Comparing methods to find a best exemplar in a multidimensional space

Titia Benders, Paul Boersma

Institute of Phonetic Sciences, University of Amsterdam, Spuistraat 210, 1012 VT, The Netherlands
Titia.Benders@uva.nl, Paul.Boersma@uva.nl

Abstract

We present a simple algorithm for running a listening experiment aimed at finding the best exemplar in a multidimensional space. For simulated humanlike listeners, who have perception thresholds and some decision noise on their responses, the algorithm on average ends up twelve times closer than Iverson and Evans' algorithm [1].

Index Terms: listening experiments, prototypes, search algorithm

1. Introduction

The best exemplars, or *prototypes*, of phoneme categories are thought to play an important role in speech perception and language acquisition [2]. However, vowel exemplars that listeners perceive to be the best exemplars are at more extreme positions in the vowel space than their average productions [3]. Perception experiments are thus necessary to find these best exemplars.

Since every phoneme has multiple acoustic properties, the search for the best exemplar of a given category has to be conducted in a multidimensional acoustic space. Iverson and Evans [1] (henceforth "I&E") note that it is difficult to obtain goodness judgments from listeners across a whole multidimensional space, and therefore present an experiment procedure (henceforth "algorithm") that is based on fast interpolation techniques from the literature on numeric computation. Their algorithm draws a number of well-chosen lines in the multidimensional space and requires listeners to provide goodness judgments for only five or six vowels per line.

In section 2 we explain why I&E's fast interpolation technique may not work well with human listeners. In section 3 we therefore present a simpler and "slower" algorithm that approaches the best exemplar with fixed decreasing step sizes instead. Simulations in section 4 confirm that this algorithm is indeed more robust to humanlike, noisy responses.

2. Iverson and Evans' procedure

Suppose that you are a participant in an experiment that aims at finding the sound that you consider to be the best possible exemplar of the vowel /a/. I&E's algorithm can help you find that best exemplar. The following two sections discuss the one– and four-dimensional cases, respectively.

2.1. In one dimension: goodness interpolation

If there is only one acoustic dimension in which sounds can vary, the algorithm starts by presenting you the two sounds with the smallest and the largest possible values on the axis (a_1 and a_2 in Fig. 1). It asks you to give goodness judgments in terms of the perceived distance to your prototype, with lower judgments indicating closer, i.e. better, sounds. These are d_1 and d_2, respectively. The algorithm then computes its first estimate of what you might find a good /a/ exemplar, by taking an average of the previous two values, weighted by your goodness ratings:

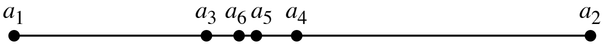

Figure 1: *Iverson and Evans' goodness interpolation.*

$$a_3 = a_1 \cdot \frac{d_2}{d_1 + d_2} + a_2 \cdot \frac{d_1}{d_1 + d_2} \tag{1}$$

If for example $d_1 = 2.5$ and $d_2 = 5.0$, the algorithm regards a_1 as closer to your prototype than a_2, and a_3 will be at one third along the route from a_1 to a_2 (see Fig. 1). The algorithm then presents you the sound a_3. If its perceived distance to your prototype, d_3, is smaller than both d_1 and d_2, the algorithm computes a new estimate of where your best exemplar could be, by means of parabolic interpolation over the hitherto best exemplar (a_3) and its two neighbours (a_1 and a_2):

$$a_4 = a_3 - 0.5\,\frac{(a_3 - a_1)^2(d_3 - d_2) - (a_3 - a_2)^2(d_3 - d_1)}{(a_3 - a_1)(d_3 - d_2) - (a_3 - a_2)(d_3 - d_1)} \tag{2}$$

When a_4 is played and judged, a similar parabolic interpolation (on the hitherto best exemplar and its two neighbours) is performed to obtain the next estimate a_5, and analogously for a_6. If in any of the last three steps the hitherto best exemplar is a_1 or a_2 (which don't have two neighbours), then the next estimate is obtained by computing a weighted average between the hitherto best exemplar and its sole neighbour, using Eq. (1). In the end, the location that the algorithm regards as the final best is the value of the set $\{a_1, a_2, a_3, a_4, a_5, a_6\}$ that you have judged to have the lowest perceived distance to the prototype.

