Expanded version of poster 4pSC13 presented at the
136th Meeting of the Acoustical Society of America
Norfolk, VA, 12–16 October 1998

Abstract in JASA 104(3 Pt 2):1855

Modified LPC resynthesis for controlling speech stimulus discriminability

Athanassios Protopapas
Scientific Learning Corporation
Berkeley, CA

Several kinds of modifications of the speech signal, including interpolation of linear predictive coding (LPC) parameters, have been used in the past to create speech stimuli that are ambiguous, i.e., fall perceptually between phonetic categories. In the present article it is demonstrated that the effects of LPC-derived log area ratio coefficients produce signals that are acoustically and perceptually intermediate between phonetic categories. Most importantly, the formulation of this method is extended to include extrapolation from these coefficients to produce pairs of stimuli that are acoustically and perceptually more distinct than the original speech signal pair. These "enhanced" stimuli can be used to gradually train nonnative or impaired listeners to make the corresponding phonetic distinctions.

In an attempt to avoid the cue-impoverished and synthetic-sounding output of formant-based speech synthesizers, researchers have often used edited natural speech to create stimuli for speech perception experiments when absolute control of individual acoustic features is not critical. The methods used to create the stimuli include period-by-period substitution (as used, for example, by Pitt & Samuel, 1993, to create ambiguous segments between /b/ and /m/ along a manner-of-articulation continuum) and waveform averaging (used by McQueen, 1991, for ambiguous fricatives between /s/ and /ʃ/). The resulting stimuli can sound quite natural though it is questionable whether their acoustic properties could derive from a possible vocal tract configuration.

Given recent advances in digital signal processing techniques, some researchers have used digitally altered natural speech to create ambiguous stimuli along continua not amenable to the substitution and averaging methods. For example, a "computer program" was used to create speech sounds ambiguous between /s/ and /ʃ/ and between /t/ and /k/, and /d/ and /g/, by Elman and McClelland (1988). The algorithm was based on linear predictive coding (LPC) resynthesis with some manual tuning (Elman, personal communication), making it possible to affect stop bursts and formant transitions in the desired manner, a feat previously only possible using synthesized speech (generally based on the formant synthesizer by Klatt, 1980). LPC is a much studied and used method and processing code is available in a great variety of development environments. However, such modification-resynthesis techniques seem to still lie beyond the grasp of most speech perception researchers, perhaps because they have not been well described. To address this gap, an LPC-based method is presented here for creating modifed speech signals based on the interpolation of coefficients. This method is then extended to encompass extrapolation for special cases such as perceptual training.

Processing Method

Consider a lossless tube equivalent model of the vocal tract (Rabiner & Schafer, 1978, pp. 82ff), comprising p tubes of equal length and fixed cross-sectional areas A_i. The shape of the model can be defined by the ratios between adjacent areas, called log area ratio coefficients g_i, and these can be derived from a speech waveform through the partial correlation (PARCOR) coefficients k_i, a byproduct of LPC analysis, using the formula (from Rabiner & Schafer, 1978, p. 444):

$$g_i = \log\left(\frac{A_{i+1}}{A_i}\right) = \log\left(\frac{1-k_i}{1+k_i}\right), \quad \text{for } 1 \le i \le p, \qquad (1)$$

where p is the order of LPC analysis. These parameters cannot be guaranteed to 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 original speech signal (to the extent that the all-pole LPC model approximates it). Small deviations from these parameters result in acoustic signals that might have been produced from slightly different vocal tracts, in the sense that the spectral characteristics of the reconstructed signal are close to those of the original analyzed signal and under the same constraints with respect to the number of formants, and their relative positions, that the model allows. Sets of parameters intermediate between those derived from two speech waveforms can then be expected to result in reconstructed speech signals acoustically intermediate between the original two, subject to the same vocal tract constraints, and perceptually ambiguous.

The technique described herein constitutes proprietary technology of Scientific Learning Corp., Berkeley, CA. Patent pending.

Correspondence regarding this article may be sent to Athanassios Protopapas at Scientific Learning Corp., 1995 University Ave., Ste. 400, Berkeley, CA 94704, e-mail protopap@scilearn.com.

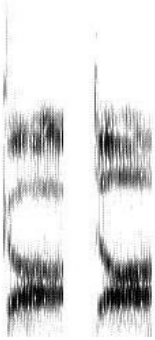

Figure 1. Spectrograms of the natural syllables [ga] (left) and [da] (right) produced by a male speaker. The displayed frequency range is 0-5.5 kHz and each stimulus is 260 ms long.

