Formal modelling of L1 and L2 perceptual learning: Computational linguistics versus machine learning

Paola Escudero [1*], *Jelle Kastelein* [2], *Klara Weiand* [2], *R.J.J.H. van Son* [1]

[1] Institute of Phonetic Sciences, University of Amsterdam, The Netherlands
[2] Department of Mathematics and Informatics, University of Amsterdam, The Netherlands

escudero@uva.nl

10.21437/Interspeech.2007-525

Abstract

In this paper, we evaluate the adequacy of two widely used machine learning algorithms and a computational linguistic proposal to model L2 perceptual development. The three proposals are, in order, Nearest Neighbor, Naive Bayesian and Stochastic OT and the Gradual Learning Algorithm. We compared the three models' outputs to those of Spanish learners of Dutch who were asked to categorize synthetic stimuli as one of the 12 Dutch vowels. The empirical results of the human learners show that L2 learners differ significantly from native listeners, but also that their perceptual spaces tend to become more native-like with L2 proficiency. The results of the simulations show that all three algorithms are able to model listeners' data to a certain extent but that Stochastic OT and the Gradual Learning Algorithm, i.e. the linguistic model, best reproduces L1 and L2 data.

1. Introduction

Anecdotal and empirical evidence shows that learning to perceive the sounds of a second language (L2) is a difficult task. Models such as the Speech Learning Model [1] and the Perceptual Assimilation Model [2] are among the most cited frameworks for explaining L2 perceptual development. However, they do not consider a formal/computational implementation of their proposals, which makes their predictions difficult to test and validate. Recently, a computational, formal linguistic framework, namely Stochastic OT [3] and the Gradual Learning Algorithm [4], has been applied to L2 perceptual learning [5, 6].

In an attempt to bridge the gap between two disciplines that deal with language learning, viz. linguistics and artificial intelligence, we compare the explanatory power of this linguistic formal model with two widely used learning algorithms within machine learning, namely Nearest Neighbor [7] and Naive Bayesian [8]. We chose these algorithms because they represent different paradigms with regard to the level of abstraction assumed in perceptual categorization, ranging from merely saving the data without any form of abstraction to storing only very abstract representations. Below, we first present empirical data from Dutch natives and L2 learners of Dutch and then the results of our simulations for each of the three different models.

[*] The first three authors' names are in alphabetical order. Kastelein and Weiand conducted the simulations and Escudero contributed with the empirical data, guidance for simulations, overview of the research, and writing of articles. All four authors contributed to the development of the mathematical analysis, while Van Son and Escudero conducted the statistical analysis.

2. Native and L2 perception data

2.1. Methodology

Listeners: We tested 22 Dutch native listeners (11 males, 11 females) and 23 Peninsular Spanish speakers (10 males, 13 females) living in the Netherlands. Fourteen of the learners had beginning Dutch proficiency and nine advanced proficiency, as determined by a general comprehension test part of Dialang (www.dialang.org), which is a language assessment system based on the Council of Europe's Common European Framework of reference for language learning. The beginning learners had spent an average of 4 months in the Netherlands, while the advanced an average of 3 years.

Stimuli & procedure: The listeners heard 113 synthesized vowels with 14 F1 values and 10 F2 values. Along the F1 dimension, the values of the stimuli ranged from 240 to 900 Hz, while along the F2 dimension, they ranged from 580 to 2700 Hz. The 14 F1 and 10 F2 values were equally distant along a mel scale. The tokens used for this paper had the same vowel duration, i.e., 200 ms.

Listeners were asked to classify each of the tokens as one of the 12 Dutch vowel monophthongs, namely /i, ɪ, y, u, ʏ, e, ɛ, ø, o, ɔ, ɑ, a/. The response options were the orthographic representations of these Dutch vowels, namely <ie, i, uu, oe, u, ee, e, eu, oo, o, a, aa> respectively, which were presented on a computer screen. The experiment was conducted using a Praat [8] experiment file which automatically recorded the listeners' mouse clicks on the vowel responses. On average, listeners took 25 minutes to complete the task.

2.2. Analysis

We used F1 and F2 values of the stimuli in Hz to analyze the listeners' perceptual vowel spaces. A representation of each vowel was constructed (and plotted) as an ellipsoid region in the vowel space, where mean F1 and F2 values for a vowel defined the centroid of the ellipse, and the tilt and size of the ellipse were determined by the Eigenspace of the vowel-specific covariance matrix.