2.2. The search in multiple dimensions

In a multidimensional space, the search for the best exemplar can consist of multiple instances of the line optimization procedure of the previous section. Evans and Iverson [4] ("E&I") search the best vowel exemplars through a sequence of five line optimizations in a four-dimensional vowel space defined by F1, F2, F3, and duration. The quadrilateral in Fig. 2 (it's *almost* a triangle) shows what the algorithm regards as the outer bounds of your auditory vowel space defined by F1 and F2. The location of your best exemplar of /a/ is the circle in the quadrilateral, but at the beginning of the experiment nobody knows its location yet. The algorithm does have an initial guess of where your best /a/ exemplar might be found (shown as a_0 in the Figure), and an estimate of the centre of the vowel space (shown as a schwa). In the first stage of the experiment, the algorithm draws a line through schwa and a_0, which intersects the vowel quadrilateral in the points a_1 and a_2. Using Eqs. (1) and (2), the algorithm will play you six vowel exemplars along the line

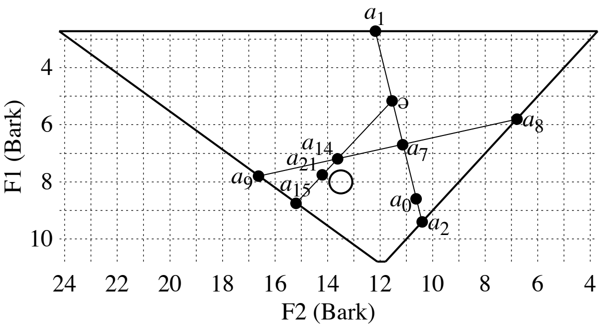

Figure 2: *Three of Evans and Iverson's line optimizations.*

piece $a_1 - a_2$ and estimate your best /a/ exemplar (shown as a_7).

In the second stage, the algorithm draws a line that runs through a_7 and is perpendicular to the first line.[1] This line intersects the quadrilateral in the points a_8 and a_9. Using (1) and (2) again, the algorithm computes a new estimate of your best /a/ exemplar along the line piece $a_8 - a_9$, given as a_{14} in the Figure. The best F1 and F2 values found are passed on to the third stage and fourth stage, in which one-dimensional optimizations of duration and F3 are performed. In the fifth stage, the algorithm draws a line through three-dimensional space: from the F1, F2 and duration values of schwa, through the F1 and F2 of a_{14} combined with the best duration found thus far. This line intersects the boundary of the F1–F2 space at a_{15}. The algorithm then searches for the best exemplar along this line, changing F1, F2 and duration simultaneously and ending up with a_{21}, which is very close to your real best exemplar.

2.3. Problems with the goodness interpolation method

The basis of the goodness interpolation method are the listener's goodness judgments to the stimuli. However, rating the closeness of some observation to an abstract ideal is an intrinsically difficult task and we doubt that listeners are able to give reliable goodness judgments. An additional problem for goodness judgments are stimulus order effects: a moderately good exemplar played after a particularly bad one will be judged to be better than that same exemplar played after a very good one. Especially listeners that are not used to participating in perception experiments might have difficulties with rating the goodness of sound stimuli. Such noise in the goodness judgments may cause a_3 to be judged worse than a_1 or a_2, or a_4 to be judged worse than a_3. Every time an acoustic value is judged worse than its neighbours, the algorithm steers off to the left or right side of that value, and will never be able to pass that value again. We therefore set out to design an algorithm that can search in a multidimensional space without goodness judgments and is robust against reasonable amounts of noise.