Consider, for example, the syllables [da] and [ga], recorded by a male speaker, the spectrograms of which are shown in Figure 1. These were analyzed using 24-pole LPC analysis on Hamming-windowed 27.21 ms frames at 9.07 ms intervals. The log area ratio coefficients were derived using Equation 1, and then sets of "intermediate" coefficients were created by linear interpolation between the resulting vectors at the desired positions. That is, one first computes the differences d_i between corresponding coefficients as

$$d_i = g_i^{[\mathrm{da}]} - g_i^{[\mathrm{ga}]}, \;\; 1 \le i \le p. \tag{2}$$

This defines a p-dimensional vector on the straight line that joins the points in p-space defined by the log area ratio coefficients for [da] and [ga]. Any point along this vector relative to $g_i^{[\mathrm{ga}]}$ would define the log area ratio coefficient set for a vocal tract model between those corresponding to the original [da] and [ga]. Specifically, for $r \in [0,1]$ one can define

$$g_i^r = g_i^{[\mathrm{ga}]} + r\,d_i, \;\; 1 \le i \le p \tag{3}$$

and the resulting coefficients can then be converted to PARCOR coefficients using the formula

$$k_i = \frac{1 - e^{g_i}}{1 + e^{g_i}}, \;\; 1 \le i \le p. \tag{4}$$

to be used for LPC resynthesis of a signal with "[da]–[ga] proportions" of r:(1-r).

Figure 2 shows the spectrograms of the resulting resynthesized signals for values of r ranging between zero and one at intervals of 0.25. Notice the intermediate positions of the third formant, one of the most important cues for the perceptual distinction between [da] and [ga] (Harris, Hoffman, Liberman, Delattre, & Cooper, 1958; Smits, Bosch, & Collier, 1996). Notice also that the higher formants, which were not identical for the natural (recorded) [da] and [ga], are not fading in and out between their values for [da] and [ga] but are gradually "shifted" as r changes, so that there is always the same number of formants of the appropriate prominence.

However, nothing restricts application of this method to $0 \le r \le 1$. Using values of r outside the range $[0,1]$ ought to result in pairs of stimuli that are acoustically more different from each other than were the natural stimuli from which the original LPC coefficients were derived. Most importantly, the exaggerated acoustic 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 recorded syllables away from their natural acoustic properties.

Figure 3 illustrates the point with a series of spectrograms for resynthesized stimuli based on a recording of the words "rock" and "lock" ([rak] and [lak]). Results are shown for values of r from -0.75 to 1.75 (based on 14-pole LPC analysis of 27.21 ms Hamming-windowed speech frames 9.07 ms apart). Notice the intermediate positions of the third formant onset and transitions between $r = 0.0$ (corresponding to the original [l]) and $r = 1.0$ (corresponding to [r]) and the more "extreme" formant tracks for r outside this interval. It appears that, for values of r 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 r 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], as intended.

Potential Applications

The ambiguous stimuli, created by interpolating (i.e., for $r \in [0,1]$) between the coefficient sets of the original recordings, can be used in speech perception experiments investigating the effects of "higher-level" factors on phonetic perception such as those mentioned in the introduction or any other experiment in which phonetically ambiguous stimuli are required. The method proposed here guarantees that the resulting stimuli will resemble natural speech in that they will be subject to the same constraints as natural speech signals, such as rate of acoustic change, number and prominence of formants, etc. For ambiguous stimuli along phonetic dimensions other than place of articulation, it may be necessary

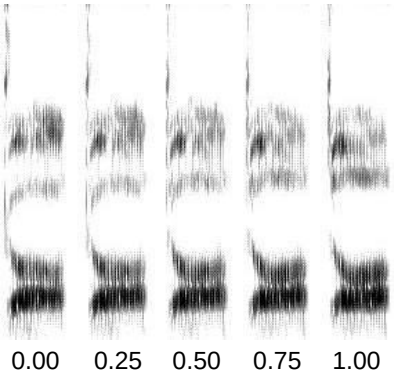

Figure 2. Spectrograms of the resynthesized syllables along a continuum from [ga] (r=0.00) to [da] (r=1.00) using the indicated values of r interpolating between the log area ratio coefficients derived from LPC analysis of the stimuli shown in Figure 1. The displayed frequency range is 0–5.5 kHz and each stimulus is 260 ms long.

