JoSS Journal of Speech Sciences

DOI: 10.20396/joss.v10i00.15863

AUTOMATIC IDENTIFICATION OF SYNTHETICALLY GENERATED INTERLANGUAGE TRANSFER PHENOMENA BETWEEN BRAZILIAN PORTUGUESE (L1) AND ENGLISH (L2)

BORGES, Atos A. S.[1*]
RODRIGUES FILHO, Washington Luis P.[2]
ROCHA, Aratuza Rodrigues Silva[3]
CARVALHO[2], Wilson Júnior de Araújo[4]
LIMA JR., Ronaldo Mangueira[5]
BARBOSA, Fábio Rocha[6]

[1]Postgraduate Program in Electrical Engineering – Federal University of Piauí – ORCID: https://orcid.org/0000-0002-4135-8864
[2]Department of Electrical Engineering – Federal University of Piauí – ORCID: https://orcid.org/0000-0002-8604-7015
[3] Faculdade Afonso Mafrense – ORCID: https://orcid.org/0000-0002-1449-1918
[4]Postgraduate Program in Applied Linguistics – State University of Ceará – ORCID: https://orcid.org/0000-0003-1606-356X
[5]Postgraduate Program in Linguistics – Federal University of Ceará – ORCID: https://orcid.org/0000-0002-6027-3161
[6]Postgraduate Program in Electrical Engineering – Federal University of Piauí – ORCID: https://orcid.org/0000-0001-5473-0584

Abstract: *Transfer phenomena between Portuguese (L1) and English (L2) produced by Brazilian learners are well documented in the literature. However, the identification and classification of these processes are made mainly through transcriptions, a slow and laborious process done by specialized linguists. The rapid identification of these phenomena would be of great value for software doing proficiency placement tests and could be used in language schools, distance education, computer-assisted pronunciation training (CAPT) or by autodidacts and researchers. The present work analyzed possible techniques and tools that can be used in the automatic identification of some transfer processes. Data for the some grapho-phonic-phonological transfer were synthetically generated in the Google Translate™ TTS system. Then we tested three classification algorithms to perform the identification: k-Nearest Neighbor, Centroid Minimum Distance and Artificial Neural Networks. The results indicate that these techniques are of great value for Linguistics and for new software applications in language learning.*

Keywords*:* Grapho-phonic-phonological transfer; k-Nearest Neighbor; Centroid Minimum Distance; Artificial Neural Networks; language learning.

---

[1]* Corresponding author: atosborges00@ufpi.edu.br

1. Introduction

Pronunciation is one of the key elements that influence the mastery of a language. Especially in the process of learning a non-native language (L2)[2], pronunciation is a central concern for those who want to communicate effectively. During the learning of a new language, an interphonology emerges. Interphonology is a linguistic system different from both that of the native language (L1) and of the L2, with both languages influencing such system (1). Students in the process of learning an L2 transfer some of their knowledge of the L1 to the new language due to the already established structure of the L1, which might jeopardize communication at times. This phenomenon, when manifested in speech or oral reading, is called grapho-phonic-phonological knowledge transfer (2). The term grapho-phonic-phonological contemplates not only the transference of phonetic-phonological knowledge (3) but also the transference of the grapheme-phoneme relationships of one language to the other (4–6). In the case of this work, we focused on the grapheme-phoneme knowledge transfer between Brazilian Portuguese (BP) as L1 to English as L2.

When the leaner finds an unknown structure in the L2, they use strategies to adapt the L2 to the closest structure already known in the L1. These phenomena can be manifested in many ways, including the change, deletion, or insertion of a segment (vowel or consonant), as well as changes at the prosodic level, such as changes in word stress, sentence stress, rhythm, and intonation. All these processes can cause misunderstandings and problems in communication. Therefore, L2 learners must overcome such phenomena in the process of developing proficiency and fluency in the new language.

There are many conditions in which these phenomena are more susceptible or attenuated, such as the orthographic depth of the languages (7), time of first exposure (8), formal education in the L2 (9), and the proficiency of the learners (10). All these factors might have a role in the occurrence of these transfer processes.

Although there is a vast literature for grapho-phonic-phonological transfer between Brazilian Portuguese and English as a Foreign Language (11), there is a shortage of works aimed at recognizing and classifying these processes in an automated way. Most studies carry out transfer identification through audio transcription, an arduous and costly task done by hand. Only two works were found proposing forms of automated identification. The first was a categorization of BP speakers by a Self-Organizing Map (SOM) regarding the transfer of stress patterns between BP-L1 and English-L2 (12). The second also aimed to identify transfer processes from BP to English-L2 of Brazilian students using a Multi-Layer Perceptron (MLP) neural network (13). A faster way to identify these processes would be truly valuable for linguists conducting these studies. In addition, a system capable of identifying deviations in pronunciation would also be useful for language learning software, helping language schools, mobile app developers and autodidacts. It has already been shown that these phenomena can even be used to predict the scores at the listening section of Brazilians in the TOEIC (Test of English for International Communication) (14).

Therefore, this work sought to investigate a few possibilities available to recognize the occurrence of some of these phenomena automatically through software identification techniques. Five phenomena related to grapheme-phoneme correspondences were chosen from the literature of BP transfer to English-L2: a) the deletion of initial [h] in words beginning with <h>, as in ‘humorist’ pronounced as [ˈjuməɻɪst] or [ˈuməɻɪst]; b) the deletion of initial [h] with a change of

[2] In this paper we do not distinguish ‘foreign language’ and ‘second language’, using the umbrella acronym L2 for any type of non-native language.

[aj] to [i] in words beginning with <hy>, such as ‘hydrant’ pronounced as [ˈidɽənt]; c) changing [aj] to [i] while keeping the pronunciation of initial [h] in words beginning with <hy>, as in ‘hydrant’ pronounced as [ˈhidɽənt]; d) pronouncing silent <k> with the insertion of an epenthetic [i] in words beginning with <kn>, such as ‘knife’ pronounced as [kiˈnajf]; and e) voicing /s/ when <s> appears between two vowels, as in ‘case’ pronounced as [kejz]. To approach the problem in diverse ways, tests were conducted using three different classification algorithms: k-Nearest Neighbor (kNN), Centroid Minimum Distance (CMD) and Artificial Neural Networks (ANNs). To collect the data, samples of native-like pronunciation and of BP-influenced pronunciation were synthetically generated in Google Translate™ text-to-speech system.

The first hypothesis we assumed was that the Google Translate™ text-to-speech system is able to simulate the grapho-phonic-phonological transfer phenomena. This way it would be possible to synthetically compose the dataset needed for the classification without any human tests in this first phase. The second hypothesis tested was that even the classic classifiers, such as kNN, CMD, and ANNs with simple and fast architectures are able to correctly identify the phenomena. If so, it would be possible to create systems capable of doing the identification task but still maintaining simplicity and low processing power, ideal for online and mobile applications.

2. Data collection

Five widely known transfer phenomena were chosen to be collected in the Google Translate™ TTS system. These phenomena are well documented and commonly found in the pronunciation of Brazilian beginning learners of English (13, 15).