3. An algorithm with pairwise comparisons

The alternative algorithm we propose is an instance of an optimization technique that is "robust" rather than "fast", namely local hill-climbing with gradually decreasing step size. The basic working is that the algorithm repeatedly asks you, the listener, to compare the quality of two sounds, namely the sound that the algorithm currently thinks is your best exemplar and a sound that is acoustically some distance removed from it. Whenever you judge the latter sound to be a better exemplar, the algorithm accordingly changes its opinion of what your best exemplar is. We illustrate this for one and for four dimensions.

3.1. In one dimension: hopping the line

Fig. 3 shows a single acoustic continuum, like Fig. 1. The algorithm's initial guess of where your best exemplar is (a_1), lies in the middle of the continuum, at 50%. The algorithm computes two new sounds, which are removed from a_1 by a distance of $50\% \cdot 0.7 = 35\%$ of the extent of the continuum: a_1' and a_1''. The algorithm then asks you which of a_1' or a_1'' is a better exemplar. If you find a_1' better, the algorithm will ask you to compare a_1' and a_1. If a_1' "wins" again, it will become the algorithm's new guess of your best exemplar; if you judge a_1 and a_1' as equally good, the average of a_1 and a_1' will become the algorithm's new best guess; and if a_1 wins, it will remain the algorithm's best guess. If on the first trial, you were to judge a_1'' better than a_1', the algorithm will make you compare a_1 and a_1'', which will result in the algorithm's new best guess being a_1, a_1'', or their average. In Fig. 3, you chose a_1' as better than a_1 and a_1'', and this is therefore promoted to be the algorithm's new best guess, a_2. In the second stage, the step size is reduced to $50\% \cdot (0.7)^2 = 24.5\%$ of the size of the continuum. This gives the points a_2' and a_2'' on both sides of a_2 (since values outside the continuum are non-acoustic, a_2' falls at the left edge). The first comparison the algorithm will have you make is that between a_2 and a_2', since the algorithm remembers and honours the direction of the previous step taken. If you then judge a_2' as worse than a_2, but a_2 and a_2'' as equally good, the algorithm's next best guess will be a_3, right in the middle of the latest two sounds. In the next stage, the step size is again multiplied by 0.7, and the algorithm will let you compare a_3 and a_3''. If you judge a_3'' better than a_3, the algorithm will turn it into a_4. Next, again with a smaller step, the algorithm plays a_4 and a_4'' to you, but you prefer a_4; the algorithm then plays a_4 and a_4', but again you prefer a_4, which therefore turns into a_5. The following smaller step size (comparing $a_5' - a_5''$, then $a_5' - a_5$) takes you to a_5', which turns into a_6. With the last smaller step size you go to a_6'', the algorithm's final guess of your best exemplar.

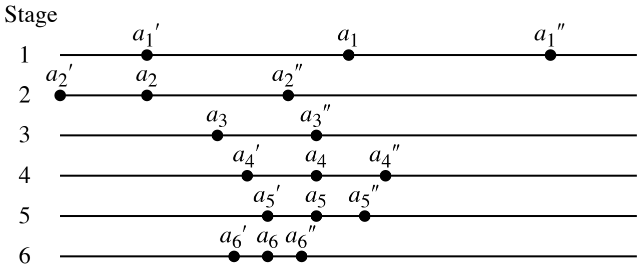

Figure 3: *Stepwise optimization along a single line.*

3.2. In multiple dimensions: interleaving

The stepwise method could perform line-by-line optimizations, but it also lends itself to *interleaved line optimizations*. This

[1] E&I do not say what they regard as perpendicular. It seems likely that in their case the two lines are perpendicular within a space in which F1 and F2 are expressed in Hertz; this overestimates the importance of F2 with respect to F1, so in Fig. 2 and our simulations of the I&E method below, we express F1 and F2 in the psycho-acoustic Bark scale.

means that it can perform a step size of 35% on the F1 continuum, then a step size of 35% on the F2 continuum, then on the duration continuum, then on the F3 continuum, after which it can do a step size of 24.5% on each of the four continua, and so on. The advantage of interleaving is that no energy is spent on small differences in e.g. F1 if the F2 is still far removed from the prototype; instead, the initial phases of the search bring all dimensions to an approximately correct value, after which more detailed optimizations become more relevant.

