Adaptive phonetic training for second language learners

Athanassios Protopapas[1] *and Barbara Calhoun*[2]

[1]Institute for Language and Speech Processing, Athens, Greece
[2]Scientific Learning Corporation, Berkeley, CA
protopap@ilsp.gr, bcalhoun@scilearn.com

Abstract

Some second language contrasts are very difficult to perceive, such as the English r/l for Japanese speakers. In phonetic training of such contrasts, variability and fading have been found to boost efficiency and generalization. Here we report on eight participants trained with an LPC-based method of processing speech stimuli that increases discriminability by exaggerating their natural acoustic differences. Mean word identification error rate dropped from 34% to 18% after 2–4 weeks of adaptive training on CV syllables. Variability in speaker voices and phonetic context, initially blocked but gradually mixed, ensured generalization over these dimensions.

1. Introduction

The problem of perceiving nonnative phonetic contrasts is well known in second language (L2) learning research. Several theories have been proposed regarding the cause of selective difficulty observed with particular contrasts, depending on one's first language (L1) background [1]. On a practical level, various approaches have been tried for training L2 learners to perceive correctly (i.e., categorically, as evidenced by identification and discrimination tests) contrasts not found in their L1 that are particularly difficult.

Several lessons have been learned from many studies over the past three decades concerning such training, notably pertaining to efficiency and generalization [2]. It is now common to include a diverse set of training stimuli, with respect to both the voices (speakers) and the phonetic context in which the trained phonemes occur. This is necessary for the induction of general (i.e., context-independent) phonetic categories and has produced positive outcomes in word identification [3,4]. In addition, identification (categorization) training is preferred over discrimination training because the latter is thought to direct attention to fine acoustic details distinguishing particular training tokens and not to the formation of phonetic categories [2,3].

Application of a long-standing psychological principle of learning [5], now corroborated by neuroscience on the formation and maintenance of categories in neural networks [6], has shown promise for L2 education. Specifically, making the acoustic dimension of interest more salient, to the point where the stimuli are discriminable by the L2 speakers, offers an entry point towards mastering a difficult distinction by adapting to the speaker's ability and gradually pushing their limits ("fading") to the point of the natural phonetic stimuli. This has been applied, with some success, in a few cases [7,8]. However, such attempts at enhancement typically take the form of amplification, truncation, or elongation of the acoustic signal, regardless of whether the critical acoustic difference is itself temporal or in amplitude.

Here we present an approach similar but with the radical variation of addressing directly the acoustic dimension of difference. At the same time, we employ the proven principles of variability and adaptive modification to increase efficiency and generalization. Furthermore, the gradual buildup from easier to more difficult conditions is not restricted to the acoustic dimension but is also extended to stimulus variability. L2 contrast learning may be more difficult, and thus likely more tiring and less motivating, when one is presented initially with a great variability of stimuli. However, variability is necessary for generalization. An adaptive schedule of variability would first train a single instance of the contrast and then gradually add voices and phonetic contexts [9]. This is the procedure in our study.

In summary, a method of training L2 learners to perceive a difficult phonetic contrast is presented, building on previous research on variability and adaptation, and adding the element of selective modification of a particular acoustic dimension of interest. The well-researched case of Japanese speakers learning the r/l contrast of American English is chosen because abundant reference data exist in the literature [4,10] as a gauge of efficiency and effectiveness. In the following sections, we describe our novel processing method for creating stimuli and we show that the proposed method indeed increases discriminability without rendering the stimuli too unnatural. Finally we present methods and outcomes for 8 Japanese speakers who participated in a preliminary training study.

2. Processing Method

Creating *over*-discriminable speech contrasts can be conceived as an extension of constructing phonetically *ambiguous* stimuli based on natural speech sounds. Creating acoustic continua between two speech sounds has a long history in psycholinguistic research, where phonetically ambiguous stimuli are needed to neutralize or delay phonetic perceptual decisions.

Methods to create "intermediate" stimuli have included period-by-period substitution, waveform averaging etc. Digital signal processing algorithms have been used for phonetic continua not amenable to these methods. For

example, algorithms based on linear predictive coding (LPC) can be used to create ambiguous speech sounds by affecting stop bursts and formant transitions in a manner previously only possible with synthesized speech. LPC-resynthesis techniques can be applied to a range of stimulus contrasts to create "intermediate" stimuli without making explicit assumptions about the nature of their acoustic differences.