The first phenomenon investigated was the deletion of initial [h] in words beginning with <h> (henceforth, H-deletion), which corresponds to the deletion of the glottal fricative [h] at the beginning of a word. As initial <h> has no corresponding sound in Portuguese, a Brazilian learner might produce [i] and [u] in the beginning of ‘hilarious’ and ‘humorist’, respectively. Therefore, other 123 similar words were selected to trigger the phenomenon. Another factor that might trigger this process is the existence of a silent <h> at the beginning of some English words, like ‘hour’ and ‘honor’.

The second phenomenon was the deletion of initial [h] with a change of [aj] to [i] in words beginning with <hy> (henceforth, HY-i). As in the previous process, the deletion of [h] occurs due to the absence of a sound corresponding to the grapheme <h> in initial position in Portuguese, especially in cognate words such as ‘hyper’, ‘hydrant’ and ‘hydrogen’. The reason for vowel shift is the fact that the grapheme <y> might be used to represent both [aj] and [i].

The third process chosen was only changing [aj] to [i] while keeping the pronunciation of initial [h] in words beginning with <hy> (henceforth, HY-hi). The HY-hi process goes in the opposite direction of the previous ones concerning the pronunciation of <h>. In H-deletion and HY-i processes, there is the deletion of initial [h], but in HY-hi the [h] is pronounced, with only a replacement of [aj] by [i], as described above. Since both processes, HY-i and HY-hi, are triggered by words beginning with <hy>, the same 150 words were used to test both phenomena.

The fourth process investigated is the pronunciation of silent <k> with the insertion an epenthetic [i] in words beginning with <kn> (henceforth, KN-kin). This transfer process is characterized by the pronunciation of [k] when <k> should be silent in words like ‘knife’ or ‘knickers’. Primarily, this phenomenon occurs because in BP the letter <k> in initial position is pronounced, and it is only silent in very few words of English origin, as is the cases of ‘know-how’ and ‘knock-out’. In turn, the insertion of the vowel [i] is a way for the learner to restructure the syllable considering the phonotactics of BP. The 108 words selected for the tests have this specific structure to serve as a trigger for the phenomenon.

The last process investigated was the voicing of /s/ when <s> occurs between two vowels (henceforth, S-z). It is the pronunciation of voiced [z] when the voiceless [s] should be pronounced. The voicing occurs in words like 'basic', 'case' or 'fantasy', and may come from the rule of pronouncing [z] when <s> is between two vowels in BP, a pattern easily transferred to the L2. Therefore, 125 words with <s> between vowels were selected to trigger this transfer phenomenon.

The corpus of this study was constructed and classified according to word frequency (high and low) and type of word (cognates, noncognates and nonwords). These categories can overlap, with the same word being classified concerning both its frequency and its type. The words were chosen from the Corpus of Contemporary American English (COCA)[3], an online and open-access corpus of English with more than a billion words from spoken and written language. The COCA corpus was also used to define the word frequency criterion, considering fewer than 1500 occurrences in the corpus as low frequency. Non-words were also incorporated to the study, all generated by the authors modifying existing words but still obeying English phonological patterns. As the pronunciations in this work were produced by software, only two recordings for each word were necessary, one with the effects of the transfer phenomenon, as if pronounced by a Brazilian learner, and the other without it, as if pronounced by an English native speaker. A varied quantity of words must be used to be able to reach statistical significance. For this reason, a total of 508 words were used, presented in Table 1, generating a total of 1016 recordings.

Table 1. Distribution of selected words for the phenomena in each category

Category	H-deletion	HY-i/HY-hi	KN-kin	S-z	Total in the category
High frequency	48	26	36	61	171
Low frequency	62	109	57	49	277
Cognate	69	94	0	69	232
Noncognate	41	41	108	41	231
Nonwords	15	15	8	15	53
Total in the process	**125**	**150**	**108**	**125**	**508**

2.1. Google Translate™ TTS System

To create a software capable of producing human-like speech, Google™ has developed a Text-to-Speech (TTS) system. The goal of text-to-speech is to generate a naturally sounding speech waveform given a text to be synthesized. It can be viewed as a sequence-to-sequence mapping problem; from a sequence of discrete symbols (the text) to a real-valued time series (waveform), which corresponds to the utterance. This process is designed to mimic human speech production, emulating the periodic (vocal cords vibration) and aperiodic (closure, burst frication) components present in human voice. The mainstream approach to speech synthesis in the recent woks of Google™ is the statistical parametric speech synthesis (SPSS) (16).

The SPSS paradigm is used together with a set of generative models to perform the mapping between the linguistic features extracted from the input text to acoustic features used in the speech production. SPSS based on hidden Markov models has grown in popularity over the last decade, becoming a popular option used today. This approach has various advantages over other techniques for speech synthesis; however, its major limitation is the quality of the synthesized speech (17). For this reason, recent researchers at Google™ have proposed the use of

[3] https://www.english-corpora.org/coca/

neural networks to perform the mapping between linguistic features and acoustic features (Tokuda and Zen, 2016; Tokuday and Zen, 2015; Zen et al., 2013; Zen et al., 2016).

In 2017, about 1/3 of all languages in Google's TTS options already used Recurrent Neural Networks (RNN) as acoustic models and almost all options of languages in Android mobile devices already used RNN-based TTS systems (22). Thus, it is possible to state that the Google Translate™ TSS structure mimics the human brain structure. The mapping of linguistic features to acoustic features using a parallel-distributed system is remarkably similar to the human reading process in the brain. Several works have demonstrated that it is possible to emulate parts of the human brain responsible for language processing using neural networks (23–25). Therefore, the tool can be seen as a connectionist simulation of the human brain processing language.

The fact that Google Translate™ TTS systems show deviations in the pronunciation when words that are not part of the training lexicon are presented is recognized and the company regularly publishes articles that develop techniques to avoid such situations (26, 27). Therefore, it is plausible to consider the tool capable of simulating the transfer processes that occur in humans learning a new language. In these cases, the system behaves as an adult learner of a foreign language in the early stages, adapting patterns already known by their neural network (the brain in the case of the learner), producing L2 forms that have undesired L1 characteristics.

To formally test the ability of a TTS system based on ANN to simulate transfer phenomena, we performed the test with the Google Translate™ audio option. This tool is free, simple, and available online in almost the entire world. To collect the samples, we selected Brazilian Portuguese as the input language and English as the output language, and the English words were written in the tool's inbox. Figure 1 illustrates this procedure with the word 'hygiene'.