An important technical detail about step size is that the size of a continuum may depend on the values on other continua. In Fig. 2, for instance, the size of the F2 continuum is smaller for high F1 values than for low F1 values. For high F1 values, the step size in the F2 continuum is scaled down accordingly.

Beside the possibility of interleaving the four dimensions, an important difference with I&E's goodness interpolation method is that our decreasing-step-size method seems to be less sensitive to noise. In Fig. 3, for instance, the search moved to the left of a_1 but could easily have moved to the right of a_1 later on (e.g. via a''_4) if the initial move to the left was a mistake. No such repair is possible in I&E's method.

4. Comparing the algorithms by simulations

We now present computer simulations that investigate the effect of noise on the outcomes of the two algorithms. Both algorithms search the best exemplar in a four-dimensional acoustic space defined by F1, F2, F3 and duration. Duration is expressed as the natural logarithm of the duration in milliseconds, with 4 as the lower limit and 6 as the upper limit. The formant space is limited by the following constraints:

$$2.73\ \text{Bark} \le \text{F1} \le 10.796\ \text{Bark} \tag{3}$$

$$\text{F1} + 1.0\ \text{Bark} \le \text{F2} \le 28.3\ \text{Bark} - 1.5 \cdot \text{F1} \tag{4}$$

$$14.5\ \text{Bark} \le \text{F3} \le 18.5\ \text{Bark} \tag{5}$$

$$\text{F2} + 0.5\ \text{Bark} \le \text{F3} \tag{6}$$

4.1. Properties of the simulation

In the simulations, the goodness interpolation algorithm searched as described in sections 2.1 and 2.2, with five stages and six responses per stage, and the stepwise algorithm searched as described in sections 3.1 and 3.2, with six cycles through the four dimensions. The auditory discrimination threshold, or JND, for the formants was set to 0.3 Bark [5]. The JND for duration was set to 0.1 in the natural logarithm domain, which corresponds to a duration ratio of 10.5%.

For the simulations, we created 10000 virtual participants, each with a best exemplar (for /a/) drawn from the following uniform distributions: F1 between 7 and 9 Bark; F2 between 13 and 15 Bark; F3 between 15.5 and 17.5 Bark; and duration between 5.35 and 5.85 ln-ms. To ensure that a best exemplar fell within the possible vowel space, F1 was sampled first, then F2 was sampled within the range allowed by this F1, and F3 was sampled last.

Each sampled best exemplar, or "participant", participated in six experimental conditions: for each of the two algorithms they performed the experiment under three noise conditions. The first noise condition simulated unhumanly perfect listeners, the two other conditions simulated increasingly more realistic listeners.

The success of an individual participant in an experimental condition is expressed as the Euclidean distance in JND units (in the four-dimensional space of F1, F2, F3 and duration) between the exemplar found by the algorithm and the true best exemplar (prototype) of the participant. The larger this distance, the lower the success of the participant in finding an exemplar close to his /a/ prototype. We express the success of an entire condition as the root mean square (rms) of the 10000 Euclidean distances. The higher the rms, the larger the average distance between the found exemplar and the best exemplar, and/or the larger the variation in the success of the participants. As a second measure of a participant's success in an experimental condition, we determine whether the found exemplar could be considered close to the best exemplar. We arbitrarily consider two vowel tokens as close to each other if the Euclidean distance in JND units between the two is equal to or lower than 4. The success of an entire condition can then be the percentage of participants for who the algorithm finds an exemplar close to their prototype.

