INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

The Influence of Parallel Processing on Illusory Vowels

Takeshi Kishiyama

Language and Information Sciences, The University of Tokyo, Japan
kishiyama.t@gmail.com

Abstract

Research has shown that listeners perceive illusory vowels inside consonant clusters that are not allowed in their L1. This phenomenon has been examined using several psycholinguistic and computational models, including hidden Markov models (HMMs), applied to human phoneme perception. However, the inference algorithm of HMMs assumes that parallel processing, which has not been proven to have psychological reality, is a valid cognitive process. This study tested the psychological reality of parallel processing by attempting to duplicate two results from previous studies: First, listeners perceive an illusory vowel in consonant clusters that are not permissible in their L1. Second, the illusory vowel is based on the characteristics of the preceding consonant, indicating that listeners integrate phonotactics and acoustic information. The experiment manipulated the number of candidates that the model can refer to, and the algorithm can be considered parallel when it allows models to use more than two candidates that are stored in memory. In addition, the transition probabilities between consonants were manipulated to represent the different phonotactics. The results showed that only the parallel processing condition reproduced the two observations above, supporting the psychological reality of parallel processing.

Index Terms: perceptual epenthesis, non-native speech perception, phonotactics, parallel processing

1. Introduction

This study examines the role of parallel processing in human phoneme perception, reproducing vowel epenthesis with computational models. While psycholinguists have examined human language comprehension through empirical experiments, there have also been studies using computational models. Hidden Markov models (HMMs) are one example of these computational models but there are several issues to consider when using them as models of cognitive processes. I will first introduce the phenomenon of perceptual vowel epenthesis and the models used to study and describe that phenomenon. I will then describe the issues with these models, followed by an overview of the methodologies of this computational study and its antecedents.

1.1. Illusory vowels and psycholinguistic models

Perceptual vowel epenthesis is a phenomenon in which speakers of languages that do not allow certain consonant clusters perceive illusory vowels between those consonant clusters [1, 2]. This phenomenon exists in a variety of languages such as Japanese, English, and Spanish [3, 4]. I will introduce two empirical studies that tested the role of phonotactics and acoustic cues in vowel epenthesis, as well as the models used to explain these experimental results.

The effects of phonotactics on vowel epenthesis was reported in the 1999 study "Epenthetic vowels in Japanese: a Perceptual Illusion?" by Dupoux et al. Native speakers of French and Japanese participated in the experiments. In that study, stimuli such as /ebuzo/ and /ebzo/ were presented to speakers of each language. French allows the /bz/ sequence, whereas Japanese does not. The study's results showed that native speakers of French did not perceive any epenthetic vowel in /ebzo/, while native speakers of Japanese perceived /u/ between /b/ and /z/, indicating that vowel illusions are caused by phonotactics at the perceptual level [1].

The effect of acoustic cues on the selection of inserted vowels was investigated using experiments with native speakers of Brazilian Portuguese and of Japanese in the 2011 study "Where do illusory vowels come from?" by Dupoux et al. Brazilian Portuguese and Japanese insert /i/ and /u/, respectively, as their default vowels, and they do not allow the /bz/ sequence. In the experiment, the vowels between /b/–/z/ in /ebizo/ and /ebuzo/ were reduced to create consonants with the acoustic cues of the vowels. Speakers of both languages were affected by the acoustic cues, and the insertion rate of the non-default /i/ increased significantly in native Japanese speakers [2].

There are two psycholinguistic models that explain the phenomenon described above: the two-step model [3] and the one-step model [2]. The two-step model converts sound to phonemes, dropping the acoustic cues before insertion, whereas the one-step model can reference the acoustic cues of consonants during insertion. The influence of acoustic cues on vowel selection can be explained only by the one-step model, making it the dominant model at present.

HMMs are the computational models that correspond to the one-step model [2]. A recent study has shown that transition probability modulates perceptual epenthesis [5] and that HMMs can express phonotactics that do not allow /z/ after /b/ as $P(\text{/z/}|\text{/b/}) = 0$, which means the transition probability from /b/ to /z/ is 0. Although HMMs have been used in recent computational studies [6, 7, 8], there are several issues in using them as models for phoneme-sequence perception. In the following section, I will compare the phoneme-perception results across models and summarize the issues presented by each model.