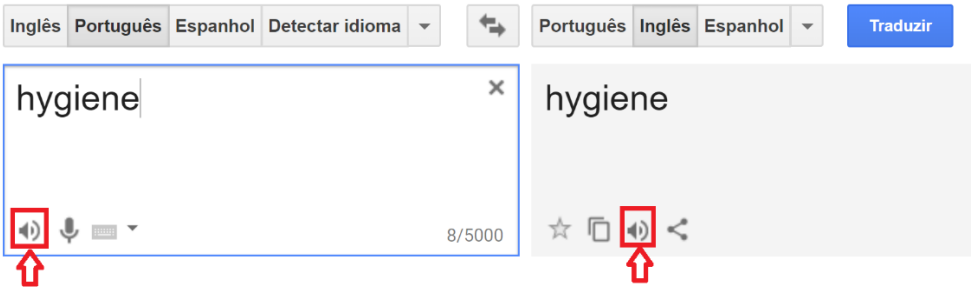

Figure 1. Illustration: Method for simulation of the transfer phenomena on the Google Translate™ platform

This way, the program generates the voice production of the English word using a system adapted for BP, thus producing some of the transfer phenomena observed in humans. After selecting the English language for the output box, which would correspond to the translation, the English word itself appears. The audio was also collected in this option to acquire the control native-like pronunciation of the word. The recordings were made using the Audacity™ software version 2.4.2 with the Microsoft Sound Mapper input mode, recording the digital productions directly from the operating system audio driver. All the data in this research were collected in August of 2018.

Although Google™ is transparent about the general principles of the algorithms used on the software, the Google Translate™ TTS system might be updated prior to the publication of this paper. This can be a limitation for reproducibility since some of the phenomena will no longer be produced by the BP voice due to improvements. Therefore, we made the recordings acquired

in 2018 and used in this study publicly available in a remote repository[4] as an open science effort. The recordings can be downloaded and verified by the peers. This is not a limitation to the study itself since its ultimate goal was not to test Google Translate™ TTS system, but rather to investigate the three classification algorithms in identifying the phenomena. The use of TTS-generated audio was simply a solution to work with reliable and easily acquired audio, but the next logical step of this research is to use actual learners' and native speakers' recordings as input.

2.2. Extraction of Acoustic Cues

To collect the samples produced by Google Translate™, we used the open-source audio software Audacity (version 2.1.2). The productions were recorded at 44.1 kHz (standard) in Wave 32-bit float PCM. However, raw speech cannot be directly used in the classification algorithms because it contains thousands of samples, which would make their processing slow, and polluted with noise, making it extremely difficult to extract knowledge from it. The solution is to represent the speech numerically with a set of coefficients obtained from the application of mathematical techniques, dividing the speech signal into multiple frames. To calculate this numeric representation, we opted to use the PRAAT software (version 6.0.21). To test different types of representation, we chose two descriptors: the mean of Formant Frequency (FF) and the mean of the Fundamental Frequency (f_0).

The sound produced in speech comes from the vibration of the vocal cords. This vibration is caused by the air flow from the lungs, creating pressure waves that propagate through the air, oscillating the air particles in a pseudo-periodic behavior. The number of "cycles" in a wave form, or the number of complete repetitions in the pseudo-periodic wave, is known as the Fundamental Frequency. This frequency is closely related to the number of times the vocal folds have opened and can be controlled by the speaker using the muscles around the vocal folds. Considering this mechanism, the fundamental frequency can be considered an indicator of vibration on the vocal cords (voicing).

Beyond its use in speech synthesis, fundamental frequency has been extensively used in speech recognition, speaker identification and speech understanding. The application in multiple-regression Hidden Markov Models as an auxiliary feature for word recognition can reduce error by 20% (28). The f_0 can be crucial for automatic speech processing in tonal languages such as Mandarin, where an effective speech recognizer needs to be able to recognize the 5 tones in addition to the usual phonetic inventory of the language (29). Widely used as a cue in speech recognition, f_0 was chosen in this work for the proposed identification task, helping to identify the phenomena that are closely related with sonorization.

Furthermore, when a vowel is produced, it is usually characterized by different resonant frequencies that vary according to their production. The sound produced by the vocal cords passes through the vocal tract, which functions as a filter. The pressure wave propagates through the vocal tract, where it resonates with greater or lesser intensity at different harmonic frequencies. The wave with maximum resonance is the one whose points of minimum and maximum vibration coincide with the length of the vocal tract. In the literature of speech production, the frequency of those waves of maximum resonance are denominated formants. In this study we used the first two formants, F1 and F2 (12, 13). It is plausible to predict that these formants, F1 and F2, carry information that characterizes the vowels produced by the Google Translate™ TTS system in a level of detail that it is possible to identify the transfers from BP to English-L2, since F1 and F2 are used by the human brain to determine vowel spectral quality and distinguish between vowels (F1 is related to vowel height and F2 to tongue advancement).

[4] https://github.com/atosborges00/ggTTS_paper

PRAAT presents the oscillogram and spectrogram of audio files. This way, it is possible to select, in each word, the exact region where each researched phenomenon occurred. This specific region was selected, cut, and saved in Wave format, resulting in a file referring to the exclusive region of incidence of transfer processes. The objective was to extract both f_0 mean and the mean of F1 and F2 from the selected region. Although two different methods are used to obtain these values, the same audio file was used for both extractions.

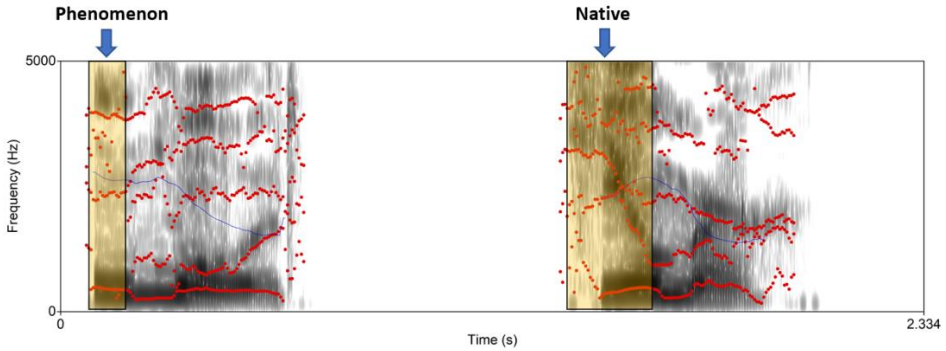

Figure 2. Illustration: Example of analysis of the word 'humor' produced by the TTS system. The arrows show the production of [u] on the left and the production of ['hju] on the right.

To extract the f_0 from the speech, PRAAT provides the option of outputting the frequency values with a collection of functions designed to implement speech analysis algorithms. In the case of fundamental frequency, the "To Pitch (ac)" command performs an acoustic periodic detection on the basis of an adapted autocorrelation method (30).

PRAAT automatically sets the f_0 value to "undefined" when the autocorrelation method cannot find a satisfactory value of correlation inside the typical values of fundamental frequency. As a mathematical strategy, we chose to switch this value to zero. With this change, the mean value of the words with unvoiced sections will differ from those without voiced sections. The classification algorithms can only rely on mathematical differences between the productions, and the addition of zeros to the mean during the unvoiced sections will make the differences in the speech production explicit to the classification algorithms by decreasing the mean value.