4.2. Results

Fig. 4 displays the success of both optimization methods in each of the three noise conditions in rms and the percentage of participants that have not found an exemplar that is close to their best exemplar. For both measures, a higher value indicates a lower success. The noise conditions and the results are discussed below.

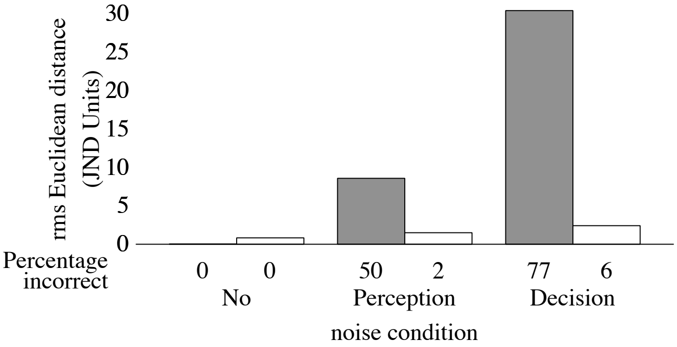

Figure 4: ***Top****: The rms of the goodness interpolation method (grey) and the stepwise method (white) in the four noise conditions.* ***Bottom****: The error rates of the goodness interpolation method (left) and the stepwise method (right), expressed as the percentage of participants failing to come close to their true best exemplar.*

The first condition is the **No Noise** condition, which simulates listeners with unhumanly perfect ears who make no response errors. For the I&E algorithm, the goodness rating these virtual listeners give to an exemplar is its Euclidean distance to the best exemplar in JND units. For the stepwise algorithm, the virtual listeners compute for each exemplar of the pairwise comparison the Euclidean distance to their best exemplar in JND units and choose the exemplar for which the distance thus computed is smaller.

The results show that I&E's goodness interpolation algorithm is the better performing algorithm in the absence of noise, although in the stepwise method, too, all 10000 participants find an exemplar that is closer than 4 JND to their best exemplar.

The **Perception Noise** condition simulates listeners with humanlike perceptual thresholds who make no response errors. These listeners perceive tokens as having acoustic values that are randomly sampled from a normal distribution around the actual value, with a standard deviation of 0.5 JND. For I&E's algorithm, our virtual listeners compute the goodness rating from these perceived values. For the stepwise algorithm, they judge two exemplars as "Equal" if these perceived values are within one "Euclidean" JND or if the perceived distances to the best exemplar are within one "Euclidean" JND.

The rms computations show that in the Perception Noise condition the performance of the goodness interpolation method is much worse than in the No Noise condition. By contrast, the stepwise method is only 1.8 times worse than in the No Noise condition, and it is almost 6 times better than the goodness interpolation method in the same (Perception Noise) condition. This indicates that the stepwise method is more robust to noise than the goodness interpolation method. Detailed inspection of the individual results shows that the success of the stepwise model is partly due to the "Equal" answers in the Perception Noise condition. When a participant answers "Equal", the algorithm infers that the best exemplar will be in between the two values, which is usually a correct inference.

The third condition is the **Decision Noise** condition, which adds some decision noise to the Perception Noise condition. For I&E's algorithm, the goodness judgments that our virtual listeners give are drawn from a Gaussian distribution with the correct goodness judgment (perceived distance to prototype) as the mean and a relative standard deviation of 10%. For the stepwise algorithm, the listener's judged distance from each of the two exemplars to the prototype is drawn from a Gaussian distribution with the perceived distance as the mean and a 10% relative standard deviation.

The Decision Noise condition simulates listeners with humanlike ears, who try to give a good response on every trial, but are subject to slight decision noise: the best human participants one can possibly get. Adding decision noise makes the rms for I&E's algorithm 3.5 times higher than with perception noise only, and the majority of the participants do not find an exemplar close to their best exemplar. The addition of decision noise makes the stepwise model about 1.5 times worse, but the overall performance of the method is still very good, with only 6% of the participants finding a bad exemplar. The rms of the stepwise method in this condition is about 12 times smaller than the rms of the goodness optimization method. These results illustrate again how sensitive the goodness optimization method is to noise, which will come naturally with human participants, and how robust the stepwise method is.

