ChaLeT: DETERMINING THE CHANCE LEVEL IN PERCEPTION EXPERIMENTS BY MEANS OF BINOMIAL TESTS

Wendy Elvira-García[1], Paolo Roseano[2 3], Assumpció Rost Bagudanch[4]

[1]Universidad Nacional de Educación a Distancia, [2]Universitat de Barcelona, [3]University of South Africa, [4]Universitat de les Illes Balears
welvira@flog.uned.es, paolo.roseano@ub.edu, assumpcio.rost@uib.es

ABSTRACT

This paper presents *ChaLeT*, an *R* script that can be used to determine statistically whether answers given by listeners in a perception test are above chance level. The script has been designed for perception tests where listeners have to choose between two categories. The script makes use of the binomial test of statistical significance to determine the chance level. It then plots the results of the perception test on a chart where coloured ribbons signal the interval where answers are below the chance level.

Keywords: perception test, chance level, binomial test, R

1. INTRODUCTION

In perceptual phonetics, researchers often use identification tests where listeners have to choose between two possible answers. In studies about intonation, for example, listeners can be asked to listen to an utterance (or part of an utterance) and to decide whether it is a statement or a yes-no question. Once the researchers have collected and counted the answers provided by the listeners, they have to decide whether the percentage of listeners that have given an answer is above the chance level or not. This step is crucial for further interpretation of the results.

In spite of being a fundamental methodological decision, the determination of the chance level is not always carried out with a method generally accepted by the scientific community. This paper puts forward a statistical tool that allows the researchers to decide whether the answers obtained in a perception test are significantly above the chance level. The tool is called *ChaLeT* (acronym of Chance Level Test) and consists of an *R* script, which runs binomial tests of statistical significance.

In the following sections, we shall first show that one of the dominant models in perceptual phonetics, the Signal Detection Theory, does not address explicitly the question of chance level (Section 1.1). Section 1.2 sets the specific goals of this paper. In Section 2, we describe the method that we propose. Namely, we give details about the input that the script needs (Section 2.1), about the statistical test it runs (Section 2.2), about the availability of the script (Section 2.3) and about the output (Section 2.4). In Section 3, we offer an example of how the script can be used in identification tests. Section 4 contains the conclusions and envisages some future developments.

1.1. Chance level in Signal Detection Theory

When facing perceptual phonetics, the important issue is to determine whether a participant may identify a signal as a particular category or not (an identification task) o to establish if he/she can distinguish two or more elements through different matching procedures (a discrimination test). There are different ways to process the results of such experiments, being one of them the Signal Detection Theory model (known as SDT), which was not developed for phonetics use (or at least, not only, since it is employed in many scientific fields) but is widely accepted in analysing phonetic perceptual data (Elman, 1979; Goldinger, 1998; Connell, 2000).

Within this framework, the researcher presents the participants some stimuli, which include both the signal (the object of study) and noise. The participants' responses to these may be hits (correct identification of the signal), misses (omission of the signal), false alarms (incorrect identification of the noise as the signal) or correct rejections (correct

detection of the noise). SDT is concerned with the hits and the false alarms to set an adequate panorama of the perception process. From these data, it can provide two basic parameters: the sensitivity of the signal (its strength) and the strategy of the participant in adopting a decision. Their combination allows displaying a picture for the strength of a signal and its degree of perceptiveness by the listeners (Wickens, 2001; Macmillan & Creelman, 2005).

Although this model provides very valuable information, it offers no clear and direct information about chance level. Of course, one of the representations of the results within SDT, the ROC curves, show in a graphic way the chance threshold, which is assumed to be at 50%. Nevertheless, presuming that chance level implies only this limit is problematic: why 50% of hits corresponds to decisions made by chance and not 49% or 55%? This seem to be a gap in the model, which lacks an essential complementary figure to the great amount of details it can provide.

The lack of a methodological consensus is even more evident if one reviews recent studies in the field. For example, if we focus on studies that present the results of identification tests (like Ladd & Morton, 1997; Remijsen & van Heuven, 1999; Post, 2000; Chen, 2003; Schneider & Linftert, 2003; Cummins, Doherty, & Dilley, 2006; Falé & Hub Faria, 2006; Schneider, Dogil, & Möbius, 2009; Dilley, 2010; Vanrell et al., 2013; Vanrell, Armstrong & Prieto, 2017, among several others), we observe that the authors have taken different methodological decisions. In some cases, chance level is not considered. In other cases, it is established in non-statistical ways. Other studies (e.g. Roseano et al., 2015) use statistics, but do not use the most appropriate test.