An LPC model of the speech signal can be formulated to be equivalent to a lossless tube "vocal tract model" (following [11], pp. 82ff). In such a model comprising p tubes of equal length $1/p$ and fixed cross-sectional areas A_i, traveling waves in each tube are subject to pressure and volume velocity continuity constraints at the boundaries between adjacent tubes, where mismatched impedance due to cross-sectional area differences results in wave reflection. The amount of backward-traveling wave reflected at each junction, r_i, is called the reflection coefficient. For the ith junction, this coefficient is related to the cross-sectional areas A_i of the two adjacent tubes according to the formula

$$r_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i}, \quad 1 \le i \le p. \tag{1}$$

Setting the radiation load at the "lips" (e.g., $A_{p+1} = \infty$ in the completely lossless case), the transfer function of this system assumes the form

$$V(z) = \frac{G}{1 - \sum_{i=1}^{p} \alpha_i z^{-i}}, \tag{2}$$

which is the steady-state system function of a slowly time-varying digital filter obtained by linear prediction analysis of order p. Moreover, the partial correlation coefficients k_i, derived in the course of solving the LPC equations to compute the predictor coefficients α_i, are related to the reflection coefficients of the lossless tube model simply as $r_i = -k_i$. In practice, the shape of the model is defined by the "log area ratio coefficients," g_i, which are derived by the formula ([12], p. 444)

$$g_i = \log\left(\frac{A_{i+1}}{A_i}\right) = \log\left(\frac{1-k_i}{1+k_i}\right), \quad 1 \le i \le p. \tag{3}$$

These parameters do not necessarily correspond to the vocal tract that produced the analyzed sound waveform, but they describe an acoustically equivalent "vocal tract" that can be used to approximately reconstruct the speech signal, to the extent that the all-pole LPC model approximates it. Small deviations in these parameters result in acoustic signals corresponding to slightly different vocal tracts. Thus, the spectral characteristics of the reconstructed signal are close to those of the original signal and under the same constraints with respect to the number of formants and their relative positions. Therefore, sets of parameters between those derived from two speech waveforms will result in signals acoustically between the original two and under the same vocal tract constraints.

Consider, for example, the syllables [rak] ("rock") and [lak] ("lock"), spoken by a male speaker. From the recording, the log area ratio coefficients are derived using Equation 3, and then "intermediate" coefficients are created by linear interpolation between the resulting vectors at each time point. That is, one first computes the differences δ_i between corresponding coefficients as $\delta_i = g_i^{[\text{rak}]} - g_i^{[\text{lak}]}$, $1 \le i \le p$, thus creating a p-dimensional vector on the straight line that joins the points in p-space defined by the coefficients for [rak] and [lak]. Each point on this vector defines the log area ratio coefficient set for a vocal tract model in between those corresponding to the original [rak] and [lak]. Specifically, for $\lambda \in [0,1]$ one can define

$$g_i^{\lambda} = g_i^{[\text{lak}]} + \lambda \delta_i, \; 1 \le i \le p, \tag{4}$$

and the resulting coefficients can then be converted to partial correlation coefficients using the formula

$$k_i = \frac{1 - e^{g_i}}{1 + e^{g_i}}, \quad 1 \le i \le p, \tag{5}$$

to be used for LPC resynthesis of a signal with "[rak]–[lak] proportions" of $\lambda : 1-\lambda$.

However, nothing restricts application of this method to $0 \le \lambda \le 1$. Values of λ outside the range [0,1] result in pairs of stimuli that are acoustically more different from each other than were the natural stimuli from which the original coefficients were derived. Most importantly, the exaggerated spectral difference between the resulting signals will be exactly along the dimension on which the natural stimuli differed in the first place. That is, an enhancement of the natural acoustic distinction will be obtained by distorting the natural syllables away from their original acoustic properties.

Figure 1 illustrates the point with a set of resynthesized stimuli based on a recording of the words "rock" and "lock" ([rak] and [lak]). Results are shown for λ between −0.75 and 1.75. Notice the intermediate positions of the third formant onset and transitions between λ=0.0 (corresponding to the original [l]) and λ=1.0 (corresponding to [r]), and the more "extreme" formant tracks for λ outside this interval. Evidently, for values of λ less than 0.0, the third formant increases in frequency and amplitude away from [r], i.e., in the direction in which [l] differs from [r]. Similarly, for values of λ greater than 1.0, the third formant approaches the second one in frequency and is increased in amplitude, thus becoming less [l]-like without affecting what is common between [l] and [r].