To obtain the mean of F1 and F2, PRAAT provides the "To Formant (burg)" command for conversion of audio objects to formant objects. This command first resamples the sound to a sampling frequency of twice the value of the parameter Maximum Formant and computes the LPC coefficients in the audio. The formant values are obtained through the poles that this algorithm computes.

It is important to warn that this methodology might result in non-typical values of formant frequencies. Normally the formant frequencies are extracted from the central region of the vowel. However, the selected region to study the phenomena was extended beyond the vowel. As the formant frequencies are obtained from the poles found in the LPC coefficients, they can be found in any region of speech, not exclusively in vowel production. Although the FF values in non-vowel regions are disperse and inconsistent, these values are useful to differentiate the mean value of the formant frequencies by the algorithms, moving the mean away from other observations without consonants. Even if this method results in implausible values for vowel production due to the influence of the FF found in consonantal regions, these differences will be evident to the classification algorithms, resulting in better separation of the groups.

3. Simulation Results

After the words produced by Google Translate™ were stored, we analyzed the productions and manually classified samples as phenomenon and no-phenomenon. As the words in this study were selected to trigger a clear manifestation of the phenomena, the identification task was trivial and performed by the authors. The recordings available in the remote repository[5] clearly indicate that when the pronunciation was BP-accented, it was heavily accented, with a clear production of the target phenomenon. The results indicate that the software indeed produces the transfer phenomena hypothesized, though not in all words. Words that trigger the processes in humans also triggered the transfer between the languages in a software using a neural network.

From the words selected for the H-deletion process, 80% triggered the transfer process in Google Translate™ TTS system. From the words selected for HY-i and HY-hi processes, 80% presented the HY-i process and only 10% presented the HY-hi process. The KN-kin process occurred in 41.67% of the words produced by the simulation. The S-z process occurred in 84% of the words selected for the study. Table 2 presents the frequency of occurrence of the phenomena in the categories of words selected for each process.

Table 2. Occurrence of the transfer processes in the samples of the BP

Process	High frequency (%)	Low frequency (%)	Cognates (%)	Noncognates (%)	Nonwords (%)
H-deletion	87.50	82.26	94.20	68.29	46.67
HY-i	57.69	85.32	81.91	75.61	80.00
HY-hi	7.69	10.09	8.51	12.20	13.34
KN-kin	8.33	52.63	-	35.48	80.00
S-z	72.13	95.92	88.41	73.17	93.34

From these results, it is possible to draw a series of conclusions regarding the occurrence of the phenomena in the TTS algorithm. For the H-deletion process, there is a clear tendency for the occurrence in cognate words when compared to noncognate words, an effect also observed, though slightly more mildly, in the HY-i phenomenon. This effect was not observed in the HY-hi or S-z phenomena, where neither presented significant differences.

Concerning word frequency, only the H-deletion process had more occurrence of the transfer phenomenon with the high frequency words; all other processes occurred more frequently in the low frequency words. The HY-hi process was the least frequent from the phenomena tested with the TTS system. Although they still occurred, the HY-i process was more dominant in the words capable of triggering both transfer processes. The shortage of samples for this process was a problem discussed later in the identification results section.

The HY-i, S-z and KN-kin phenomena presented a high level of occurrence with the nonwords. The unexpected result is the low occurrence of the H-deletion process in nonwords. The overall incidence of the HY-hi process was low, which also accounts for its low occurrence in nonwords. However, there is another unknown factor in the TTS algorithm influencing the production of the nonwords intended to trigger the H-deletion process.

In general, the results were compatible with the data already observed in humans (13). The higher incidence in cognates and low frequency words have been registered with beginning students; therefore, the neural networks behind the TTS system in Google Translate™ presented similar transfer patterns when exposed to similar inputs.

[5] https://github.com/atosborges00/ggTTS_paper

To better visualize the dataset and observe the differences between the pronunciations, we used F1 and F2 values from the regions of interest in the audio collected in BP and English to plot a visual representation of the productions. In the graphs in figure 3, we plotted the native-like pronunciation produced by the English option, as well as the productions from the BP option that produced the phenomenon and those without the phenomenon.

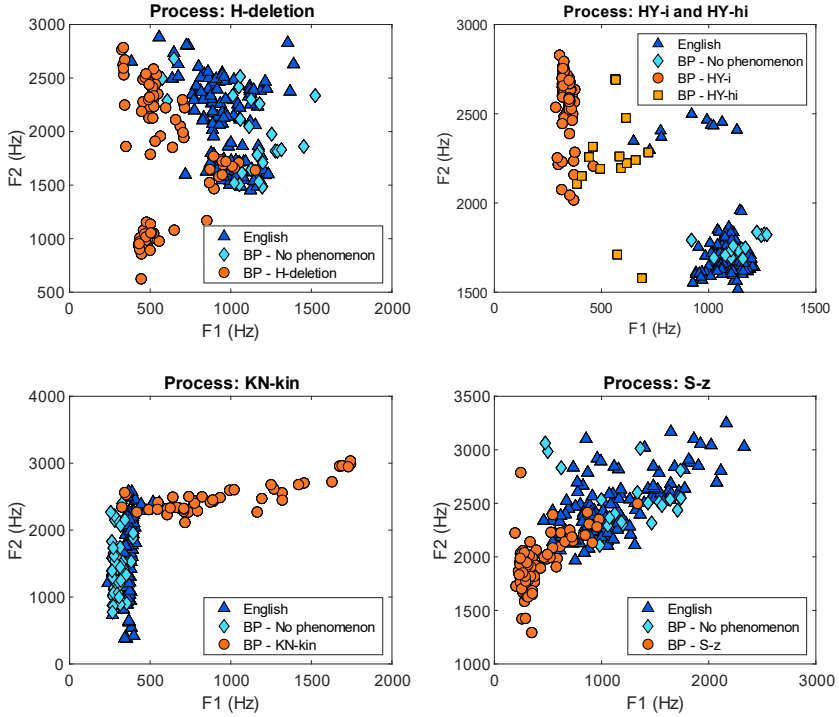

Figure 3. Dispersion plot: Distribution of Formant Frequencies in the productions of the simulated phenomena

It is possible to notice that the words are clustered within similar pronunciations. In the H-deletion process, it is possible to notice the formation of clusters around some formant values. These clusters are caused by the vowels that are pronounced when the initial [h] is deleted, each vowel presenting characteristic formant frequencies. The native-like pronunciation, on the other hand, presents the [h] sound, resulting in different FF means.

The agglomeration in the HY-i process is noteworthy, providing information of the vowel [i] being produced. The HY-hi process is in an intermediate region between pronunciations, with the production of the initial <h>, but changing [aj] with [i].

The KN-kin phenomenon presented low variation in the average of the second formant frequency, while the native pronunciation presents a variation of values approximately from 100 Hz to 3000 Hz. The opposite occurs with the mean of first formant, with a wide range of frequencies for the pronunciation with the transfer phenomenon. This behavior may be due to the appearance of antiformants in the production of the consonant <n>, which appear when the nasal cavity is involved in the sound production.

