Distributional training of speech sounds can be done with continuous distributions

Karin Wanrooij[a] **and Paul Boersma**
Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, The Netherlands
karin.wanrooij@uva.nl, paul.boersma@uva.nl

Abstract: In previous research on distributional training of non-native speech sounds, distributions were always *discontinuous*: typically, each of only eight different stimuli was repeated multiple times. The current study examines distributional training with *continuous* distributions, in which all presented tokens are acoustically different. Adult Spanish learners of Dutch were trained on either a discontinuous or a continuous bimodal distribution of the Dutch vowel contrast /ɑ/–/aː/. Both groups improved their perception of the contrast; this shows that continuous training works equally well as discontinuous training. Using the more natural continuous distributions is therefore recommended for future distributional learning experiments.

PACS numbers: 43.71.Ft, 43.71.Es, 43.71.An [SGS]
Date Received: November 15, 2012 **Date Accepted:** January 28, 2013

1. Introduction

Earlier research has shown that adult learners can improve their discrimination or classification of a non-native speech sound contrast simply by listening for a few minutes to a bimodal distribution representing this contrast (Maye and Gerken, 2000, 2001; Hayes-Harb, 2007; Gulian *et al.*, 2007; Escudero *et al.*, 2011; Wanrooij *et al.*, 2013). This phenomenon is called "distributional learning." The stimuli differ from one another in steps along an acoustic continuum. For a bimodal distribution, two stimuli with acoustic properties near the end points of the continuum (e.g., the two stimuli with F1 values of 11.9 and 14.0 ERB in Fig. 1; left) are presented more often than the other stimuli (as represented by the varying line lengths in the figure). Through the differences between the stimuli in their frequency of presentation, listeners supposedly start to treat these two most frequently presented stimuli (and their acoustic neighbors, which are presented slightly less often) as exemplars of two different speech sounds.

1.1 Discontinuous and continuous distributions

In all previous studies on distributional learning, bimodal distributions were based on stimuli with 8 or 10 different values for voice onset time (e.g., Maye and Gerken, 2000, 2001; Hayes-Harb, 2007; Maye *et al.*, 2002; Maye *et al.*, 2008; Yoshida *et al.*, 2010), vowel formants (e.g., Gulian *et al.*, 2007; Escudero *et al.*, 2011; Wanrooij *et al.*, 2013), or fricative frequencies and formant transitions (Cristià *et al.*, 2011), and these stimuli were repeated in certain proportions. In Fig. 1 (left), for instance, the eight stimuli (the thin vertical lines) are spaced at equal distances along the F1 continuum, and some stimuli are presented more often than others (the height of the vertical lines), while acoustic values in between those of the eight stimuli are never presented. We therefore label such distributions "discontinuous."

In a natural environment, however, acoustic values are never repeated exactly. Rather, naturally occurring speech tokens can have any value (between certain bounds)

[a] Author to whom correspondence should be addressed.

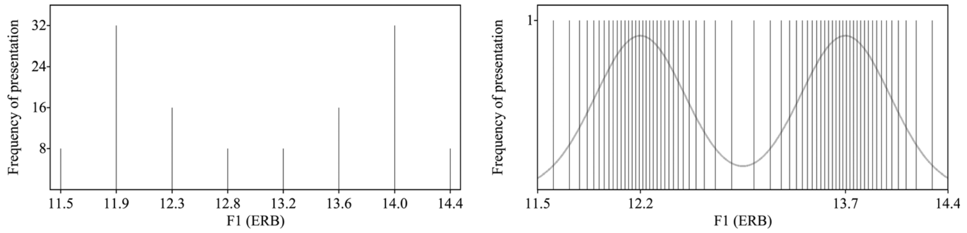

Fig. 1. At the left, a discontinuous stimulus distribution; at the right, a continuous stimulus distribution. Each vertical line represents a stimulus with a specific F1 value. The height of each vertical line shows how often the stimulus is presented to the listener. The gray curve in the right picture is the underlying probability density function (see Sec. 2.2).

along the relevant acoustic dimension. When applying this idea to a bimodal stimulus distribution for distributional training, we obtain Fig. 1 (right), where the stimuli (the thin vertical lines) are spaced more densely around 12.2 and 13.7 ERB and more sparsely elsewhere, and each stimulus is presented only once. We therefore label such distributions "continuous."