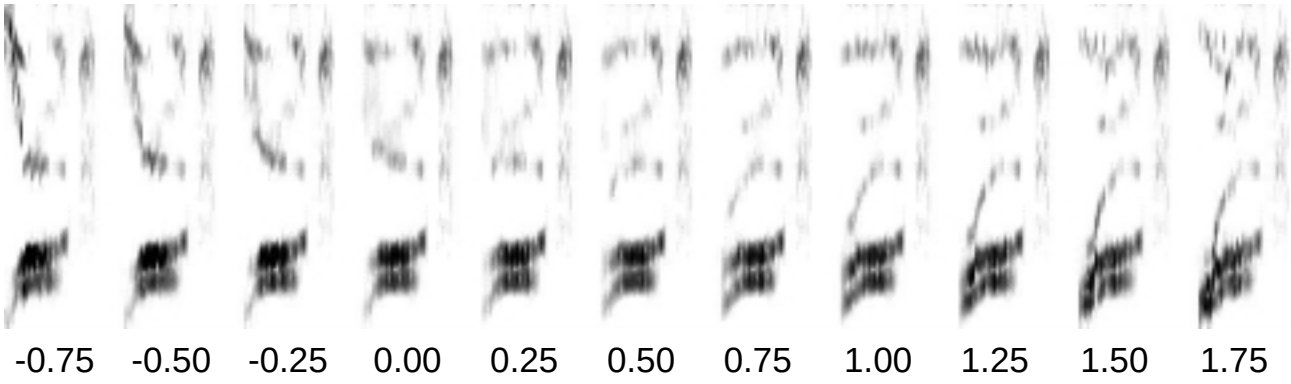

Figure 1. Spectrograms of the resynthesized syllables along a continuum on the line defined by the vector of log area ratio coefficients from [lak] to [rak]. The indicated values of λ were used with Equation 4 to interpolate and extrapolate from the two sets of LPC-derived log area reflection coefficients. The displayed frequency range is 0–5.5 kHz. Each stimulus is 265 ms long.

Informal listening of these stimuli indicates that they sound natural for $0 \le \lambda \le 1$ within a reasonable range of processing parameters (LPC order, processing window length, frame rate, sampling rate). The resynthesized stimuli become progressively less natural-sounding as λ moves away from the [0,1] interval, necessitating some additional fine-tuning, such as setting the LPC order after some trial and error, imposing an amplitude envelope on the resynthesized signal to avoid extreme fluctuations, and smoothing the reflection coefficients. Splicing only the distinctive portion of the ambiguous resynthesized signal onto the natural remaining utterance improves the naturalness of the entire stimulus. To be successful, such splicing must be done at an appropriate point, preserving the fundamental period across the juncture. However, it must be noted that once the base stimuli are carefully selected, the processing procedure can be automated. Furthermore, it is not necessary to perform such processing over a large set of recordings because a few exemplars for each voice and phonetic contrast will suffice to induce proper generalization. Thus the need for manual adjustment does not restrict application of the method to unrealistic laboratory settings or to cases too specific to be of general use.

3. Perceptual Evaluation

Perceptual testing of the resynthesized stimuli with native speakers is necessary to ensure that they can be accurately identified as the intended phonemes and perceived in a categorical manner along the continuum. An identification test was conducted by presenting participants with a single stimulus in each trial (i.e., a syllable from a [ra]–[la] continuum), asking them to classify it as one of two categories ("ra" or "la" in this example). The percentage of one response category is plotted against position in the acoustic continuum (λ). For native speakers, the resulting curve is typically flat around both endpoints with an abrupt transition at some point in between. This abrupt perceptual transition is often considered the hallmark of phonetic perception.

In order to ascertain that the abrupt perceptual transition is not an artifact of having only two response categories, the participants' ability to *discriminate* between pairs of stimuli drawn from the same continuum is also tested. Two stimuli are presented in each trial for a judgment of "same" or "different." Here, each pair of stimuli were synthesized with λ values differing by 0.3. Therefore, a data point plotted at 0.55 shows discrimination of a stimulus synthesized with λ =0.4 from one with λ =0.7. Typically, for native speakers, two stimuli given the same phonetic label in the identification test are difficult to discriminate, whereas stimuli from opposite sides of the transition boundary are very easy.

In the context of the present method an additional test is of interest. Specifically, discriminability between stimuli should increase with increased difference in λ value. That is, discrimination should be much better for stimuli 2.0 λ-units apart (one with $\lambda = -0.5$ and the other with $\lambda = 1.5$) than for stimuli 0.5 λ-units apart (one with $\lambda = 0.25$ vs. one with $\lambda = 0.75$).

Figure 2 shows the identification and discrimination performance of 3 adult native English speakers on the stimuli from two [ra]–[la] continua, one with a male and one with a female voice, for λ between -0.7 and 1.7 in steps of 0.1. The abrupt perceptual transition between [r] and [l] labeling and the peak in discrimination at the same point indicate that these stimuli are perceived appropriately. Note also that the exaggerated stimuli are consistently labeled as exemplars of their respective category (left column, points outside the [0,1] range), and that stimulus pairs separated by at least the natural [r]–[l] distance of 1.0 (right column) are perfectly discriminable. The increased discrimination for some pairs 0.3 λ-units apart outside the [0,1] range (middle column) is in part due to artifacts introduced by the extrapolation and in part because exaggeration may cause phonetic distortion (here especially on the [r] side). This is expected because the processing is meant to push phonetic exemplars away from their natural position and thus possibly to the fringes or outside their respective