1.2. The psychological reality of computational models

HMMs are a generalization of the traditional phoneme-perception model, but they are problematic in that they fail to reproduce the traditional phenomena. Unless the psychological reality—the validity of a model as a cognitive model— can be guaranteed, comparing models will not contribute to the elucidation of cognitive processes. In this section, I will introduce phenomena [9] and models [10] that can be generalized by HMMs, describe the relationship between each model and HMMs [11], and discuss the issues presented by using HMMs.

One example of the phenomenon of single-phoneme perception is the *perceptual magnet effect* [9], in which it becomes difficult for participants to discriminate sounds around the *prototype*, which has the highest $P(S_{\text{new}}|c)$, the likelihood of the

http://dx.doi.org/10.21437/Interspeech.2021-89

new sound S_{new} as a phoneme c. The exemplar-based perceptual model [10] reproduced this phenomenon, supporting its psychological reality. The process of inferring the phoneme $\hat{c}$ for the new sound S_{new} in the exemplar-based perceptual model is equivalent to that of a Naive Bayes model [11]. This perceptual model can be considered a special case of HMMs.

A Naive Bayes model finds the candidate $\hat{c}$ that maximizes the product of $P(c)$ and $P(S_{\text{new}}|c)$ as

$$\hat{c} = \arg\max_{c} P(c)P(S_{\text{new}}|c) \tag{1}$$

when it maps the new sound as a phoneme. By contrast, an HMM calculation is

$$\hat{c} = \arg\max_{c} P(c|c_{t-1})P(S_{\text{new}}|c) \tag{2}$$

and the $P(c)$ is conditioned on the previous phoneme c_{t-1}, yielding $P(c|c_{t-1})$. Therefore, HMMs are a generalization of Naive Bayes models: The special case that does not allow this conditioning is equal to the equation (1).

HMMs are important to consider not only because they can partially integrate phonotactics into a model, but because they are a generalization of the computational model of phoneme perception. This means that phoneme perception can be treated as a special case of phoneme-sequence perception, which allows the model to be consistent across the input. There are, however, three problems with validation when using HMMs.

First, HMMs do not support psychological reality. Even though the phenomena of phoneme-sequence perception include the effect of phonotactics on perception [1] and the effect of acoustic information on vowel selection [2], these effects have not been replicated using HMMs. In addition to the lack of replication, phonotactics have prevented HMMs from replicating previously-established empirical evidence [6].

The necessity of phonotactics in phoneme perception was examined in a 2018 experiment that used consonant clusters with acoustic cues of the vowel /ah(a)pa/ [6]. First, the research confirmed that native Japanese speakers select and insert vowels based on acoustic details [12], as shown by the 2011 Dupoux et al. study. Then, they created several models: HMMs, with phonotactic representation; Naive Bayes models, which don't represent phonotactics; and models that use acoustic information only. The Naive Bayes models and HMMs yielded a low correlation to the results of previous behavioral experiments, indicating that phonotactics are not necessary to explain the phenomenon [6].

The third and final problem with using HMMs for validation is the unsupported assumption of parallel processing. The Viterbi algorithm [13] used for inference in HMMs stores and uses multiple c_{t-1} phonemes as candidates in the computational process. Such parallel processing cannot be assumed without its validation via models of cognitive processes. It is still being tested in other linguistics domains such as lexical comprehension [14] and sentence processing [15]. Therefore, it is necessary to verify the psychological reality of parallel processing by reproducing empirical evidence via computational models of phoneme-sequence perception.

1.3. Present Study

I examined the validity of phonotactics and parallel processing, which is controversial when we treat HMMs as models of cognitive processes. For this purpose, I set up an algorithm with three conditions: a Naive Bayes model, a serial HMM, and a parallel HMM. Each condition calculates the candidate that maximizes the likelihood of each input, but a Naive Bayes model cannot reference the previous phoneme as Figure 1a. By contrast, a serial HMM can look up one preceding candidate as Figure 1b, and a parallel HMM can refer to two or more candidates simultaneously as Figure 1c. This makes it possible to control the use of phonotactics and parallel processing.