In the S-z plot we observe the agglomeration of words that present [s] sonorization, while samples without the phenomenon show greater dispersion. The pronunciation of the consonant [z] involves the vocal cords; therefore, there are resonant frequencies that can be interpreted as formant frequencies by the extraction algorithm. This is a good characterization of the phenomenon, with a distinction between the native-like pronunciation and that with the phenomenon, even with no vowels directly involved.

The distribution of mean f_0 values obtained in the simulations can be visualized in Figure 4, presenting the distribution for native-like pronunciations, as well as the mean values obtained in the BP audio option with and without the transfer processes.

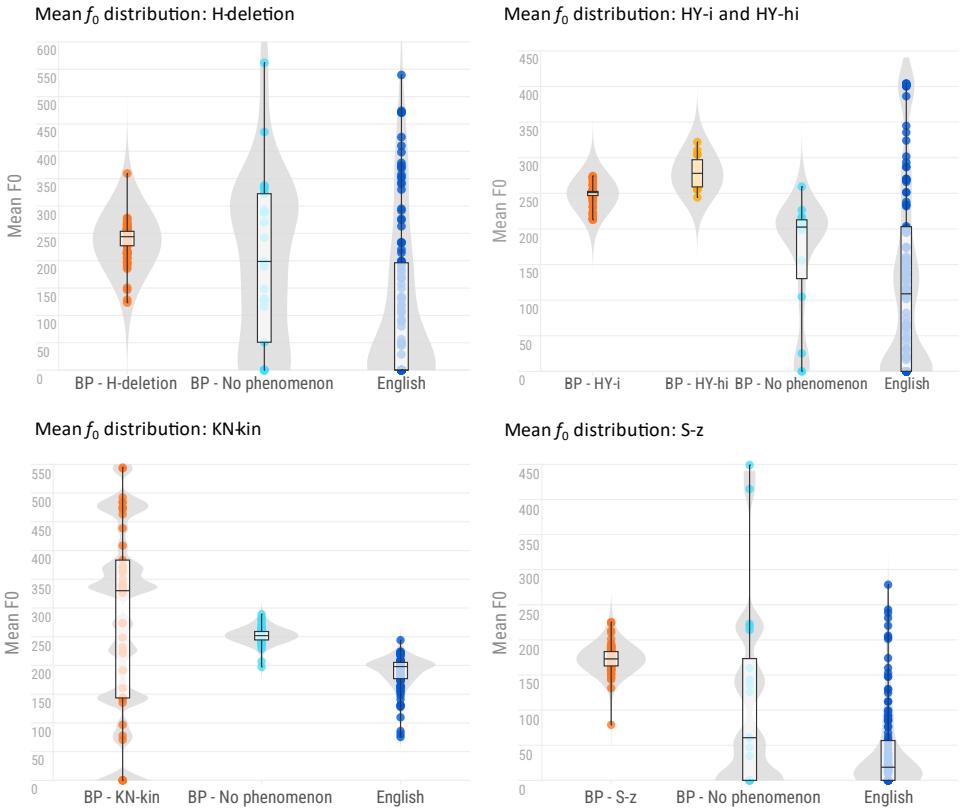

Figure 4. Boxplot with violin plot: Distribution of mean f_0 values in the utterances of the simulated phenomena

From the graphs presented, it is possible to draw a series of conclusions about the phenomena's behavior. In all processes, except for KN-kin, the phenomena were characterized by the concentration of mean f_0 values, indicating the presence of well-defined fundamental frequencies due to the voiced nature of the section, while samples from native-like pronunciations were more dispersed. The HY-i and HY-hi processes were also concentrated around different mean values, reinforcing the differences in the manifestations of the two processes. In KN-kin, the same reasons caused the mean f_0 concentration; however, with native-like samples presenting well-defined fundamental frequencies.

These contrasts in the distribution of mean f_0 highlight the differences between the productions, adding more evidence to the hypothesis of transfer process simulation and providing useful information to be used in the identification algorithms. These dynamics can be useful for the algorithms searching for mathematical disparities between the with-phenomena and the native-like pronunciations.

4. Identification Techniques

After the collection of the samples and extraction of f_0 mean, and F1 and F2 mean, three supervised algorithms were implemented to perform the automatic identification of the phenomena. The following diagram illustrates the process.

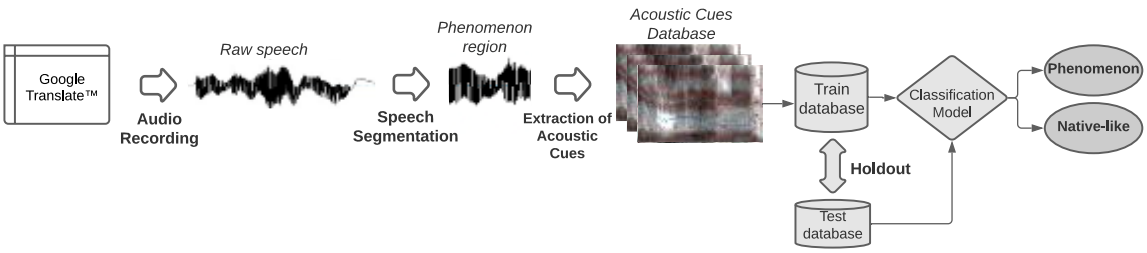

Figure 5. Flowchart: Construction and validation of the proposed identification system

As the three algorithms are supervised, the manually classified datasets were divided into a training subset (or memory subset) and a testing subset. The training subset is used as reference to the algorithm, presenting enough information about the behavior of the samples to allow for learning and generalization. With the training process completed, all three classification algorithms were tested with the testing subset. The samples of the training subset were never presented during the testing process or added in the reference data. This way we could test the accuracy and generalization levels of the models for new samples.

4.1. k-Nearest Neighbor

The idea behind the KNN method is simple. The most frequent class among neighbors closest to the sample to be classified is assigned to it (31). In other words, the classes of the nearest neighbors of the new sample are computed and the more common class is probably the class of the new instance.

Mathematically, it can be defined as: Let $V = \{v_1, v_2, \dots, v_3\}$ be a set of training patterns and $C_1, C_2, \dots, C_p$ the classes in which the set V was divided. The rule of the nearest neighbor can be defined as:

$$if\ d(v, v_a) \leq d(v, v_j), j = 1,2 \dots n\ and\ v_a \in C_i\ then\ v \in C_i$$

The distance $d(v, v_j)$ can assume different forms, e.g. Minkovski distance, Euclidean distance, Mahalanobis distance (32). The KNN algorithm computes the distance not only for the nearest neighbor, but also for the k nearest neighbors. The KNN algorithm performs the classification of the new samples according to the following pseudo-code (33):

Algorithm 1 k-Nearest Neighbors
Input: Dataset with unknow samples to be classified with dimensions (m, n).
Output: Vector with the m classifications of the samples from the dataset.
1: **for** i = 1 : *number test samples* **do**
2: Compute the Euclidean distance between the current vector v_i and other vector from the reference dataset defined as:

$$d(v_i, v_j) = \sqrt{\sum_{h=1}^{n} (x_{ih} - x_{jh})^2} \qquad (1)$$

3: Sort the vectors by the distance and store in $v^* = \{v_1^*, \dots, v_N^*\}$
4: Select the nearest k instances: $v^*[1:k]$
5: Determine the class of the sample as the most frequent class among the k nearest neighbors $v^*[1:k]$
6: **end for**

To avoid ties in the number of neighbors in each class, it is recommended that the *k* be odd. The optimum number for *k* must be obtained through tests.