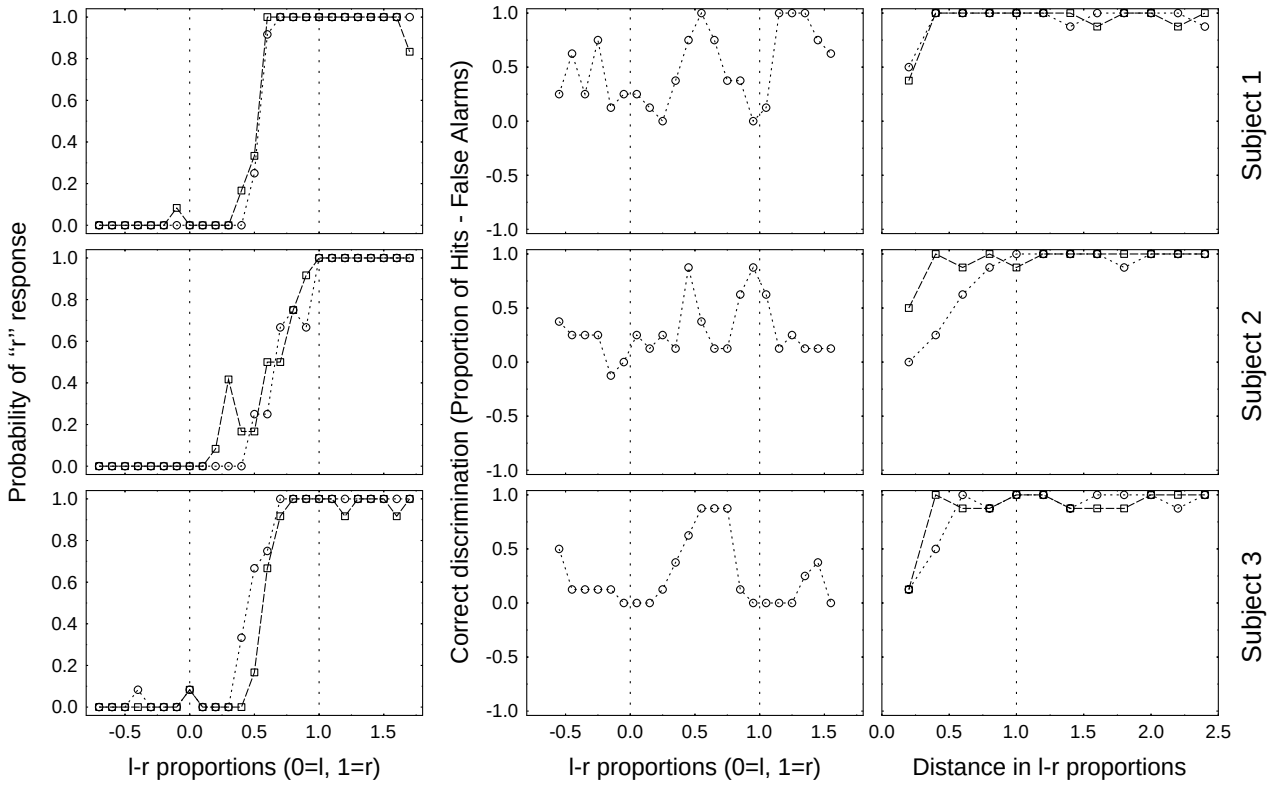

Figure 2. Categorical perception curves for 3 native American English speakers with the resynthesized stimuli along a [ra]– [la] continuum for two stimulus voices (male: squares on dashes; female: circles on dots). Each row shows data from a single participant. Left: Identification of stimuli for λ between -0.7 and 1.7 in 0.1 steps. Middle: Discrimination performance (hits−false alarms over total number of trials) of stimulus pairs 0.3 λ -units apart. Right: Discrimination of stimuli symmetric with respect to λ =0.5 for increasing values of λ -distance.

phonetic category. What is important is that the acoustic differences between these stimuli are of the same kind as between natural tokens.

4. Training Study

According to our hypothesis, given sufficient exaggeration, listeners unable to discriminate the natural stimuli will be able to make accurate disctinctions of the resynthesized stimuli, and this can help them gradually learn the acoustic distinction between the natural tokens. In this section we describe a training study with eight Japanese listeners, including pre- and post-testing for discrimination and identification in words and syllables.

4.1 Method

4.1.1 Participants

Eight native Japanese females, most of them students, completed the training study and post-testing. They were recruited through advertisements on the UC Berkeley campus and paid an hourly compensation. They had typically started learning English in Japan, focusing on grammar and reading. They all used mostly English at work but typically spoke Japanese (or both Japanese and English) with friends, family, and at home.

Table 1. The current age, time in the United States at recruitment, and age of first English instruction, for the eight participants who completed training and pre- and post-testing.