In the current study, we aimed to examine whether previous observations obtained with discontinuous distributions might have been artifacts of the unnatural sampling method. After all, it is known that input variability can influence category formation and discrimination (Lively *et al.*, 1993; Rogers and Davis, 2009), so that one could hypothesize that the observed changes in participants' behavior after training were due to the artificially sparse (eightfold) sampling of the acoustic space. To find out whether the effects reported in the distributional learning literature have not been methodological artifacts, one would have to test whether adult listeners also improve classification performance through listening to a more ecologically valid continuous distribution, with more variation in acoustic values, and without stimulus repetition. This is done in the present article, which compares three groups of participants: one group was presented with a discontinuous training (hence, the Discontinuous group), another group with a continuous training (hence, the Continuous group), and the third group was a control group that listened to classical music (the Music group). As explained in Sec. 2.1, the Discontinuous and Music groups were taken from Wanrooij *et al.* (2013).

1.2 A vowel contrast and its appropriate participant group

For the acoustic continuum we chose the Dutch vowel contrast /ɑ/–/aː/. For this contrast, appropriate listeners are native speakers of Spanish. This group is known to have difficulty classifying the two Dutch vowels when the durational difference (/aː/ is longer; Adank *et al.*, 2004) is eliminated, so that only the spectral difference (/aː/ has higher first and second formants; Pols *et al.*, 1973; Adank *et al.*, 2004) can be used to classify the vowels correctly (Escudero and Wanrooij, 2010; Escudero *et al.*, 2011; Wanrooij *et al.*, 2013). To train the Spanish listeners on this spectral difference only, the manipulated acoustic dimensions in both distributions (i.e., discontinuous and continuous) were the first and second formant values (F1 and F2), and the duration of the training vowels was kept constant (see also Escudero *et al.*, 2011; Wanrooij *et al.*, 2013). Note that Fig. 1 shows the discontinuous and continuous distributions of the F1 values only; because F2 values varied linearly with those of F1, the pictures for the discontinuous and continuous F2 distributions look identical.

2. Method

The method was identical to that of Escudero *et al.* (2011) and Wanrooij *et al.* (2013). Participants performed a pre-test (Sec. 2.3), a training phase (Sec. 2.2) and a post-test (Sec. 2.3).

2.1 Participants

Participants were adult native speakers of Spanish who were learning Dutch. Only the Continuous group was new, and it consisted of 50 participants. The Discontinuous group was taken from an earlier study as follows. To ensure a high-level benchmark for the Continuous group, we chose for our Discontinuous group the group that had shown the most improvement of all four groups that received discontinuous distributional training in two recent studies in our lab (Escudero *et al.*, 2011; Wanrooij *et al.*, 2013). These two studies used identical pre- and post-tests and an identical procedure as those used for the Continuous group in the present study, and in both studies the results were the same. Specifically, in both studies three groups of Spanish listeners participated, one presented with a discontinuous bimodal distribution representing the Dutch contrast /ɑ/–/aː/ (the Bimodal group), one exposed to a discontinuous *enhanced* bimodal distribution of the same contrast (the Enhanced group), and one, the control condition, presented with classical music (the Music group). In the enhanced bimodal distribution, the perceptual distance between the end point acoustic values of the training stimuli was larger than that in the non-enhanced bimodal distribution. Accordingly, the difference between the two speech sounds was "exaggerated" and thus presumably easier to perceive (Kuhl *et al.*, 1997; Liu *et al.*, 2003).

With new Spanish participants in the Bimodal and Enhanced groups (50 in each group), Wanrooij *et al.* (2013) replicated the results obtained for the participants (53 in each group) in Escudero *et al.* (2011), i.e., that (1) the Enhanced group improved significantly in accuracy of classification of Dutch /ɑ/ and /aː/, (2) the Music group did not show significant progress, and (3) the Enhanced group improved significantly more than the Music group. Table 1 shows the difference scores (i.e., post-test minus pre-test classification accuracy in percentages) for each group (i.e., Enhanced, Bimodal, Music) in both studies.

It can be observed that the Enhanced group in Wanrooij *et al.* (2013) had the highest absolute improvement after training of all four groups, with a 95% confidence interval (CI) that appeared narrower and further from zero than in Escudero *et al.* (2011). Therefore we used this group as a stringent standard against which to compare our Continuous group. In addition, we compared the results of the Discontinuous and Continuous groups to the Music group's results as obtained in Wanrooij *et al.* (2013). In Escudero *et al.* (2011) and Wanrooij *et al.* (2013), a music condition was preferred over a unimodal control condition for ethical reasons, because all participants were learners of Dutch and previous research had shown that a unimodal distribution may reduce discrimination performance (Maye *et al.*, 2002; Wanrooij *et al.*, 2012). Table 2 lists the mean age, age range, and length of residence in the Netherlands as a measure of previous exposure to Dutch, for the Discontinuous (12 male, 38 female), Continuous (15 male, 35 female), and Music (6 male, 44 female) groups separately.