Once the centre and extents of the perceptual ellipses were established, we computed the Normalized Midpoint Distance (NMD) between pairs of ellipses, which is a measure that serves to quantify the distance between any two vowels in the vowel space. The NMD is determined by taking the Euclidean distance in Hz space between the centroids of a pair of vowels, given by the formula shown below, where *xi* and *yi* are the coordinates of the centroid μi of vowel *i*.

$$D_{\mu_1,\mu_2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

This distance is then normalized by the average of the distances between the centroid and the edge of the ellipse along the line connecting the two centroids. The radius along a line at angle θ, i.e., the distance between an ellipse's centroid and its edge along the line connecting the two centroids, was calculated with the formula below, where a and b are the lengths of the major and minor axes and Φ is the rotation of the ellipse.

$$r_i = \frac{1}{\sqrt{\left(\frac{\cos(\theta - \phi_i)}{a_i}\right)^2 + \left(\frac{\sin(\theta - \phi_i)}{b_i}\right)^2}}$$

Given the total distance between the centroids and the radii of the ellipses along the line connecting them, the NMD can then be calculated by the formula below, where v_1 and v_2 are vowel 1 and vowel 2 respectively.

$$NMD_{v1,v2} = \frac{\sqrt{2}D_{\mu1,\mu2}}{\sqrt{r_1^2 + r_2^2}}$$

Thus, the size of a vowel space in terms of non-overlapping vowel areas can now be expressed as the difference between pairs of NMD values. An average minimal distance is defined by first selecting a central vowel as the vowel with the smallest summed squared distances with all the other vowels, which in this case gives 11 pairwise differences per listener group. Thus, the root mean squared distance, i.e., the root of the average squared distance between this vowel and all the other vowels, was used to measure how well the vowels in a certain perceptual vowel space are separated. That is, with this computation, the amount of vowel overlap on a perceptual space can be computed.

2.3. Results

This section shows the perceptual categorization of the 12 Dutch vowels by the native Dutch listeners and the L2 learners. For purposes of clear visualization, only the resulting ellipses for the three corner vowels are plotted together with the middle vowel. However, for the statistical analysis presented below, the categorizations for all 12 vowels were considered. The middle vowel needed to be the same for all groups and we chose /øː/ because its summed-squared distance was the lowest for the learners, and the second lowest by a small margin for the natives. Figures 1 and 2 show the vowel ellipses of the natives and beginning L2 learners, respectively.

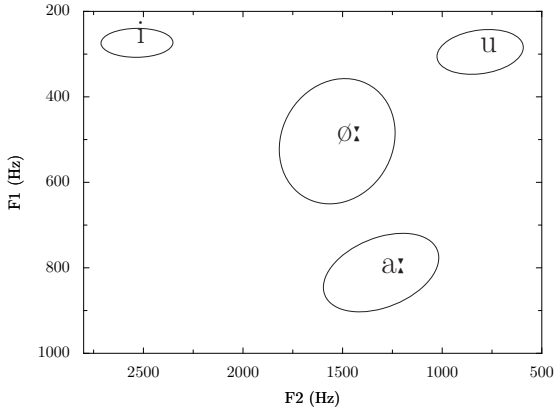

Figure 1: *Native Dutch listeners' categorization. Linear average (across 11) pairwise NMD difference = 246.29. x axis: F1 in Hertz, y axis: F2 in Hertz.*

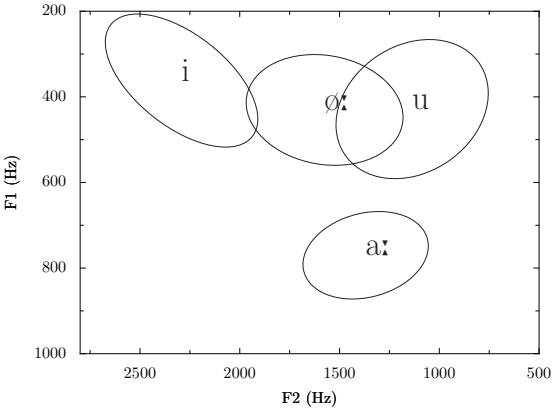

Figure 2: *Begnning L2 listeners' categorization. Linear average pairwise NMD difference* = 74.08

As we can see, the vowel space of the natives is much larger than that of beginning learners. This large perceptual space difference is confirmed by a Wilcoxon Matched Pairs Signed Ranks test on the 11 distances computed separately for the two groups (W+ = 61, W-= 5, $p \leq 0.00977$). The question now is whether L2 learners ever approximate a perceptual space similar to that of native listeners. Figure 3 shows the ellipses for the same vowels for the 7 advanced learners of Dutch.