Participant	Age	Stay in US	Age English
S1	29	15 mo	13
S2	28	6 mo	12
S3	24	1 yr	N/A
S4	26	4 mo	12
S5	24	1 mo	12
S6	25	9 mo	12
S7	28	1.5 mo	13
S8	29	2 yr	12

4.1.2 Stimuli

The training stimuli were resynthesized syllables along [rV]–[lV] continua, for the three vowels [a], [i], and [u]. Two natural pairs from each of two speakers, one male and one female, digitized at 11.025 kHz / 16 bits, with 22.05-ms windows of signal overlapping 50% were subjected to 20th-order LPC processing resulting in a set of stimuli for each pair with λ ranging between –1.75 and +1.75. The same stimuli were used for syllable

identification and discrimination testing before and after training (but only one voice was used in discrimination).

The stimuli for the word identification tests included 128 minimal [r]–[l] pairs with initial singleton [r] or [l] and a variety of following vowels. They were recorded by the same two speakers who recorded the training stimuli and by two additional speakers, one male and one female, to test for generalization over voices.

4.1.3 Procedure

Participants were tested before and after training on the resynthesized [ra]–[la] stimuli, including identification (6 repetitions of the 50 stimuli) and discrimination (2 repetitions of each of 88 combinations of two stimuli, including same and different), as described above for the American listeners. They were also tested on word identification (each of the 256 words presented once). An additional word identification test was used with the last few participants, including words from minimal r/l pairs with the [r] or [l] in a word-final or medial position or in a word-initial cluster. The entire testing battery typically took two to three hours and was administered in sessions over two or three testing days. No feedback was provided during the testing sessions.

Training was initiated on a day following completion of pre-testing and was continued until each participant reached a high level of competence (explained below). In each training trial, a stimulus was presented and the participant had to press a key or mouse button to indicate perception of the initial consonant as [r] or [l]. Immediate feedback was provided after each trial. An unrestricted number of optional practice trials, added at the initial participants' request, were available between blocks of training trials. In practice trials, participants pressed the [r] or [l] button and the corresponding stimulus was played at the current level of modification.

Scheduling of stimuli followed the principle of gradual variability, through stages of phonetic context (vowels) and voice (speaker) session blocking according to the participant's ability. Stages progressed from an individual instance (recording) with the (clearer) female voice, vowels segregated by blocks, to two instances with the same voice, to create intra-speaker variability aimed at removing attention from accidental acoustic characteristics or processing artifacts. The next stages followed the same process with the male voice, and then a mixture of both instances from both voices. In the final stage, blocking by vowel was also removed and stimuli from all voice × vowel conditions were presented mixed.

Over the course of training, acoustic processing decreased from an exaggerated level to normal speech and finally to a slightly ambiguous level (λ within (0,1)). Within each training session, stimuli were presented at the lowest level of "exaggeration" the participant could handle, using an adaptive staircase procedure (modified 12-up 2-down) to track perceptual progress. Upon reaching a processing level less than zero (i.e., with acoustic distance less than natural) at the end of a session, the participant progressed to the next stage.

Participants trained in a quiet booth at the Scientific Learning Corp. offices for at most an hour each day, typically covering 3 or 4 sessions of 250 trials each. Stimuli were presented over Sony MDR-600 or Sennheiser HD-60 headphones at about 75 dB SPL.

4.2 Results

Each participant trained as long as necessary at each stage before proceeding to the next one. The total number of training days for each is shown in Table 2. Figure 4 (on the next page) shows a typical training curve illustrating the progression through stages of voice and phonetic context variability. Note that successive stages are mastered with decreasing difficulty with the exception of the final stage, with all phonetic contexts mixed, which required a larger number of trials.

4.2.1 Stimulus discriminability

An important issue in the application of our method is the pre-training discrimination performance on pairs of stimuli taken symmetrically around the acoustic [ra]–[la] midpoint. This is shown in Figure 3 averaged over all eight participants. It indicates whether the acoustically "exaggerated" stimuli are indeed more discriminable to the Japanese listeners and thus likely to have facilitated learning of the phonetic distinction. Note that the discrimination between resynthesized tokens less than 1.0 λ-units apart (the natural distance) is very poor. However, discrimination of stimuli spaced further apart is increasingly improved, peaking after 1.5 (i.e., for the pair of stimuli with λ values of -0.25 and 1.25).

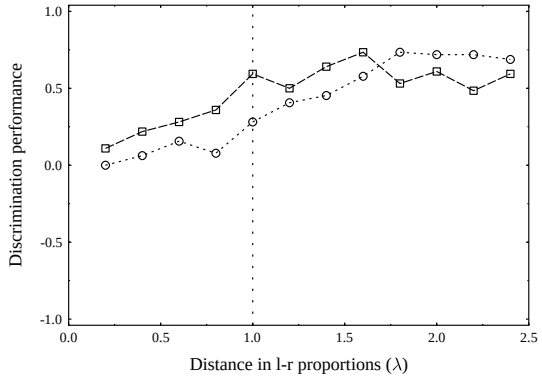

Figure 3. Discrimination of stimuli symmetrically around λ =0.5 as a function of λ-distance, averaged over the eight Japanese training participants. The y-axis corresponds to the proportion of hits minus false alarms.