1.2. Goals

The general goal of this paper is helping researchers who run perception tests to establish when the results obtained pass chance level. Our specific objectives are 1) putting forward a standard test which can be used to determine chance level in tests where listeners have to choose between two categorical answers, 2) visualizing the results of such test in a clear and easily interpretable way.

2. METHOD

2.1. Input

The results of the perception test must be collected in an Excel sheet, which will serve as input for the script that runs statistical analysis. Table 1 shows the format of the Excel sheet, and a template is available online and can be downloaded from the webpage https://github.com/wendyelviragarcia/chance_levels_for_perceptual_tests.

It has to be pointed out that the input must be the *number* of answers given, not the *percentage*, since the statistical test we use (Section 2.2) is sensitive to the size of the sample. Another important requirement is that the number of answers is the same for all stimuli.

Table 1: Expected input of the pipeline.

Stimuli	Answer	nAnswers
1	Question identified as question	90
2	Question identified as question	91
3	Question identified as question	85
4	Question identified as question	77
5	Question identified as question	57
6	Question identified as question	47
7	Question identified as question	38
8	Question identified as question	20
9	Question identified as question	18
10	Question identified as question	20
11	Question identified as question	9
1	Statement identif. as statement	6
2	Statement identif. as statement	11
3	Statement identif. as statement	7
4	Statement identif. as statement	24
5	Statement identif. as statement	46
6	Statement identif. as statement	65
7	Statement identif. as statement	78
8	Statement identif. as statement	84
9	Statement identif. as statement	85
10	Statement identif. as statement	90
11	Statement identif. as statement	91

2.2. Statistical test

In order to determine the chance level, we consider every question asked to a judge as a Bernoulli trial (also known as binomial trial) (Papoulis, 1984), which in statistics is a random experiment with exactly two possible outcomes, and where the probability of each outcome is the same every time the experiment

is conducted. The result of a sequence of Bernoulli trials is called Bernoulli process. The expected outcome of a Bernoulli process is a Bernoulli distribution, which is a special case of a binomial distribution. The binomial distribution, on its turn, is the basis for the *binomial test* of statistical significance, which is the standard test used when the null hypothesis is that two categories are equally likely to occur (such as "head" and "tail" in a coin toss).

The classical statistical example for which a binomial test is used is determining whether a coin used for tossing is fair. If the coin is fair, the probability of getting heads is 50%. If I toss the coin 100 times and I get 100 "heads", the coin is evidently not fair. The same holds if I get 98 heads, or 97, 96, 95… Nevertheless, if I toss the coin 100 times and I get 51 heads and 49 tails, it does not necessarily imply that the coin has been tampered. In other words, if I get 51 heads and 49 tails, the result is still within the chance level. The question, therefore, is: how many heads/tails do I have to get in order to be sure that the coin is not fair?

This is the same question we were asking ourselves in Section 1.1. In perception tests, we need to know how many "correct" answers we need to get in order to be sure that listeners do not answer random.

The *binomial test* gives an answer to such questions. We set the significance level at 1% or 5%, we run a binomial test, and if the p-value we get is below .05, we reject the null hypothesis and we conclude that the coin is not fair. In perception tests, we would conclude that answers are above chance level.

2.3. ChaLeT

We designed an *R* (R Core Team, 2019) script called *ChaLeT* (Chance Level Test), which is available online (https://github.com/wendyelviragarcia/chance_levels_for_perceptual_tests) and that implements a binomial test (2.2) on the results of an identification test (2.1).

2.4. Output

The output of *ChaLeT* is a chart like the ones commonly used to plot the result of identification tests (like those in Figures 3 and 4 below). On the horizontal axis, one finds the stimuli and on the vertical axis the number of "correct" answers. In addition to this traditional elements of the plot, *ChaLeT* draws two coloured ribbons which signal the interval where answers are below the chance level according to binomial tests with a significance level of p<.01 (outer mauve ribbon) or p<.05 (inner plum-coloured ribbon) (Figure 5). If the points corresponding to a stimulus fall within this ribbon, judges' identification is below chance level. Researchers will then need to draw the phonological consequences of this.

3. EXAMPLIFICATION

In order to give an example of how *ChaLeT* can help the researchers interpret their results, in this section, we replicate in a simplified way an identification test presented in a well-known study about Catalan intonation (Vanrell et al., 2013) and we apply *ChaLeT* to the results.

Vanrell et al. (2013) show that Majorcan Catalan listeners distinguish information and confirmation seeking questions by means of two distinct nuclear falling pitch accents, which are represented schematically in Figure 1.