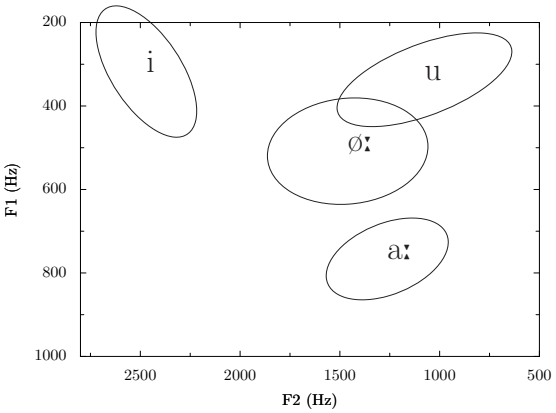

Figure 3: *Advanced L2 listeners' categorization. Linear average pairwise NMD difference = 142.67.*

A visual comparison between Figures 1 and 3 suggests that the advanced L2 learners have a perceptual space which approximates the native vowel space better than that of the beginning learners. We used the same statistical test as above to compare the 11 pairwise differences of the Dutch natives and the advanced learners. The test did not reach significance (W+ = 55, W- = 11, $p \leq 0.0537$). In contrast, a comparison of the 11 pair wise differences of the advanced learners with those of the beginning learners shows that they are significantly different (W+= 58, W- = 8, $p \leq 0.0244$). These two tests together can be taken to mean that the advanced learners' perceptual space is at least different from that of the beginning learners.

3. Modelling

3.1. The three models

The Nearest Neighbor (kNN) algorithm is a so-called lazy learner, meaning that it does not require complicated learning procedures, but rather performs the necessary calculations at the time of classification of new instances. For the algorithm to work, each instance must be expressible as occupying a location in a high-dimensional Euclidean space. Learning simply occurs through storing each example with its corresponding label, which results in a space populated by all instances stored up until the current point in time. To classify a new instance, the algorithm calculates the Euclidean distance to each example in the instance space and looks at the closest neighboring point. The label of this point is assigned to the new instance. Several different labels may be stored at a single point. In such a case, one label is probabilistically chosen on the basis of the resulting point-specific vowel frequency distribution.

The Naive Bayes (NB) classifier is a traditional statistical classification algorithm which relies on Bayes' well known formula for calculating conditional probabilities:

$$p(a \mid b) = \frac{p(a)p(b \mid a)}{p(b)}$$

During training, the classifier attempts to maximize the likelihood of the training data by using the relative frequency estimation over attribute values, under an assumption of conditional independence between variables. This gives the conditional probability of individual attribute values, $a_1, \ldots, a_n$ given a class, c_j, $P(a_i|c_j)$. In the vowel perception case of the present study,, each vowel constitutes a single class and the classifier builds the model by estimating the relative frequency of attribute-value occurrences for the class given in the training data. When trying to classify an incomming stimulus, , the resulting frequency distribution model is used to calculate the probability of observing each of the candidate vowels, given the stimulus. Then, a class label is chosen from this set on the basis of this probability distribution.

The Gradual Learning Algorithm (GLA) is an error-driven learning algorithm used in combination with Stochastic Optimality Theory (SOT) [3], which is a linguistic theory originally used in phonology but also in other areas of linguistics. Within this theory, the *optimal candidate* for classification is selected from a set of generated candidates by means of a hierarchy of soft constraints. SOT differs from traditional OT [9] in that constraint rankings are not discrete and ordinal but rather arranged on a continuous scale, which means that the distance between constraints can be learned from training data. The GLA takes a set of constraints and input-output pairs augmented with frequency information and subsequently adjusts the ranking values of the constraints so that the number of errors, i.e., cases where the optimal candidate for an input does not match the output, is minimized. In addition, the ranking value varies due to a noise component that is added during evaluation, which makes the selection of the optimal candidate non-deterministic.