Thus it appears that listeners who have not learned to utilize a particular acoustic cue (or set of cues) in making a phonetic distinction can in fact perform better on the basis of this acoustic cue (or set of cues) if it is sufficiently exaggerated to become salient.

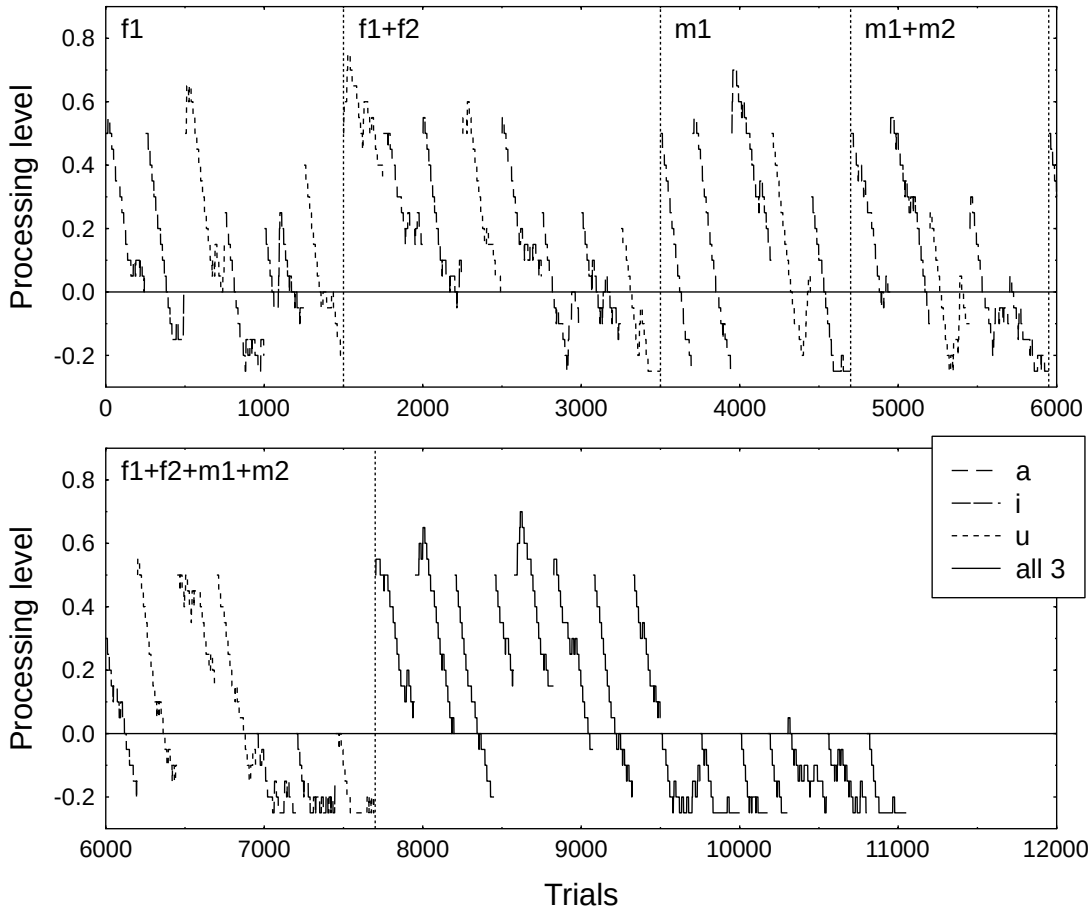

Figure 4. Training progression for participant S6. The discontinuous lines over the first 7700 trials correspond to the three phonetic contexts (vowels) which were blocked until mastered with each voice separately and then with both voices combined. The continuous lines at the final stage correspond to the mixed-content condition. f: female voice; m: male voice. The numbers 1 and 2 refer to the token pair used in training (only one first, then both in mixed presentation for acoustic generalization). "Processing level" is distance from natural, in λ -units, negative values indicating interpolation (ambiguity) rather than exaggeration.

4.2.2 Phonetic categorization

Figure 5 shows the performance of the eight Japanese participants before and after training on identification (left) and discrimination (right). In the pre-training test, U-shaped and flat curves were obtained, in contrast to the "categorical" identification curves obtained from the native English speakers (Figure 2). The pre-training discrimination curves also indicate very poor performance, the proportion of hits minus false alarms never exceeding 0.5 in the natural and ambiguous range, in striking contrast to the natives. Note that most of these Japanese listeners also appear unable to use the artifacts and distortions present in the extreme stimuli in making their judgments, as shown by their low performance in the "exaggerated [r]" range. This is further evidence of their lack of appropriate phonetic categories relative to which some stimuli may be judged to be worse exemplars by native speakers.