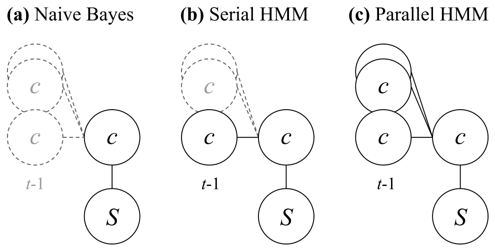

Figure 1: *A Naive Bayes model (a) cannot refer to the previous phoneme, whereas a serial HMM (b) can, and a parallel HMM (c) can store and refer to multiple phonemes.*

The conditions were manipulated to see which algorithm could reproduce the empirical evidence of previous studies. In Experiment 1, I created computational models for the 1999 Japanese- and French-language experiments of Dupoux et al. and examined if the effects of phonotactics on perception could be reproduced. In Experiment 2, I created computational models for the 2011 Dupoux et al. Japanese- and Brazilian Portuguese-language experiments to reproduce the effects of acoustic cues on vowel selection. If both phonotactics and parallel processing are required, only the parallel HMM condition reproduces the previously-observed empirical results.

In my experiments, only a parallel HMM reproduced the empirical evidence, indicating that phonotactics alone are not sufficient and that parallel processing is also necessary.

2. Experiment 1

2.1. Model and data

As a condition of language, the model represented the phonotactics of Japanese with the transition probability $P(c|c_{t-1})$. Also, if c and c_{t-1} are the same phonemes, it was set to 0.6 or 0.7 to model the durations. Other transition probabilities were equally distributed. For example, the transition probability from /b/ to /o/, /e/, /u/, /b/, and /z/ was set to 0.1, 0.1, 0.1, 0.7, and 0, respectively, in the models of Japanese. Here, $P(/z/|/b/)$ was set to zero to represent phonotactics that prohibit the consonant clusters, whereas $P(/b/|/b/)$ represented their duration.

For the condition of vowel length in the model's input stimulus, two data from each phoneme of /ebuzo/ were generated (/eebbuuzzoo/). Each phoneme, /o/, /e/, /u/, /b/, and /z/, was normally distributed, with respective means of 500, 600, 700, 800, and 900. The standard deviation was set to 25. Then, the /u/-duration was adjusted to 2, 1, or 0 to have the condition of vowel length.

Having set parameters and data in this way, the transition probability $P(c|c_{t-1})$ can be calculated. Furthermore, the data generated from each normal distribution is one-dimensional, allowing us to calculate the likelihood of sound $P(S_{new}|c)$. It's enough for simulation because both equations (1) and (2) can be calculated. The data and code for replication are available at: `https://github.com/kishiyamat/interspeech-2021-replication`.

2.2. Procedure

The models described above were given data with different vowel lengths for /u/, and each algorithm inferred the phonemes. The inference was performed 100 times for all combinations of factors (models, data, and algorithms). Since the results of inference look like /eebbzzoo/, duplicates were grouped as /ebzo/. Trials that resulted in /ebuzo/ were marked as 1 (epenthesis), and all other cases were marked as 0 (no epenthesis). This is because, in Experiment 1, the only possible vowel in the distribution of /b/ is /u/, and therefore, if there is an insertion, /ebuzo/ is always produced. Since there are 100 results from each trial, the number of insertions becomes the vowel insertion rate (%).

2.3. Results

In the Dupoux study, the vowel insertion rate of native Japanese speakers was high even for stimuli without vowels, as shown in Figure 2a. In my study, only the parallel HMM reproduced this trend, with a vowel insertion rate of 50–60% for stimuli without vowels (viterbi 2 and 3 in Figure 2b). Furthermore, there was a significant language effect ($p < .001$), a vowel duration effect ($p < .001$), and an interaction effect between language and the linear component of vowel duration ($p < .005$), all of which were as reported in the Dupoux study.

By contrast, the Naive Bayes model and the serial HMM used here (naive_bayes 1 and viterbi 1 in Figure 2b) did not insert vowels for stimuli without vowels. There was only a vowel duration effect ($p < .001$), which is not a replication of the results of the Dupoux study.