3.2. Simulations

For the Dutch native simulation, the three algorithms were trained on 100% of the data from the 22 native Dutch listeners shown above. Figure 4 shows the resulting ellipses plots for each model. In the caption, we see the linear average pairwise NMD difference which each model yields between parentheses. The results of the same statistical test show that the simulated Dutch NB vowel space is significantly larger than that of the other two models (KNN: W+ = 66, W-= 0, $p \leq 0.00098$; SOT-GLA: W+ <= 66, W-= 0, $p \leq 0.00098$). In addition, the vowel space of simulated Dutch SOT-GLA is significantly larger than that of simulated Dutch KNN (W+ = 61, W-= 5, $p \leq 0.0098$). As for the human data, the Dutch NB vowel space is significantly larger than that of the Dutch human listeners (W+ = 66, W-= 0, $p \leq 0.00098$), while the Dutch humans were found not to differ significantly from either the KNN or SOT simulated Dutch (in both cases: W+ = 50, W-= 16, $p \leq 0.148$). This seems to suggest that both the KNN and the SOT-GLA simulations yield vowel spaces that compare well with the Dutch human data.

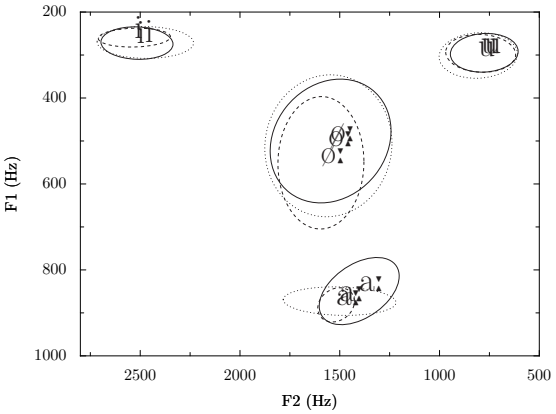

Figure 4: *Simulated Dutch listeners. KNN = solid (211.80), NB = dashed (462.61), SOT-GLA = dotted (260.15).*

The simulations of the L2 learners started as simulated native speakers of Spanish because we assume that the initial state of L2 acquisition is a *copy* of the L1 categorization strategies [6]. Thus, the models were first trained on data from monolingual Spanish listeners classifying the same tokens as the Dutch natives and L2 learners but choosing from the five Spanish vowels, /i, e, a, o, u/. Thus, the classifications in the Spanish models were mapped onto the Dutch vowel inventory using a probabilistic transformation. This mapping scheme was manually constructed on the basis of observed correspondences between the Spanish vowel space of the monolingual native speakers on the one hand, and the Dutch vowel space of the beginning learners on the other, and subsequently refined. Figure 5 shows the beginning L2 results for the three models (Average pair wise NMD differences in the caption).

When simulating beginning L2 learners, the NB model yields a significantly different perceptual space from those of both the KNN (W+ = 65, W- = 1, $p \leq 0.002$) and the SOT (W+ = 58, W-= 8, $p \leq 0.024$) models. In addition, the latter two models produce significantly different vowel spaces (W+ = 58, W-= 8, $p \leq 0.024$). As for the human L2 beginners, none of the models yields vowel spaces which are significantly different from the beginners' vowel space (NB: W+ = 43, W- = 23, N = 11, $p \leq 0.4131$; , SOT: W+ = 37, W- = 29, N = 11, $p \leq 0.7646$; KNN W+ = 27, W- = 39, N = 11, $p \leq 0.6377$), which seems to suggest that all three models succeed in modelling human beginning L2 learners.

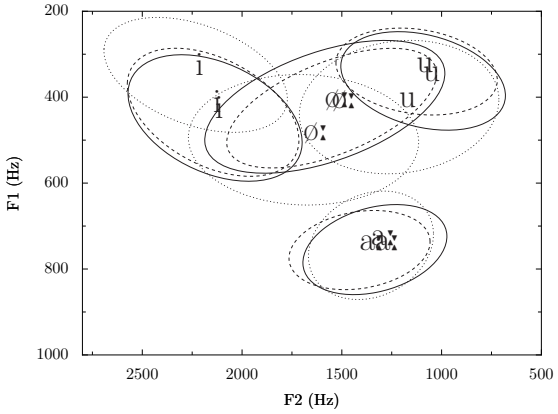

Figure 5: *Simulated Beginning learner of Dutch. KNN = solid (71.4025), NB = dashed (102.32), SOT-GLA = dotted (94.74).*

To simulate advanced learners of Dutch, we fed the native Spanish models with an additional 30% of randomly chosen training examples from the native Dutch listeners' data. It was assumed that 30% amounts to a third of the listeners' lives, i.e. approximately 8-10 years of L2 exposure. For these learners, the categorization output is either a Dutch vowel based on their L2 experience or a vowel from a Spanish-to-Dutch vowel mapping, according to the probabilistic mapping scheme that was used before. Figure 6 shows the simulated advanced learners' results (average pair wise NMD difference in the caption).