Several works have already used this algorithm and its variations to perform all kinds of classification. The popularization of KNN happened in the 90's with some new applications of the algorithm (33). Since then, several works have used the algorithm for various types of identification or classification, including recognition of aspects human language (34–36).

4.2. Centroid Minimum Distance

The CMD algorithm works based on a basic principle about the dataset. The classification is based on the distances from the unknown samples to the center of mass of the already known classes. If the new sample is near to the center of mass of a class, also known as centroid, there is a high probability that the sample belongs to that class. The following pseudo-code demonstrates the steps of the algorithm.

Algorithm 2 Centroid Minimum Distance
Input: Dataset with unknow samples to be classified with dimensions (m, n).
Output: Vector with the m classifications of the samples from the dataset.
1: **for** i = 1 : *number of new samples* **do**
2: Find the center of mass m_j of each class ω_j with N_j elements defined as:

$$m_j = \frac{1}{N_j} \sum_{x \in \omega_j} x_k \qquad (2)$$

with k = 1, 2,…, N_j
3: Compute the Euclidean distance between the current vector v_i and the center of mass of all classes in the data set defined as:

$$d(v_i, m_j) = \sqrt{\sum_{h=1}^{n} (x_{ih} - m_{jh})^2} \qquad (3)$$

4: Sort the vectors by the distance and store in $v^* = \{v_1^*, \dots, v_N^*\}$
5: Determine the class of the sample as the class of the closest center of mass.
6: **end for**

The simplicity of this method is an important advantage for the implementation and universalization of the algorithm. It does not require huge processing power, making the implementation possible in most devices. It was already used in the identification of species of plants (37) and types of skin cancer (38).

4.3. Artificial Neural Networks

An artificial neural network is a system composed of ordered neurons in layers interconnected through synaptic weights. These synaptic weights ponder the connection between two neurons, or between an input and a neuron assuming a higher value according to the influence of that connection to the output of the network. ANN has input nodes that receive stimuli from the external medium and output neurons that provide the network response. Usually, a layer between the input and output neurons is used, known as the hidden layer. The use of the hidden layer structure enables ANN to solve non-linearly separable problems, approximating a function f: $I \rightarrow O$, $I \subseteq R^n, O \subseteq R^m$ where I is the training set and O is the target set. The neural network used in this research has a Multi-Layer Perceptron (MLP) architecture.

The term "learning" for an ANN is the act of establishing the output of the network by presenting a set of examples during the training stage. In this step, the adjustments of the synaptic weights occur to obtain the relations between input and output. In supervised learning (type of learning used in this work), the presented data patterns contain information about the stimuli applied in the input and the desired output in the last layer of the network. The precision of the model built by the network must be constantly measured. The Mean Square Error between the expected value d_k and the output of the neurons y_k is defined as:

$$\varepsilon_{ms} = \frac{1}{2N}\sum_{t=1}^{N}\sum_{k=1}^{M}[d_k(t) - y_k(t)]^2 \qquad (4)$$

where N is the number of training samples, M is the number of neurons in the output layer and t is the number of the current iteration.

The mean square error must be computed in each epoch and used to perfect the model through a training algorithm. The Levenberg-Marquardt algorithm was applied to perform this error minimization. This algorithm is defined as:

$$\boldsymbol{w}(p+1) = \boldsymbol{w}(p) - \boldsymbol{L}^{-1}\boldsymbol{J}^T(\boldsymbol{w})\boldsymbol{e}(\boldsymbol{w}) \qquad (5)$$

where $\boldsymbol{w}$ is the representation of the weights, $\boldsymbol{J}$ is the Jacobian matrix, $\boldsymbol{e}$ is the vector containing the errors and $\boldsymbol{L}$:

$$\boldsymbol{L} = \boldsymbol{J}^T(\boldsymbol{w})\boldsymbol{J}(\boldsymbol{w}) + \mu\boldsymbol{I} \qquad (6)$$

with μ being a scalar known as regularization constant and $\boldsymbol{I}$ the identity matrix. When μ is close to zero, the algorithm behaves similarly to the Gauss-Newton method for minimization. However, when μ assumes a high value, the behavior is close to the Back-Propagation algorithm. To summarize, the algorithm sequence is presented as:

Algorithm 3 Multi-Layer Perceptron ANN
Input: Dataset with unknow samples to be classified with dimensions (m, n).
Output: Vector with the m classifications of the samples from the dataset.
1: **for** t = 1 : *maximum number of epochs* **do**
2: Computes the feedforward propagation to obtain the output $d_k(t)$ for each k
3: Computes the mean square error ε_{ms} for all k samples
4: Adjust the $\boldsymbol{w}(p+1)$ weights by the Levenberg-Marquardt rule
5: **end for**
6: Computes the feedforward propagation to obtain the final classifications

This type of double-layered neural network with iterative training is called Multi-Layer Perceptron Artificial Neural Network. Its applications to speech processing are well stablished in the literature, with demonstrated accuracy and generalization capabilities (13).

5. Identification Results

To evaluate the identification performance, we validated the results with a score computed using precision and recall measures. This score is called the F1-score ("F" coming from F-score in statistics, not to be confused with first formant values) and uses precision and recall measures, both defined as:

- **Precision:** defined by the proportion of true positives in relation to the total number of samples predicted to be in that class, including false positives. Mathematically, it is defined as the number of true positives (T_P) divided by the sum of true positives and false positives (F_P).

$$P = \frac{T_P}{T_P + F_P} \quad (7)$$

- **Recall:** defined by the proportion of true positives in relation to all samples that in fact belong to that class, including false negatives. It is defined as the number of true positives (T_P) divided by the sum of true positives and false negatives.

$$R = \frac{T_P}{T_P + F_N} \quad (8)$$

The F1-score is then calculated as the harmonic mean between precision and recall.

$$F1 = \frac{2PR}{P + R} \quad (9)$$

In summary, the results correspond to the average F1-score for each of the 50 iterations using randomized holdout for training, cross-validation and testing subsets. The F1-score ± 1 standard deviation for the three algorithms is distributed in Table 3, presenting the performance in the test sets using both mean f_0 and the mean of the first two formant frequencies.

Table 3. F1-scores obtained by the algorithms in each phenomenon studied.