2.2 Training: Stimuli and procedure

The stimuli in the continuous and discontinuous training distributions were made with the Klatt synthesizer in the computer program PRAAT (Boersma and Weenink, 2011). Each stimulus in both distributions had a fundamental frequency (F0) contour that fell from 150 to 100 Hz. Also the stimulus duration was 140 milliseconds (ms), and the inter-stimulus interval was 750 ms. Total training time was nearly 2 minutes.

Table 1. Difference score (= post-test minus pre-test percentage correct) for groups of Spanish listeners presented with enhanced, bimodal and musical training phases in two previous studies. Ninety-five percent confidence intervals are given between parentheses.

Previous study	Enhanced	Bimodal	Music
Escudero *et al.* (2011)	6.04 (+2.76… +9.31)	0.80 (−2.22… +3.83)	−0.15 (−3.50… +3.21)
Wanrooij *et al.* (2013)	6.63 (+4.05… +9.20)	3.83 (+0.97… +6.68)	2.00 (−0.50… +4.50)

Table 2. Participants' age (standard deviation between parentheses), age range, and length of residence (in years) in the Netherlands. *The Discontinuous and Music groups were taken from Wanrooij *et al.* (2013).

Group	Mean age	Age range	Length of residence
Music*	38.0 (9.0)	19.0–60.0	6.3 (6.8)
Discontinuous*	37.3 (8.0)	21.0–56.0	5.4 (5.0)
Continuous	33.2 (9.8)	21.6–63.2	3.1 (4.9)

The stimuli in the *discontinuous* distribution are described in detail in Escudero *et al.* (2011) and Wanrooij *et al.* (2013). The F1 values (range: 11.5–14.4 ERB) and F2 values (range: 15.3–18.2 ERB) varied in eight steps of approximately 0.4 ERB apart. Stimuli 1 through 4 with the lower F1 and F2 values can be thought of as representing the Dutch vowel /ɑ/, and stimuli 5 through 8 with the higher F1 and F2 values can be thought of as representing the Dutch vowel /aː/. Stimuli 1, 4, 5, and 8 in the tails (see Fig. 1, left) were each presented eight times, stimuli 2 and 7 at the peaks each occurred 32 times, and stimuli 3 and 6 were each presented 16 times. Thus the total number of presentations was 128.

To make a *continuous* distribution that would correspond as closely as possible to the discontinuous one, we first had to match the shapes of the distributions. For this, we approximated the ratio of the least to most frequent stimuli; i.e., this ratio is 1 to 4 in earlier studies with discontinuous distributions and is approximately 1 to 4 in the current continuous distribution (see Fig. 1). Further, we created the underlying continuous distribution as the sum of two Gaussian curves the means of which were positioned at 25% and 75% of the F1 range (and consequently also of the F2 range), and the standard deviations of which were set to 11% of the total F1 (or F2) range. This distribution is the probability density function shown in Fig. 1 (right).

The next step was the determination of the F1 and F2 values for each stimulus. We created the same total number of stimuli (128) as for the discontinuous distribution. This time none of the stimuli was repeated, so that each stimulus had a unique combination of F1 and F2 values. As the procedure for the calculation of the F2 values is the same as that for the F1 values, we restrict the description to the F1 values, as follows.

After determining the precise shape of the underlying continuous distribution (the gray curve in Fig. 1, right), the F1 values of the 128 stimuli (the thin vertical lines in Fig. 1, right; for the purpose of clarity only 64 stimuli are shown) were calculated in the following way. First, the area under the curve was normalized, i.e., it was set to 128, the number of stimuli. Then the distribution was sampled evenly, i.e., the F1 values were chosen in such a way that the area between consecutive F1 values under the curve was always 1. Thus there were 127 unit areas between the 128 F1 samples. The additional leftmost area (running from the left edge of the F1 continuum to the first F1 sample) and rightmost area (running from the last F1 sample to the right edge of the F1 continuum) were 0.5 each.

The task of the participants in the training phase was merely to listen. Participants in the Discontinuous and Continuous groups were instructed to listen to the vowels carefully because they would perform a post-test afterward. Participants in the Music group were asked to relax while listening to the music and were informed that they would perform a post-test afterward.