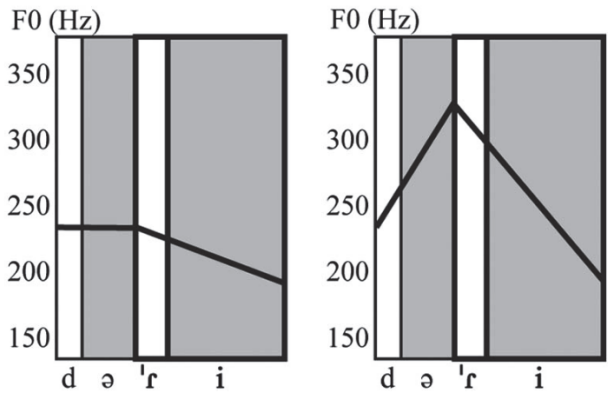

Figure 1: Simplified F0 contours of nuclear pitch accents of a confirmation-seeking (left) and an information-seeking yes-no question (right) (Source: Vanrell et al., 2013).

In order to carry out an identification test, the researchers manipulated F0 to create a continuum between the two stimuli. Figure 2 represents the original contours (solid lines) and contours obtained by means of manipulation in the two directions (dotted lines). In this case, the total number of steps was 11 and the distance between each step and the following was always the same.

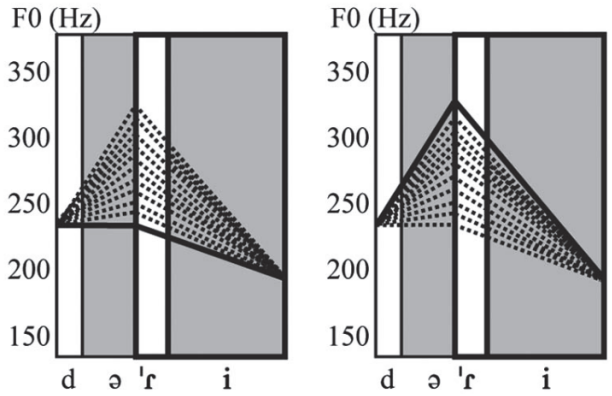

Figure 2: Simplified F0 contours of pitch accents obtained from a confirmation-seeking (left) and an information-seeking yes-no question (right) (Source: Vanrell et al., 2013).

In the identification test, listeners listened to all 22 stimuli and had to relate each stimulus to one of two categories (i.e. they had to decide if what they listened to was a confirmation-seeking or an information-seeking yes-no question). The results of the perception test were plotted on a diagram (Figures 3) where the horizontal axis contained the stimuli. On the vertical axis the researchers represented the percentage of "correct" answers, where "correct" means that "the judge considers that the stimulus corresponds to the category of the recording used to create the stimulus the judge hears".

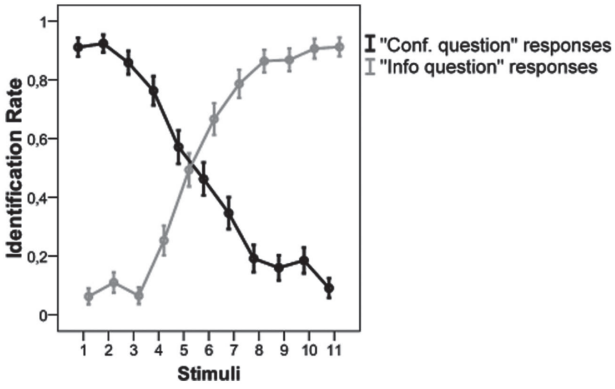

Figure 3: Results for the identification presented in Vanrell et al. (2013) (Source: Vanrell et al., 2013).

If we run *ChaLeT* with a dataset (Table 1) that approximately corresponds to the results obtained in the identification test presented above, we obtain a graph like the one in Figure 4. The inner coloured ribbons indicates the fringes where the *binomial test* says that answers are not different from chance level, at different significance levels (see Section 2.4).

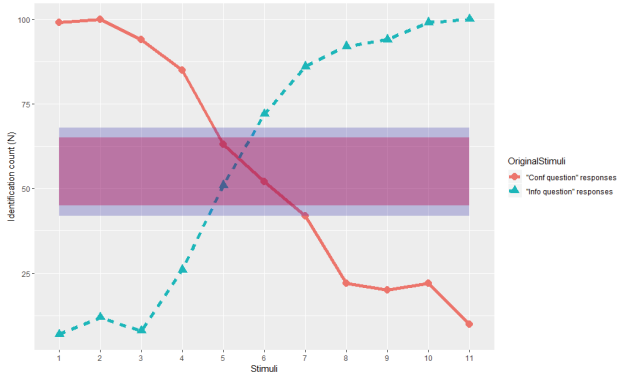

Figure 4: Results for an identification test similar to Vanrell et al.'s (2013) plotted by means of *ChaLeT*, with coloured ribbons marking chance level.