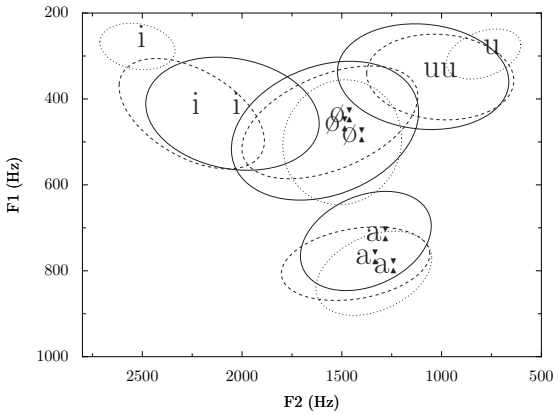

Figure 6: *Simulated Advanced learners. KNN = solid (86.18), NB = dashed (116.26), SOT-GLA = dotted (225.59).*

Thus, for advanced learners the SOT-GLA model yields a significantly different vowel space from those of both the KNN (W+ = 64, W- = 2, N = 11, $p \leq 0.00293$) and the NB (W+ = 60, W- = 6, N = 11, $p \leq 0.01367$) models. The latter two models also differ significantly from one another (W+ = 59, W- = 7, N = 11, $p \leq 0.01855$). As for the human advanced learners, neither the NB (W+ = 51, W- = 15, N = 11, $p \leq 0.123$) nor the SOT-GLA models (W+ = 55, W- = 11, N = 11, $p \leq 0.05371$) yield vowel spaces which are significantly different from the advanced listeners' vowel space, while the KNN does (W+ = 60, W- = 6, N = 11, $p \leq 0.01367$).

4. Discussion and conclusions

Our results show that it is possible to quantify the dimensions of perceptual vowel spaces in both humans and simulated native and non-native listeners and how they are learned. Moreover, it has been shown that the characteristics of each type of perceptual vowel space can be compared meaningfully.

Our quantitative comparison of human native versus L2 learners shows that beginning learners start with a smaller perceptual space than that of the natives but that with experience with the new language, advanced L2 learners can develop to mimic the native perceptual space.

With respect to the main aim of the present study, i.e. to compare models of human perceptual development, the choice of models/algorithms was specifically designed to give a continuum between amounts of information reduction in the perceptual storage of vowel tokens. That is, the KNN stores every vowel exemplar while the NB reduces all examples to probability distributions and SOT-GLA reduces them to discrete abstract categories and constraints. Our simulations show that NB does not succeed in reproducing the Dutch humans, KNN does not succeed in reproducing advanced L2 learners, while SOT-GLA succeeds in reproducing all three sets of human data, i.e. native, beginning and advanced.

We argue that our simulations indicate that models of human sound perception which aim at explaining both L1 and L2 learning should be based on algorithms/frameworks which allow for strong abstraction and generalization, such as the SOT-GLA model which perfomed well in all our simulations.

5. References

[1] Flege, J. E. Second language speech learning theory, findings, and problems. In W. Strange (Ed.), Speech perception and linguistic experience: Issues in cross-language research. Timonium, York Press, MD, 1995, pp. 233-277.

[2] Best, C. T. A direct realist view of cross-language speech perception. In W. Strange (Ed.), Speech perception andlinguistic experience: Issues in cross-language research. Timonium, York Press, MD, 1995, pp. 171-206.

[3] Boersma, P. Functional phonology: Formalizing the interactions between articulatory and perceptual drives. Holland Academic Graphics, The Hague, 1998.

[4] Boersma, P. & Hayes, B. Empirical tests of the gradual learning algorithm. Linguistic Inquiry, Vol. 32, 2001, pp. 45–86.

[5] Escudero, P. & Boersma, P. Bridging the gap between L2 speech perception research and phonological theory. St. in Sec. Lang. Acq., Vol. 26, 2004, pp. 551-585.

[6] Escudero, P. Linguistic perception and L2 acquisition: Explaining the attainment of optimal phonological *categorization*. LOT dissertation series 13, Utrecht University, 2005.

[7] Lin, J-H & Scott Vitter, J. A theory for memory-based learning. Technical report DUKE–TR–1993–29, 1993.

[8] Mitchell, T.M. Machine Learning. 1997.

[9] Boersma, P. & Weenik, D. Praat: doing phonetics by computer [Computer program]. Retrieved in 2006 from http://www.praat.org.

[10] Prince, A. & Smolensky, P. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers Center for Cognitive Science, 1993.