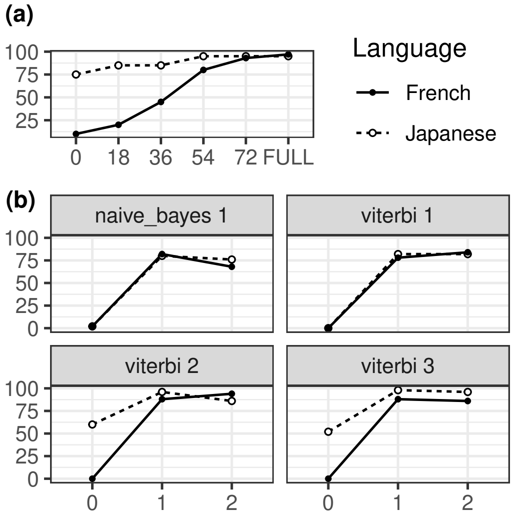

Figure 2: *The vowel insertion rate as a function of vowel length in Figure 1 from Dupoux et al., 1999, "Epenthetic vowels in Japanese: a Perceptual Illusion?" Journal of Experimental Psychology: Human Perception and Performance, Vol. 25, p. 1571. was created by the author in (a). The rate was also drawn for each algorithm based on the inference results in this study as in (b). The difference between viterbi 2 and viterbi 3 is the number of candidates the model can refer to.*

Next, I counted the number of inference results from each algorithm when the vowel duration was zero (Figure 3). The model of French does not change depending on the algorithm used, whereas the model of Japanese shows different trends depending on the algorithm used. In the Naive Bayes model, the phonotactics were ignored, yielding /ebzo/ without epenthesis. By contrast, the serial HMM inferred /ebo/, which not only did not insert a vowel, it also removed /z/.

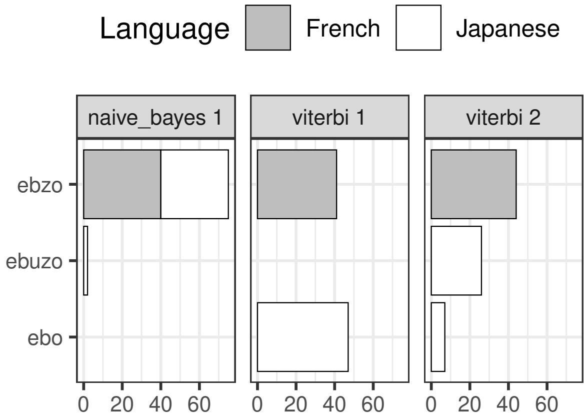

Figure 3: *In condition where the duration of /u/ was set to zero, the phoneme patterns that appeared more than 10 times were counted for each condition of algorithm and language.*

2.4. Discussion

In Experiment 1, only the parallel HMM reproduced the empirical evidence from the 1999 Dupoux et al. study. In this section, the following three points are discussed: (i) the cause of failure in the serial HMM, (ii) the process of the parallel HMM, and (iii) the cause of failure in the Naive Bayes model.

The serial HMM could not assign /z/ after /eb/ because $P(\text{/z/}|\text{/b/})$ is zero in the Japanese-language model. Furthermore, in Experiment 1, the data in /z/ is not likely to be labeled as a vowel, so no transition could take place after /b/. This setting of /z/ results in the inference that /b/ continues, yielding /ebo/. Even if a vowel were to have been inserted, this inference would have contradicted the empirical evidence that the epenthesis is based on the acoustic cues of the previous consonant.

In contrast, the parallel HMM keeps both /b/ and /u/ as previous candidates when it assigns /z/. Therefore, it can compare the likelihood of $P(\text{/z/}|\text{/b/})$ and $P(\text{/z/}|\text{/u/})$ after having the sound of /z/. This allows the model to select /u/ as the previous candidate, which yields /ebuzo/ and replicates the results from Dupoux et al.

Finally, the Naive Bayes model ignored the phonotactics and selected the most plausible candidate at each point, yielding /ebzo/ even when the modeled language was Japanese. Thus, the Naive Bayes model made the same inferences for both Japanese and French, which is contrary to the empirical evidence of the Dupoux et al. study.

The next step is to examine the effect of acoustic cues on vowel selection, where the insertion of a vowel is supposed to depend on the acoustic cues of the previous consonant.