The situation is substantially improved in the post-training data, particularly for the identification task. Most labeling curves attain or approach a normal (i.e., native) appearance of a categorical curve with two distinct categories and a more or less abrupt transition between them. The discrimination data are still distinct from native performance, and this may have to do with the training task. These results show great improvement but imperfect ultimate formation of phonetic categories and warrant further research of training tasks and methods to induce more native-like representations.

4.2.3 Word identification

However, the acid test of the method is not in the categorical perception of the syllables used for training, but in the correct identification of naturally spoken words, since this is the skill required for communication. Table 2 shows the pre- and post-training performance of the eight participants in the identification of words from minimal [r]–[l] pairs, recorded from the same speakers used for training as well as from different speakers. Note that the range of vowel contexts in this test greatly exceeds the three contexts used in training. Still, word error rate typically dropped by half for these individuals, regardless of initial performance or amount of training.

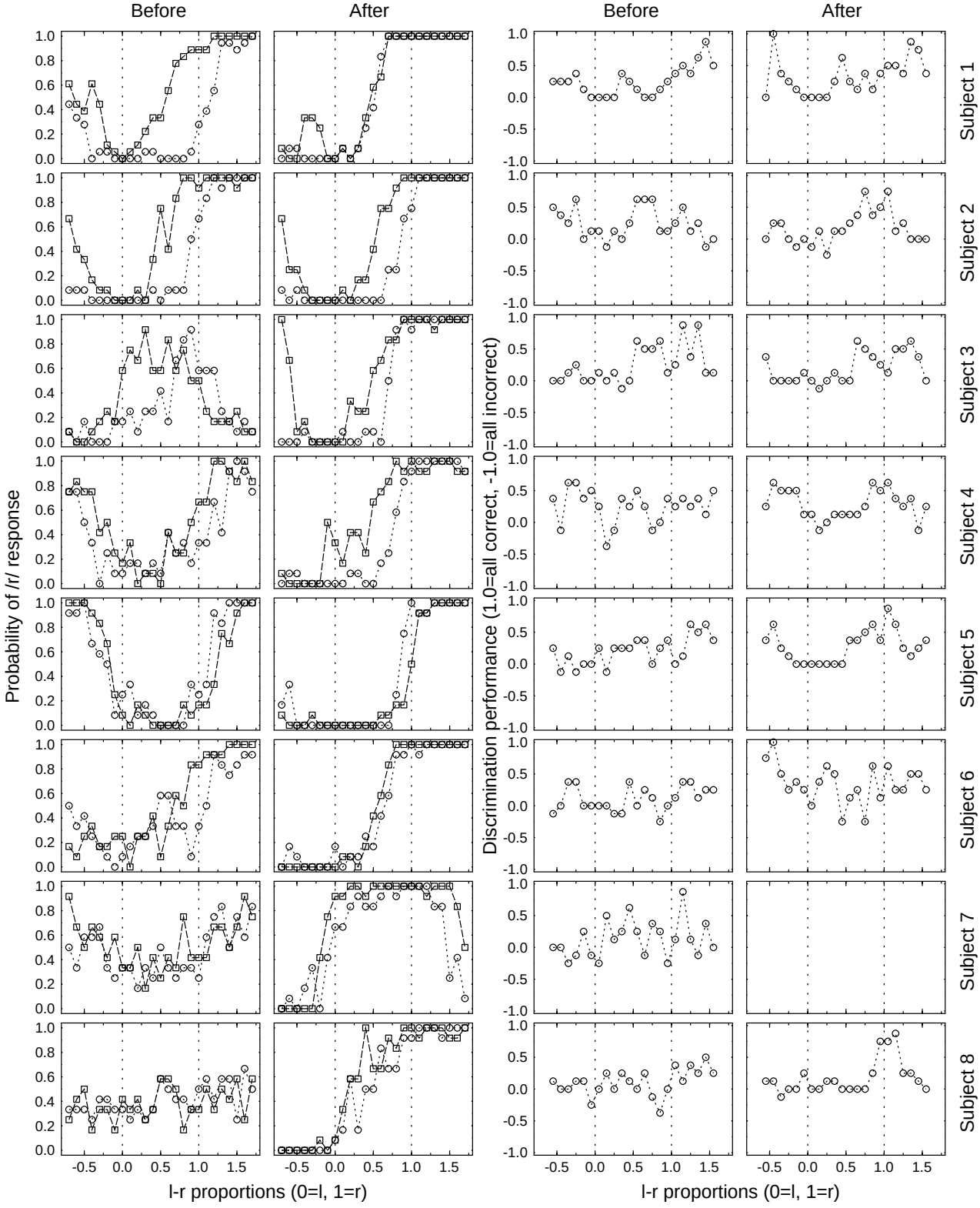

Figure 5. Test of categorical perception of the resynthesized syllables by the eight Japanese participants, including identification (left) and discrimination(right), before and after training. Each row shows data from a single participant for stimuli in a male voice (squares on dashed line) and a female voice (circles on dotted line). Discrimination performance is a proportion of hits minus false alarms for pairs of stimuli 0.3 λ -units apart.