A last point to note here is that all of the experiments with I&E's algorithm have taken 30 trials per participant. The average number of trials needed in the Decision Noise condition with our stepwise algorithm was 37.04, with a standard deviation of 2.72. The robustness of the stepwise method thus comes at the cost of being only slightly slower.

5. Discussion and conclusion

In our current implementation of the stepwise method, it is only possible to move along one-dimensional axes. Imagine, though, that the perceptual space were formed in such a way that the good exemplars of the vowel /a/ lie on a steep ridge that runs diagonally through the F1–F2 space, with the best exemplar on the highest point of this ridge. If the method ends up on this ridge, but not on its highest point, a change in either F1 or F2 will decrease the goodness and only a change in F1 and F2 simultaneously will lead to the best exemplar. It might thus be necessary to add "diagonal" axes to the stepwise method.

Oglesbee and De Jong [6] ("O&J") present a third method to find best exemplars in a multidimensional space. Similarities with our method are that it performs interleaved optimizations of the multiple continua and works with comparisons of three exemplars. Their strategy for finding the next best guess is intermediate between I&E's and ours: like I&E's, it is based on goodness interpolation (between the two exemplars judged best), but like ours, it has a bias towards the best judged exemplar. However, a large difference between O&J's method on the one hand and I&E's and our methods on the other hand, is that O&J's method does not "zoom in": the two end points (of the three under consideration) are always separated by a distance of two-thirds of the acoustic continuum. Depending on the width and skewness of the listener's goodness function on the continuum, this can lead to extremely slow convergence, large oscillations, or getting stuck at an edge of the continuum.

In this paper we have compared two methods to find the best exemplar of a phoneme category in a multidimensional acoustic space: Iverson & Evans' goodness interpolation algorithm and a stepwise algorithm based on pairwise comparisons. We first argued that the stepwise optimization method requires an easier response from participants, namely a choice between two exemplars, than the goodness interpolation algorithm, in which participants have to give goodness judgments to individual exemplars. Our simulated experiments subsequently showed that if listeners have perceptual thresholds and some decision noise in their responses, as most human participants will have, the stepwise method outperforms the goodness interpolation algorithm. In general, these results show that mathematically fast models, such as Iverson & Evans' goodness interpolation algorithm, require high precision in the responses, in order to come close to the participant's best exemplar. Humans are unlikely to give such responses. For finding real humans' best exemplars in a multidimensional space a simple, robust algorithm, as our stepwise algorithm, will do better.

6. Acknowledgments

We would like to thank Don van Ravenzwaaij, Tessa Verhoef and Bart de Boer for useful discussions.

7. References

[1] Iverson, P. and Evans, B.G., "A goodness optimization method for investigating phonetic categorization", Paper presented at the 15th International Conference of Phonetic Sciences, 2003.

[2] Kuhl, P.K., "Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not", Perception & Psychophysics, 50(2):93–107, 1991.

[3] Johnson, K., Flemming, E. and Wright, R., "The hyperspace effect: phonetic targets are hyperarticulated", Language, 69:505–528, 1993.

[4] Evans, B.G. and Iverson, P., "Vowel normalization for accent: an investigation of best exemplar locations in northern and southern British English sentences", Journal of the Acoustical Society of America, 115(1):352–361, 2004.

[5] Kewley-Port, D. and Zheng, Y., "Vowel formant discrimination: towards more ordinary listening conditions", Journal of the Acoustical Society of America, 106(5):2945–2957, 1999.

[6] Oglesbee, E. and De Jong, K., "Searching for best exemplars in multidimensional stimulus spaces", Journal of the Acoustical Society of America, 122(4):EL101–EL106, 2007.