3. Experiment 2

3.1. Model and data

In Experiment 2, the modeled languages are Brazilian Portuguese and Japanese, which insert /i/ and /u/, respectively, as their default vowels. This difference between the two languages is represented by increasing the transition probability to their default vowels. Also, since each language devoices /i/ and /u/ [16, 17], these vowels have more variance than other vowels. This difference is represented by the wider standard deviations of /i/ and /u/ in the models for each language.

Two data were generated from each phoneme's distribution. In Experiment 2, each phoneme, /o/, /e/, /u/, /b/, /z/, and /i/, had a normal distribution, with 500, 600, 700, 800, 1000, and 900 as their respective means, and 25 as the standard deviation. The standard deviation of the default vowel for each language was changed to 50. Furthermore, the coarticulation of either /i/ or /u/ was given to the data of /b/. In the /i/-coarticulation condition, for example, /b/ was the most likely candidate, while /i/ was the next most likely candidate. The natural condition without coarticulation was also added for comparison.

3.2. Procedure

The models described above used data with different types of coarticulation, and each algorithm inferred the phonemes. After assigning 1 for /ebizo/ and -1 for /ebuzo/, these values were averaged, so that the value indicates a preference for /i/ as it approaches 1 and a preference for /u/ as it approaches -1.

3.3. Results

In the 2011 study, the preference was affected by coarticulation, yielding /i/ for the /i/-coarticulation condition and /u/ for the /u/-coarticulation condition in both languages (Figure 4a). In the present study, only a parallel HMM could reproduce this trend (Figure 4b). By contrast, the Naive Bayes model and the serial HMM failed to do so.

Furthermore, there was a significant language effect ($p <$ .001), a coarticulation effect ($p < .001$), and an interaction between those two factors ($p < .05$), all of which were as reported in the Dupoux et al. study.

The ratio of inserted vowels to articulatory combinations seems similar in the Naive Bayes model and the serial HMM, as shown in in Figure 4b. As I aggregated the frequency of inference results to the /i/-coarticulation, the Naive Bayes model rarely inserted /i/, while /i/ insertions are common in the serial HMM, which is contrary to the Dupoux et al. evidence.

3.4. Discussion

In Experiment 2, only the parallel HMM condition reproduced the empirical evidence of the 2011 study. Unlike Experiment 1, the Naive Bayes model also inserted vowels in some cases, but this is thought to be statistical noise resulting from the increase in the standard deviation of the vowels. The serial HMM model also contradicted the empirical evidence, in that the Japanese-language modeling inserted /i/ even with the /u/-coarticulation condition.

4. General Discussion

The present study supports the psychological reality of HMMs and of parallel processing by reproducing two phenomena. Two points are discussed here: (i) the differences from previous studies in which phonotactics failed to reproduce the phenomena, and (ii) the relationship between the HMM and the one-step model.

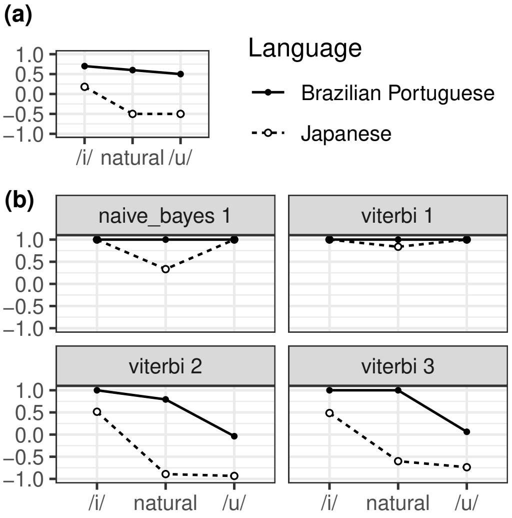

Figure 4: *Mean /i/ minus /u/ difference scores as a function of coarticulation type in Figure 3 from Dupoux et al., 2011, "Where do illusory vowels come from?" Journal of Memory and Language, Vol. 64, No. 3, p. 204 was created by the author in (a) Mean /i/ minus /u/ difference scores as a function of coarticulation type was also drawn for each algorithm in this study as in (b).*

In the 2018 study [6], the phonotactics made HMMs fail to reproduce the phenomena, whereas the acoustic model succeeded. As long as HMMs are necessary for the current study to reproduce previous experimental results, the acoustic model, which is a special-case HMM, is not sufficient. The parameters of the HMM in the previous study may have been inappropriate for the task, which could explain the differences in the previous and the present studies' results. In the 2018 study, the only difference between the models was in the language model's parameters. The HMM language model needs to be adjusted to include not only transition probabilities but durations. If it was set inappropriately for the task, the failure of the HMM can be explained as overfitting.