This improvement (which, incidentally, appears strongly related to length of stay in the US) was substantially equivalent for the untrained voices, indicating perfect generalization in the inter-speaker dimension. Analysis of the performance by phonetic context (trained vs. not trained vowel; data not shown) similarly indicates perfect generalization over this dimension as well.

Table 2. Number of training days, and word error rates before (Pre) and after (Pst) training, for the voices used in training (Er) and for two other voices (UEr), for each training participant.

Particip.	Days	ErPre	ErPst	UErPr	UErPst
S1	12	18.7	7.8	–	3.2
S2	22	37.5	21.9	–	18.8
S3	33	32.4	14.8	17.6	8.4
S4	19	32.4	27.0	30.0	17.6
S5	17	41.0	24.6	38.0	20.0
S6	12	41.8	21.9	37.2	16.8
S7	13	36.7	20.7	27.2	17.2
S8	8	33.2	5.9	12.0	1.2

On the test for word identification with [r] and [l] positions other than word-initial singletons, preliminary data from a few participants (not shown) indicate that training did not transfer along this dimension (cf. [4]). This is in accordance with the prediction that generalization is only seen along dimensions in which variability is present during training. This in no way undermines the importance of our findings because it is expected that such generalization would arise given appropriate stimulus variability in the training set.

5. Conclusion

We have applied a new method of speech processing that exaggerates acoustic differences between stimuli to make them more discriminable and thus facilitate training, combined with a "fading"-like schedule of adaptive training and increasing stimulus variability. We have presented promising preliminary data from eight Japanese speakers trained to perceive the English r/l distinction, who show large gains in (untrained) word identification, generalized over the dimensions of training variability. Further work planned or already underway, examines the long-term gains of the training, as well as the relative contribution of each training principle, namely acoustic modification, variability, and adaptive training.

Acknowledgements

The research reported herein was conducted at and supported by Scientific Learning Corporation (Berkeley, CA), who also holds a patent on the method for creating exaggerated stimuli and training; additional patent pending. We thank Talya Salz, Anne Pycha, and Kristin deVivo for recruiting and testing participants, and Bruce McCandliss for stimulating and helpful discussion.

The processing method and results of the initial perceptual evaluation were first presented at the 136th Meeting of the Acoustical Society of America (Norfolk, VA, October 1998).

References

[1] Best C (1995). A direct realist view of cross-language speech perception, in W Strange (Ed), *Speech Perception and Linguistic Experience*, York Press, Baltimore, MD, pp. 171–204.

[2] Logan S and Pruitt J (1995). Methodological issues in training listeners to perceive non-native phonemes, in W Strange (Ed), *Speech Perception and Linguistic Experience*, York Press, Baltimore, MD, pp. 351–377.

[3] Pisoni D and Lively S (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning, in W Strange (Ed), *Speech Perception and Linguistic Experience*, York Press, Baltimore, MD, pp. 433–459

[4] Lively S, Logan J and Pisoni D (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories, *The Journal of the Acoustical Society of America*, 94: 1242–1255.

[5] Terrace H (1963). Discrimination learning with and without "errors." *Journal of the Exeperimental Analysis of Behavior*, 6: 1–27.

[6] McClelland J, Thomas A, McCandliss B, and Fiez J (in press). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data, in J Reggia, E Rupin and D Glanzman (Eds), *Brain, behavioral, and cognitive disorders: The neurocomputational perspective*, Elsevier, Oxford.

[7] Jamieson D and Morosan D (1986). Training non-native speech contrasts in adults: Acquisition of the English /δ/-/θ/ contrast by francophones, *Perception & Psychophysics*, 40: 205–215.

[8] Kubo R, Pruitt J and Akahane-Yamada R (1998). *Isolating the critical segment of AE /r/ and /l/ to enhance non-native perception*, Proc 16th Int Congress Acoustics, Seattle WA, pp. 2965–2966.

[9] Pruitt J (1995). *The perception of Hindi dental and retroflex stop consonants by native speakers of Japanese and American English*, PhD thesis, Dept Psychology, University of South Florida.

[10] Logan J, Lively S and Pisoni D (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. The Journal of the Acoustical Society of America, 89: 874–886.

[11] Rabiner L and Schafer R (1978). *Digital Processing of Speech Signals*, Englewood Cliffs, NJ, Prentice-Hall.