Algorithm	Processes			
	H-deletion	HY-i/HY-hi	KN-kin	S-z
kNN	**0.9459 ± 0.0234**	0.8039 ± 0.0958	0.9341 ± 0.0368	**0.9578 ± 0.0223**
CMD	0.9342 ± 0.0248	**0.8848 ± 0.0401**	0.8774 ± 0.0443	0.9356 ± 0.0293
ANN	0.9437 ± 0.0254	0.7729 ± 0.1221	**0.94044 ± 0.0334**	0.9450 ± 0.0250

The results presented by the three algorithms were in general satisfactory for the identification goal. The differences in performance for the techniques were expected and the best results are highlighted. For the H-deletion process the ANN and kNN showed similar results, both providing a high level of accuracy and precision for the identification and separation of native-like samples from samples with the phenomenon, followed closely by the CMD algorithm. The ANN also showed good performance for the KN-kin process, followed by the kNN algorithm, which presented results within the error margins.

The CMD algorithm presented the best performance for the HY-i/HY-hi processes, with a noticeable advantage. These two processes were a challenge for the algorithms due to the shortage of samples for the HY-hi phenomenon. The results presented in the Simulation Results section showed that the HY-hi process has formant values in the middle region between HY-i and native samples. Both the decision frontiers of the ANN and kNN algorithms were heavily influenced by the surrounding samples, while the CMD provided a fixed-point centroid independent from the samples around it, tracing a better indication of the region where the HY-hi samples were supposed to be.

In the S-z processes the three algorithms presented similar results, with kNN having the highest F1-score but showing no significant advantages for the other classifiers. The distribution of formant frequencies for this process did not provide any advantage in the identification strategy for any of the algorithms, all presenting high levels of accuracy and precision in identification.

6. Conclusions

After the evidence presented by the results, a series of conclusions about the three initial hypotheses could be drawn. The first hypothesis assumed was that the Google Translate™ text-to-speech system is able to simulate the grapho-phonic-phonological transfer phenomena. For this investigation, the collected data suggest that it is in fact possible to simulate the five proposed transfer phenomena. The frequency of occurrence differed for the phenomena and for different categories of words, but all the investigated processes were present at some level in the synthetic productions of the TTS algorithm[6].

For the second hypothesis, regarding the identification problem, the results indicated that ANNs, CMD and kNN can identify the transfer processes produced by the TTS algorithm using the audio descriptor with high levels of accuracy and precision, providing ways to automatically identify the five processes with confidence. The CMD algorithm revealed to be especially efficient in identifying the HPS processes. The challenge with the low number of samples was overcome by the CMD algorithm with a robust identification to surrounding samples of more dominant classes. We could not determine which algorithm had an overall best performance, as the differences in the results were mostly within the error margin.

The results are a proof-of-concept about the usage of algorithms with low computational complexity to identify the transfer phenomena in oral speech, an achievement made possible by the use of prior knowledge about the processes and what patterns emerge when a transfer phenomenon occurs. However, application in human speech production by L2 learners still needs to be tested to assure this method as a viable option for developers designing Computer Assisted Pronunciation Training software. The technique also needs to improve the acoustic cues extraction, automatically selecting the region of interest for a real-time classification.

It is also necessary to expand the investigation with more phenomena and to acquire a greater number of samples for each process investigated. Expanding the number samples and testing new phenomena with human-generated audio will provide new information for the development of a simple and efficient identification software. Further investigation can provide significant new information and ideas not only for software development but also about the phenomena themselves.

REFERENCES

1. Rocha ARS. Os efeitos da instrução explícita em fonologia na produção e percepção de consoantes da língua inglesa. [Dissertation - Masters]. Fortaleza, Brazil: Programa de Pós-Graduação em Linguística Aplicada, Universidade Estadual do Ceará; 2012. [accessed 25 Mar 2018] Available from: http://www.uece.br/posla/dmdocuments/AratuzaRodriguesSilvaRocha.pdf

2. Zimmer MC, Alves UK. A produção de aspectos fonético-fonológicos da segunda língua: instrução explícita e conexionismo. Rev Ling Ensino. 2006;9(2):101–43. [accessed 25 Mar 2018] Available from: http://www.rle.ucpel.tche.br/index.php/rle/article/view/168

3. Hayes-Harb R, Nicol J, Barker J. Learning the phonological forms of new words: effects of orthographic and auditory input. Lang Speech. 2010;53(Pt 3):367–81. doi: 10.1177/0023830910371460

4. Bassetti B, Escudero P, Hayes-Harb R. Second language phonology at the interface between acoustic and orthographic input. Appl Psycholinguist. 2015 Jan;36(1):1–6. [accessed 24 Oct 2021] Available from:

[6] As was already mentioned, it is possible that the current version of Google Translate™ TTS system generates fewer samples with the transfer phenomena than it did in August 2018 due to its constant improvement.

https://www.cambridge.org/core/journals/applied-psycholinguistics/article/second-language-phonology-at-the-interface-between-acoustic-and-orthographic-input/349B5CD70A06209C334EB78454305D25

5. Gonçalves AR, Silveira R. Orthographic effects in speech production: A psycholinguistic study with adult Brazilian-Portuguese English bilinguals / Efeitos ortográficos na produção da fala: um estudo psicolinguístico com adultos bilíngues falantes de Português Brasileiro e Inglês. Rev Estud Ling. 2020 May 27;28(3):1461–94. [accessed 24 Oct 2021] Available from: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/16454

6. Silveira R. PL2 production of english word-final consonants: the role of orthography and learner profile variables. Trab Em Linguística Apl. 2012 Jun;51:13–34. [accessed 24 Oct 2021] Available from: http://www.scielo.br/j/tla/a/xRmprsgPSBS8v6Wfw36tpTJ/?lang=en

7. Erdener VD, Burnham DK. The Role of Audiovisual Speech and Orthographic Information in Nonnative Speech Production. Lang Learn. 2005;55(2):191–228. [accessed 27 Oct 2021] Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0023-8333.2005.00303.x

8. Flege JE, Bohn O-S, Jang S. Effects of experience on non-native speakers' production and perception of English vowels. J Phon. 1997 Oct 1;25(4):437–70. [accessed 24 Mar 2018] Available from: http://www.sciencedirect.com/science/article/pii/S0095447097900528

9. Flege JE, Liu S. THE EFFECT OF EXPERIENCE ON ADULTS' ACQUISITION OF A SECOND LANGUAGE. Stud Second Lang Acquis. 2001 Dec;23(4):527–52. [accessed 24 Mar 2018] Available from: https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/the-effect-of-experience-on-adults-acquisition-of-a-second-language/4671D40F6250DB3E2F1FCF9B90190ED5

10. Bassetti B, Atkinson N. Effects of orthographic forms on pronunciation in experienced instructed second language learners. Appl Psycholinguist. 2015 Jan;36(1):67–91. [accessed 27 Oct 2021] Available from: https://www.cambridge.org/core/journals/applied-psycholinguistics/article/abs/effects-of-orthographic-forms-on-pronunciation-in-experienced-instructed-second-language-learners/335A18457216E019DF582B372319FA05

11. Silveira R, Gonçalves AR. Efeito da ortografia. In: Kupske F, Alves UK, Lima Jr. RM, editors. Investigando os sons de línguas não nativas: uma introdução. Editora da Abralin; 2021. [accessed 28 Oct 2021] Available from: https://www.doi.org/10.25189/9788568990117

