)		2424	0.31	0.21	0.48

As one can notice, the results for the Listener test set were much lower. The recognition performance for all categories of mispronunciations was severely reduced. The best results were found for the expected phones, which showed a true positives ratio of 0.76. The average true positive ratio, considering only the data with pronunciation errors, was 0.31. Such performance would certainly be unsuitable for Computer Assisted Pronunciation Training applications. This result shows that the performance of the models, in terms of recognition, is deeply affected by the characteristics of the corpora. The previous corpus was recorded in a quiet room, with a high-fidelity microphone, thus leading to audios with very good signal-to-noise ration. However, as the Listener Corpus was compiled through crowdsourcing, with barely no control on the recording settings, the data is too noisy. Considering that the speech database that was used for training the acoustic model contained mainly clean speech, the final model seems not to be able to generalize well on noisy data.

A subset of the Listener corpus, containing only isolated-words, was also analyzed. The purpose of this subset is to best compare the results with the Induced corpus. A summary of the value is presented in Table 5

Conclusions

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank the Instituto de Telecomunicações in Coimbra for for the support with the corpora for training the acoustic models.

Table 5 Recognition results for each phone in the Listener Corpus (Isolated-Words). Results are grouped by mispronunciation pattern; the values for True Positives (TP) and Type-I/Type-II errors are presented.

#	Category	Counts	TP	Type-I	Type-II
0	Expected phone	2265	0.45	0.03	0.02
1	Initial epenthesis	2	0.50	0.50	0.00
2	Coda epenthesis	5	0.60	0.20	0.20
3	Terminal devoicing	0	-	-	-
4	Th-fronting	11	0.36	0.09	0.55
5	Palatalization	10	0.20	0.00	0.80
6	Deaspirtation in plosives	4	0.00	1.00	0.00
7	Vocalization of laterals	20	0.50	0.20	0.30
8	Vocalization of nasals	142	0.61	0.16	0.23
9	Velar paragoge	1	1.00	0.00	0.00
10	Consonantal change	25	0.52	0.24	0.24
11	Vowel change	97	0.60	0.19	0.22
12	General deletion	12	-	-	-
13	General insertion	21	0.81	0.14	0.05
Total/Avg		2518	0.82	0.09	0.09
Total/Avg (wo Exp.)		253	0.57	0.18	0.25

References

1. Fund, I.M.: World Economic Outlook. International Monetary Fund, Washington, DC, USA (2015)
2. EducationFirst: EF English Proficiency Index 2015. Education First Ltd., Lucerne (2015)
3. GlobalEnglish: The 2013 Business English Index & Globalization of English Report, p. 15. Pearson Always Learning, Pearson (2013)
4. 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)
5. Gales, M., Young, S.: The application of hidden markov models in speech recognition. Foundations and trends in signal processing **1**(3), 195–304 (2008)
6. Weide, R.: The CMU Pronouncing Dictionary 0.7a. Carnegie Mellon University (2008). http://www.speech.cs.cmu.edu/cgi-bin/cmudict
7. Strik, H.: Pronunciation adaptation at the lexical level. In: ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition (2001)
8. Zimmer, M.: A Transferência do Conhecimento Fonético-Fonológico do Português Brasileiro (L1) Para O Inglês (L2) na Recodificação Leitora: Uma Abordagem Conexionista. Dissertação de Doutorado. Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre (2004)
9. Zimmer, M., Silveira, R., Alves, U.: Pronunciation Instruction for Brazilians: Bringing Theory and Practice Together. Cambridge Scholars, Newcastle (2009)
10. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
11. Fink, G.A.: Markov Models for Pattern Recognition: from Theory to Applications. Springer, London (2014)
12. Juang, B., Rabiner, L.: In: Brown, K. (ed.) Automatic Speech Recognition - A Brief History of the Technology, p. 24. Elsevier, Amsterdam (2005)
13. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J.: Sphinx-4: A flexible open source framework for speech recognition (2004)
14. Jurafsky, D., Martin, J.: Speech and Language Processing : an Introduction to Natural Language Processing Computational Linguistics, and Speech Recognition, 2ª edn. Prentice Hall, New Jersey, USA (2000)