In comparison with the graphs traditionally used in perception studies (like the one in Figure 3), the graph in Figure 4 adds an important piece of information: it tells the researchers for which stimuli perception is at chance level. It goes without saying that this piece of information is relevant to draw conclusions about perception.

4. CONCLUSIONS

This paper describes a method for establishing the chance level in a perception test where listeners have to choose between two categories. The method is based on binomial test and is implemented by means of an *R* script called *ChaLeT*. The output of the script is a graph that represents the results of the identification task and marks chance level graphically by means of coloured ribbons.

Future versions of *ChaLeT* will need to use multinomial tests, so that the tool can be employed in perception tests with more than two possible answers.

5. REFERENCES

Chen, A. (2003). Reaction time as an indicator of discrete intonational contrasts in English. *Proceedings of Eurospeech*, Geneva, 97–100.

Connell, B. 2000. The perception of lexical tone in Mambila, *Language & Speech*, 43, 2163–2182.

Cummins, F., Doherty, C., & Dilley, L. (2006). Phrase-final pitch discrimination in English. In R. Hoffmann & H. Mixdorff (Eds.), *Proceedings of Speech Prosody* (pp. 5467–5470). Dresden: TUDpress Verlag der Wissenschaften.

Dilley, L. C. (2010). Pitch range variation in English tonal contrasts: Continuous or categorical?, *Phonetica*, 67, 63–81.

Elman, J. (1979). Perceptual origins of the phoneme boundary effect and selective adaptation to speech: A

signal detection theory analysis, *The Journal of the Acoustical Society of America*, 65, 190–207.

Falé, I., & Hub Faria, I. (2006). Categorical perception of intonational contrasts in European Portuguese. In R. Hoffmann & H. Mixdorff (Eds.), *Proceedings of Speech Prosody* (pp. 69–72). Dresden: TUDpress Verlag der Wissenschaften.

Goldinger, S. D. (1998). Signal detection comparisons of phonemic and phonetic priming: The flexible-bias problem, *Perception & Psychophysics*, 60, 952–965.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical?, *Journal of Phonetics*, 25, 313–342.

Macmillan, N. A. & Douglas Creelman, C. (2005). *Detection Theory: A User's Guide*. London: Erlbaum.

Papoulis, A. (1984). Bernoulli Trials. In A. Papoulis & S. Unnikrishna Pillai (Eds.), *Probability, Random Variables, and Stochastic Processes* (pp. 57–63). New York: McGraw-Hill.

Post, B. (2000). *Tonal and Phrasal Structures in French Intonation*. The Hague: Holland Academic Graphics.

R Core Team (2019). *R: A language and environment for statistical computing* (version 3.6.2). Retreived from http://www.R-project.org/

Remijsen, B., & van Heuven, V. (1999). Categorical pitch dimensions in Dutch: Diagnostic test. In J. J. Ohala, Y. Hasagawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 1865–1868). San Francisco: University of California.

Roseano, P., Fernández Planas, A. M., Elvira-García, W., Cerdà Massó, R., & Martínez Celdrán, E. (2015). Diferencias perceptivas entre los acentos tonales prenucleares en catalán. In A. Cabedo Nebot (Ed.), *Perspectivas actuales en el análisis fónico del habla: tradición y avances en la fonética experimental*, (pp. 163–173). València: Universitat de València.

Schneider, K., & Linftert, B. (2003). Categorical perception of boundary tones in German. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the XVth International Congress of Phonetic Sciences* (pp. 631–634). Barcelona: Causal Productions.

Schneider, K., Dogil, G., & Möbius, B. (2009). German boundary tones show categorical perception and a perceptual magnet effect when presented in different contexts. *Proceedings of Interspeech* (pp. 2519–2522). Brighton, UK, September 6–10.

Vanrell, M. M., Mascaró, I., Torres-Tamarit, F., & Prieto, P. (2013). Intonation as an encoder of speaker's certainty: Information and confirmation yes-no questions in Catalan. Language and Speech, 56(2), 163–190.

Vanrell, M. M., Armstrong, M., & Prieto, P. (2017). Experimental evidence for the role of intonation in evidential marking. *Language and Speech*, 60 (2): 242–259.

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. Oxford: Oxford University Press.