12. Silva ACC, Macedo ACP, Barreto GA. A SOM-Based Analysis of Early Prosodic Acquisition of English by Brazilian Learners: Preliminary Results. In: Laaksonen J, Honkela T, editors. Advances in Self-Organizing Maps. Berlin, Heidelberg: Springer; 2011. p. 267–76. (Lecture Notes in Computer Science). doi: 10.1007/978-3-642-21566-7_27

13. Rocha ARS. Identificação de processos de transferência do português do brasil para o inglês (L2) por meio de rede neural artificial MLP [PhD]. Fortaleza, Brazil: Programa de Pós-Graduação em Linguística Aplicada, Universidade Estadual do Ceará; 2017. [accessed 25 Mar 2018] Available from: http://www.uece.br/posla/dmdocuments/Aratuza%20R.%20Silva.pdf

14. Zimmer MC, Bittencourt HR. Produção e percepção oral em L2: os processos de transferência do conhecimento grafo fônico-fonológico do português brasileiro (L1) para o inglês (L2) e o desempenho em listening (L2). Cad Estud Lingüíst. 2008;50(1). [accessed 25 Mar 2018] Available from: https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8637237

15. Zimmer MC. A transferência do conhecimento fonético-fonológico do português brasileiro (L1) para o inglês (L2) na recodificação leitora: uma abordagem conexionista [PhD]. Porto Alegre, Brazil: Faculdade de Letras, Pontifícia Universidade Católica do Rio Grande do Sul; 2003. [accessed 25 Mar 2018] Available from: http://www.leffa.pro.br/tela4/Textos/Textos/Teses/marcia_zimmer.pdf

16. Zen H, Tokuda K, Black AW. Statistical parametric speech synthesis. Speech Commun. 2009 Nov 1;51(11):1039–64. [accessed 25 Mar 2018] Available from: http://www.sciencedirect.com/science/article/pii/S0167639309000648

17. Zen H. Acoustic Modeling in Statistical Parametric Speech Synthesis - From HMM to LSTM-RNN. In Fukushima, Japan; 2015. [accessed 25 Mar 2018] Available from: https://research.google.com/pubs/pub43893.html

18. Tokuday K, Zen H. Directly modeling voiced and unvoiced components in speech waveforms by neural networks. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. p. 5640–4. doi: 10.1109/ICASSP.2016.7472757

19. Tokuday K, Zen H. Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. p. 4215–9. doi: 10.1109/ICASSP.2015.7178765

20. Ze H, Senior A, Schuster M. Statistical parametric speech synthesis using deep neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013. p. 7962–6. doi: 10.1109/ICASSP.2013.6639215

21. Zen H, Agiomyrgiannakis Y, Egberts N, Henderson F, Szczepaniak P. Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices. In: arXiv:160606061 [cs]. San Francisco, CA, USA; 2016. [accessed 25 Mar 2018] Available from: http://arxiv.org/abs/1606.06061

22. Zen H. Generative Model-Based Text-to-Speech Synthesis. 2017. [accessed 25 Mar 2018] Available from: https://research.google.com/pubs/pub45882.html

23. Elman JL. Learning and development in neural networks: the importance of starting small. Cognition. 1993 Jul 1;48(1):71–99. [accessed 24 Mar 2018] Available from: http://www.sciencedirect.com/science/article/pii/0010027793900584

24. Li P, Farkas I, MacWhinney B. Early lexical development in a self-organizing neural network. Neural Netw. 2004 Oct 1;17(8):1345–62. [accessed 26 Mar 2018] Available from: http://www.sciencedirect.com/science/article/pii/S0893608004001534

25. MacDonald M, Christiansen M. Reassessing working memory: A reply to Just & Carpenter and Waters & Caplan. Psychol Rev. 2002 Feb 1;109:35–54; discussion 55. doi: 10.1037//0033-295X.109.1.35

26. Gonzalvo X, Podsiadlo M. Text-To-Speech with cross-lingual Neural Network-based grapheme-to-phoneme models. In 2014. [accessed 25 Mar 2018] Available from: https://research.google.com/pubs/pub45183.html

27. Li B, Zen H. Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis. In 2016. p. 2468–72. doi: 10.21437/Interspeech.2016-172

28. Fujinaga K, Nakai M, Shimodaira H, Sagayama S. Multiple-regression hidden Markov model. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (Cat No01CH37221). 2001. p. 513–6 vol.1. doi: 10.1109/ICASSP.2001.940880

29. Chen K, Yang C. The Effect of Fundamental Frequency on Mandarin Intelligibility by L2 Learners in Quiet and Noise Environments: A Pilot Study. In: Yang C, editor. The Acquisition of Chinese as a Second Language Pronunciation: Segments and Prosody. Singapore: Springer; 2021. p. 213–32. (Prosody, Phonology and Phonetics). [accessed 26 Jun 2021] Available from: https://doi.org/10.1007/978-981-15-3809-4_10

30. Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: IFA Proceedings 17. 1993. p. 97–110.

31. Fix E, Hodges JL. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. Int Stat Rev Rev Int Stat. 1989;57(3):238–47. [accessed 29 Mar 2018] Available from: http://www.jstor.org/stable/1403797

32. Morariu N. Using Pattern Classification and Recognition Techniques for Diagnostic and Prediction. Adv Electr Comput Eng. 2007 Apr 2;7(1):63–7. [accessed 25 Mar 2018] Available from: http://dx.doi.org/10.4316/AECE.2007.01014

33. Aha DW. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. Int J Man-Mach Stud. 1992 Feb 1;36(2):267–87. [accessed 29 Mar 2018] Available from: http://www.sciencedirect.com/science/article/pii/002073739290018G

34. Petrushin V. Emotion recognition in speech signal: Experimental study, development, and application. In: Proc ICSLP. 2000. p. 222–5.

35. Velican V. Automatic Recognition of Improperly Pronounced Initial 'r' Consonant in Romanian. Adv Electr Comput Eng. 2012 Aug 31;12(3):79–84. [accessed 29 Mar 2018] Available from: http://dx.doi.org/10.4316/AECE.2012.03012

36. Yan Z, Xu C. Combining KNN algorithm and other classifiers. In: 2010 9th IEEE International Conference on Cognitive Informatics (ICCI). 2010. p. 800–5. doi: 10.1109/COGINF.2010.5599804

37. Scaranti A, Bernardi R. Identificação de Órgãos Foliares utilizando as Wavelets de Daubechies. In Presidente Prudente, Brazil; 2010. [accessed 29 Mar 2018] Available from: http://iris.sel.eesc.usp.br/wvc/anais_WVC2010/artigos/poster/72803.pdf

38. Frutuoso RL, Santos JRVD, Siqueira R da S, Oliveira AC de. Uso de algoritmos de reconhecimento de padrões aplicados ao problema de câncer de pele do tipo melanoma. In SBIC; 2016. p. 1–6. [accessed 29 Mar 2018] Available from: http://abricom.org.br/eventos/cbic_2013/bricsccicbic2013_submission_321