Although an HMM has been proposed for the one-step model, HMM is inadequate as a model for linguistic knowledge and it should be generalized. The HMM assumes a geometric distribution for duration and stationarity for the acoustic model. The HMM is a restricted model of the corresponding one-step model, which does not include these assumptions. However, since these assumptions are resolved by allowing the hidden semi-Markov model to assume that duration has a Poisson distribution and by allowing the segment model [18] to assume non-stationarity, further extension and verification are necessary.

5. Acknowledgements

The author is grateful to Yuki Hirose and Yohei Oseki for helpful discussions. For the English language review, the author would like to thank Enago (www.enago.jp).

6. References

[1] E. Dupoux, K. Kakehi, Y. Hirose, C. Pallier, and J. Mehler, "Epenthetic vowels in Japanese: A perceptual illusion?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, pp. 1568–1578, 1999.

[2] E. Dupoux, E. Parlato, S. Frota, Y. Hirose, and S. Peperkamp, "Where do illusory vowels come from?" *Journal of Memory and Language*, vol. 64, no. 3, pp. 199–210, 2011.

[3] I. Berent, D. Steriade, T. Lennertz, and V. Vaknin, "What we know about what we have never heard: Evidence from perceptual illusions," *Cognition*, vol. 104, no. 3, pp. 591–630, 2007.

[4] B. Kabak and W. J. Idsardi, "Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints?" *Language and speech*, vol. 50, no. 1, pp. 23–52, 2007.

[5] A. Kilpatrick, S. Kawahara, R. Bundgaard-Nielsen, B. Baker, and J. Fletcher, "Japanese perceptual epenthesis is modulated by transitional probability," *Language and Speech*, p. 0023830920930042, 2020.

[6] A. Guevara-Rukoz, "Decoding perceptual vowel epenthesis: Experiments & modelling," Ph.D. dissertation, Ecole Normale Supérieure (ENS), 2018.

[7] T. Schatz and N. H. Feldman, "Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception," in *Proceedings of the Conference on Cognitive Computational Neuroscience*, 2018.

[8] J. Gong, M. Cooke, and M. L. G. Lecumberri, "A quantitative model of first language influence in second language consonant learning," *Speech Communication*, vol. 69, pp. 17–30, 2015.

[9] P. K. Kuhl, "Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not," *Perception & psychophysics*, vol. 50, no. 2, pp. 93–107, 1991.

[10] F. Lacerda, "The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory," in *Proceedings of the XIIIth international congress of phonetic sciences*, vol. 2. Stockholm University Stockholm, 1995, pp. 140–147.

[11] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference." *Psychological review*, vol. 116, no. 4, p. 752, 2009.

[12] A. Guevara-Rukoz, I. Lin, M. Morii, Y. Minagawa, E. Dupoux, and S. Peperkamp, "Which epenthetic vowel? phonetic categories versus acoustic detail in perceptual vowel epenthesis," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. EL211–EL217, 2017.

[13] G. D. Forney, "The Viterbi Algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[14] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[15] M. C. MacDonald, N. J. Pearlmutter, and M. S. Seidenberg, "The lexical nature of syntactic ambiguity resolution." *Psychological review*, vol. 101, no. 4, p. 676, 1994.

[16] M. S. Han, "Unvoicing of vowels in Japanese," *Study of Sounds*, vol. 10, pp. 81–100, 1962.

[17] J. M. Camara Jr and J. Mattoso, *The Portuguese Language*. Chicago: University of Chicago Press, 1979.

[18] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on speech and audio processing*, vol. 4, no. 5, pp. 360–378, 1996.