2.3 Pre- and post-tests: Stimuli and procedure

The pre- and post-tests, which were equal to those used in Escudero *et al.* (2011) and Wanrooij *et al.* (2013), were identical classification tasks, which were the same for all participants. Listeners heard an X-stimulus and two subsequent response options

Table 3. Pre- and post-test percentages correct, and difference (= post- minus pre-test percentage correct) per group. Standard deviations between participants in each group are given between parentheses. *Discontinuous and Music groups from Wanrooij *et al.* (2013).

	Pre	Post	Difference
Music*	61.73 (11.12)	63.73 (13.31)	2.00 (8.81)
Discontinuous*	60.43 (11.71)	67.05 (13.48)	6.63 (9.06)
Continuous	62.40 (10.74)	72.08 (13.12)	9.68 (10.13)

A and B. They were forced to choose which option was from the same vowel category as X.

The X-stimuli were chosen to be natural vowels to promote classification rather than discrimination; they were a subset of the vowels reported in the corpus by Adank *et al.* (2004), which were produced by male and female speakers of standard Northern Dutch. The response options A (F1 = 12.5 ERB, F2 = 16.1 ERB) and B (F1 = 13.3 ERB, F2 = 17.4 ERB) were chosen to be synthetic; they were created with the computer program PRAAT (Boersma and Weenink, 2011) and had an equal duration of 140 ms to prevent participants from resorting to durational differences between /ɑ/ and /aː/ (recall Sec. 1.2).

In each test, participants were asked to classify 80 X-stimuli. Listeners were told that the next trial would only appear after their response, but they were encouraged to answer as quickly as possible and to guess if they were unsure. To test hearing and understanding of the test, the participants performed a practice test before the pre-test and before the post-test.

3. Results

Table 3 gives the pre- and post-test percentages correct (i.e., the percentage of correct classifications of the 80 test stimuli) and the difference (i.e., the post- minus pre-test percentage correct) for the Music, Discontinuous, and Continuous groups. An ANOVA on pre-test accuracy did not display a significant difference between the three groups [$F(2,147) = 0.40$, $p = 0.67$]. This supports the equality of the groups before training.

The difference between pre- and post-test accuracy is a measure of improvement after training. For the Continuous group, this difference was 9.68% (95% CI = +6.80%… +12.55%), which was significantly different from zero [one-sample $t(49) = 6.75$, $p < 0.001$]. As reported in Wanrooij *et al.* (2013), the difference score also differed from zero significantly for the Discontinuous group [one-sample $t(49) = 5.17$, $p < 0.001$], and it did not for the Music group [one-sample $t(49) = 1.61$, $p = 0.12$] (95% CIs: see Table 1). This confirmed that both the Discontinuous and the Continuous groups improved their accuracy percentages robustly after training. An ANOVA with difference scores as the dependent variable revealed a significant difference between groups [$F(2,147) = 8.54$, $p < 0.001$]. *Post hoc* *t*-tests on the difference scores using Tukey's HSD for multiple-comparison corrections showed a significant difference between the Music and Discontinuous groups of +4.63% (CI = +0.20%… +9.05%, $p = 0.04$) and between the Music and Continuous groups of +7.68% (CI = +3.25%… +12.10%, $p < 0.001$), and no significant difference between the Discontinuous and Continuous groups (difference = +3.05%, CI = −1.38%…+7.48%, $p = 0.24$). Thus participants who received distributional training improved more than participants who listened to music instead, although we cannot say with confidence that the progress of the Continuous group (9.68%) was larger than that of the Discontinuous group (6.63%).

4. Conclusion

We showed that listeners' performance in classifying a non-native phoneme contrast can be improved not only by training them with a discontinuous distribution but also by training them with a continuous distribution. We can therefore erase the fear that earlier results demonstrating an effect of training with discontinuous distributions (e.g., Maye and Gerken, 2000, 2001; Maye *et al.*, 2002, 2008; Hayes-Harb, 2007; Gulian *et al.*, 2007; Yoshida *et al.*, 2010; Escudero *et al.*, 2011; Cristià *et al.*, 2011; Wanrooij *et al.*, 2013) could have been artifacts of the discontinuous sampling method; after all, these results have now been replicated with the arguably more natural continuous distributions, so it has become more likely that the observed perceptual improvements are a realistic result of bimodal training. However, as both types of sampling have now been shown to exhibit distributional learning effects and continuous distributions can be considered more ecologically valid than discontinuous distributions, we recommend for future distributional learning experiments not to artificially reduce the variation in the stimuli to 8 or 10 auditory values but to solely employ continuous distributions.

Acknowledgments

This research was supported by Grant No. 277.70.008 from the Netherlands Organization for Scientific Research (NWO) awarded to the second author. Participant recruitment and testing for the Music and Discontinuous groups were also supported by NWO Grant No. 275.75.005 awarded to Paola Escudero. Further, we would like to thank Paola Escudero for assisting with participant recruitment, and Marieke van den Heuvel and Gisela Govaart for testing the participants in the Continuous group.

References and links

Adank, P., Van Hout, R., and Smits, R. (**2004**). “An acoustic description of the vowels of Northern and Southern standard Dutch,” J. Acoust. Soc. Am. **116**, 1729–1738.

Boersma, P., and Weenink, D. (**2011**). “Praat: Doing phonetics by computer [computer program],” http://www.praat.org (Last viewed 3/11/2013).

Cristià, A., McGuire, G. L., Seidl, A., and Francis, A. L. (**2011**). “Effects of the distribution of acoustic cues on infants’ perception of sibilants,” J. Phonetics **39**, 388–402.

Escudero, P., Benders, T., and Wanrooij, K. (**2011**). “Enhanced bimodal distributions facilitate the learning of second language vowels,” J. Acoust. Soc. Am. **130**, EL206–212.

Escudero, P., and Wanrooij, K. (**2010**). “The effect of L1 orthography on non-native and L2 vowel perception,” Lang. Speech **53**, 343–365.

Gulian, M., Escudero, P., and Boersma, P. (**2007**). “Supervision hampers distributional learning of vowel contrasts,” in *Proceedings of the 15th International Congress of Phonetic Sciences*, Saarbrücken, pp. 1893–1896.

Hayes-Harb, R. (**2007**). “Lexical and statistical evidence in the acquisition of second language phonemes,” Second Lang. Res. **23**, 1–31.

Kuhl, P., Andruski, J., Chistovich, I., Chistovich, L., Kozhevnikova, E., Ryskina, V., Stolyarova, E., Sundberg, U., and Lacerda, F. (**1997**). “Cross-language analysis of phonetic units in language addressed to infants,” Science **227**(5326), 684–686.

Liu, H.-M., Kuhl, P., and Tsao, F.-M. (**2003**). “An association between mothers’ speech clarity and infants’ speech discrimination skills,” Dev. Sci. **6**(3), 1–10.

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). “Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories,” J. Acoust. Soc. Am. **94**, 1242–1255.

Maye, J., and Gerken, L. (**2000**). “Learning phonemes without minimal pairs,” in *Proceedings of the 24th Annual Boston University Conference on Language Development*, edited by S. C. Howell, S. A. Fish, and T. Keith-Lucas (Cascadilla Press, Somerville, MA), pp. 522–533.

Maye, J., and Gerken, L. (**2001**). “Learning phonemes: How far can the input take us?” in *Proceedings of the 25th Annual Boston University Conference on Language Development*, edited by A. H.-J. Do, L. Domínguez, and A. Johansen (Cascadilla Press, Somerville, MA), pp. 480–490.

Maye, J., Weiss, D., and Aslin, R. (**2008**). “Statistical phonetic learning in infants: Facilitation and feature generalization,” Dev. Sci. **11**(1), 122–134.

Maye, J. C., Werker, J. F., and Gerken, L. A. (**2002**). “Infant sensitvity to distributional information can affect phonetic discrimination,” Cognition **82**, B101–B111.

Pols, L. C. W., Tromp, H. R. C., and Plomp, R. (**1973**). “Frequency analysis of Dutch vowels from 50 male speakers,” J. Acoust. Soc. Am. **53**, 1093–1101.

Rogers, J. C., and Davis, M. H. (**2009**). "Categorical perception of speech without stimulus repetition," in *Proceedings of Interspeech 2009*, Brighton, pp. 376–379.

Wanrooij, K., Escudero, P., and Raijmakers, M. (**2013**). "What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning." J. Phonetics. In press.

Wanrooij, K., Van Zuijen, T., and Boersma, P. (**2012**). "MMN declines after distributional vowel training," poster presentation at *The 6th Conference on Mismatch Negativity (MMN) and its Clinical and Scientific Application*, May 1–4, New York, http://home.medewerker.uva.nl/k.e.wanrooij/bestanden/MMN2012.pdf (Last viewed 3/11/2013).

Yoshida, K. A., Pons, F., Maye, J., and Werker, J. F. (**2010**). "Distributional phonetic learning at 10 months of age," Infancy **15**(4), 420–433.