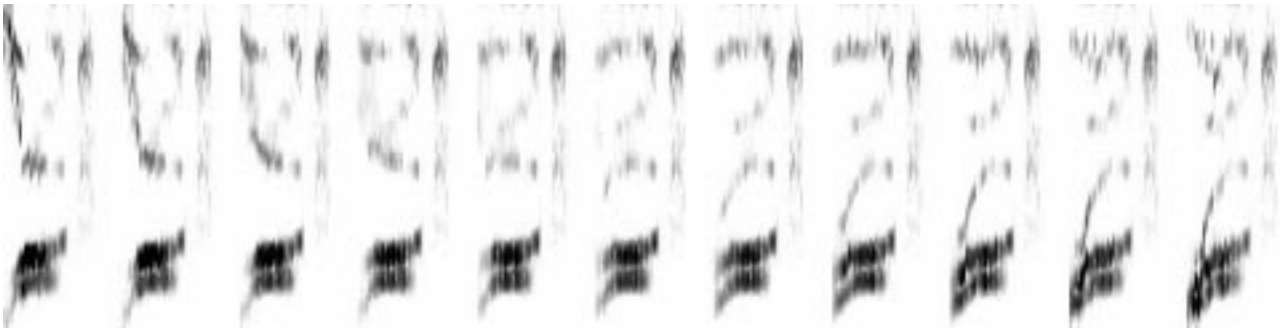

Figure 3. Spectrograms of the resynthesized syllables along a continuum on the line defined by the vector of log area ratio coefficients from [lak] to [rak]. The indicated values of *r* were used with Equation 3 to interpolate and extrapolate from the two sets of LPC-derived log area reflection coefficients. The displayed frequency range is 0-5.5 kHz and each stimulus is 265 ms long.

to supplement the interpolation of LPC coefficients with interpolation of residual energy (gain) in a similar manner.

The potential utility of stimulus pairs of *exaggerated* acoustic discriminability becomes more clear in light of recent advances in our understanding of brain plasticity and its role in phonological representations (Guenther & Gjaja, 1996; Merzenich et al., 1993). It is a common assumption that in the normally developing person in the linguistic context of his/her native language, the (language-dependent) acoustic cues signaling phonetic distinctions are processed in such a way that, over time, stable phonological representations develop that nonlinearly map acoustic features to linguistic (phonemic) categories. In contrast, in some specifically language impaired (SLI) and dyslexic children, phonological representations have been argued to be weak, possibly as a result of impaired auditory reception (see discussions in Bishop, 1992; Farmer & Klein, 1995; Gathercole & Baddeley, 1993; Masterson, Hazan, & Wijayatilake, 1995; Wagner & Torgesen, 1987). Training SLI children with acoustically modified speech stimuli has been shown to result in substantial improvements in speech perception and language skills (Tallal et al., 1996; Merzenich et al., 1996). The method presented here for creating "overdiscriminable" stimulus pairs may prove helpful in the context of such training, by allowing training to become (a) more specific to each child, and (b) more specific to each phonetic contrast. Individual customization of training sets is made feasible because, given LPC-derived coefficients for a large set of syllables, it is technically feasible to exaggerate those pairwise distinctions at which each child is most deficient. In addition, stimulus specificity means that each syllable is not generically "enhanced" but is specifically removed from the one with which it is most confusable because the coefficient extrapolation is done along the difference vector between particular stimuli.

An additional possible use of such overdiscriminable stimuli lies in the context of second-language learning. Consider, for example, a language in which an English phonemic contrast is neutralized, such as Japanese or Korean, in which there is no phonemic distinction between [r] and [l]. Lifelong experience with such a language has resulted in the Japanese speakers' inability not only to produce the [r]–[l] contrast, but also to perceive it (Miyawaki et al., 1975). Attempts to train Japanese speakers to perform this distinction have met with limited success (Strange & Dittman, 1984; Lively, Logan, & Pisoni, 1993; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; Logan, Lively, & Pisoni, 1991), perhaps because of self-reinforcing properties of early phonetic category learning. However, as recent modeling efforts have shown, it may be possible to split an established categorical representation in two if the distinction between the members of the two subcategories are initially exaggerated (McClelland, 1998). That is, it may be easier to create two separate categories for [r] and [l] in place of the single Japanese category by training with [r] and [l] artificially modified to be more distinct (i.e., perceptually more different from each other). It is also a well known fact in behavioral training that learning is facilitated when initiated at a level of difficulty where the task can be performed, if with difficulty, relative to the case in which training is initiated at a level that precludes successful performance. Obviously, the use of the present LPC-based modification method lends itself perfectly to such an application.

In the following section preliminary data are presented on the perception of the resynthesized stimuli by native and nonnative listeners. The resulting identification and discrimination curves confirm with our expectations and indicate that the proposed applications are at least worth exploring.

Perceptual Evaluation

Informal listening of these stimuli indicated that they can sound quite natural for $0 \leq r \leq 1$ if the processing parameters (e.g., LPC order, frame length, sampling rate) are carefully adjusted and the derived parameters (energy and pitch) are manually tuned, as necessary. The resynthesized stimuli become progressively less natural sounding as r moves away from the $[0, 1]$ interval, with some high-frequency artifacts and increased amplitude variation, necessitating additional adjustments in the intermediate parameters, constraints on the amplitude of the resynthesized signals, coefficient smoothing, or sometimes repetition of the procedure based on different recordings. Simultaneous two-dimensional inter/extra-polation along both the log area co-

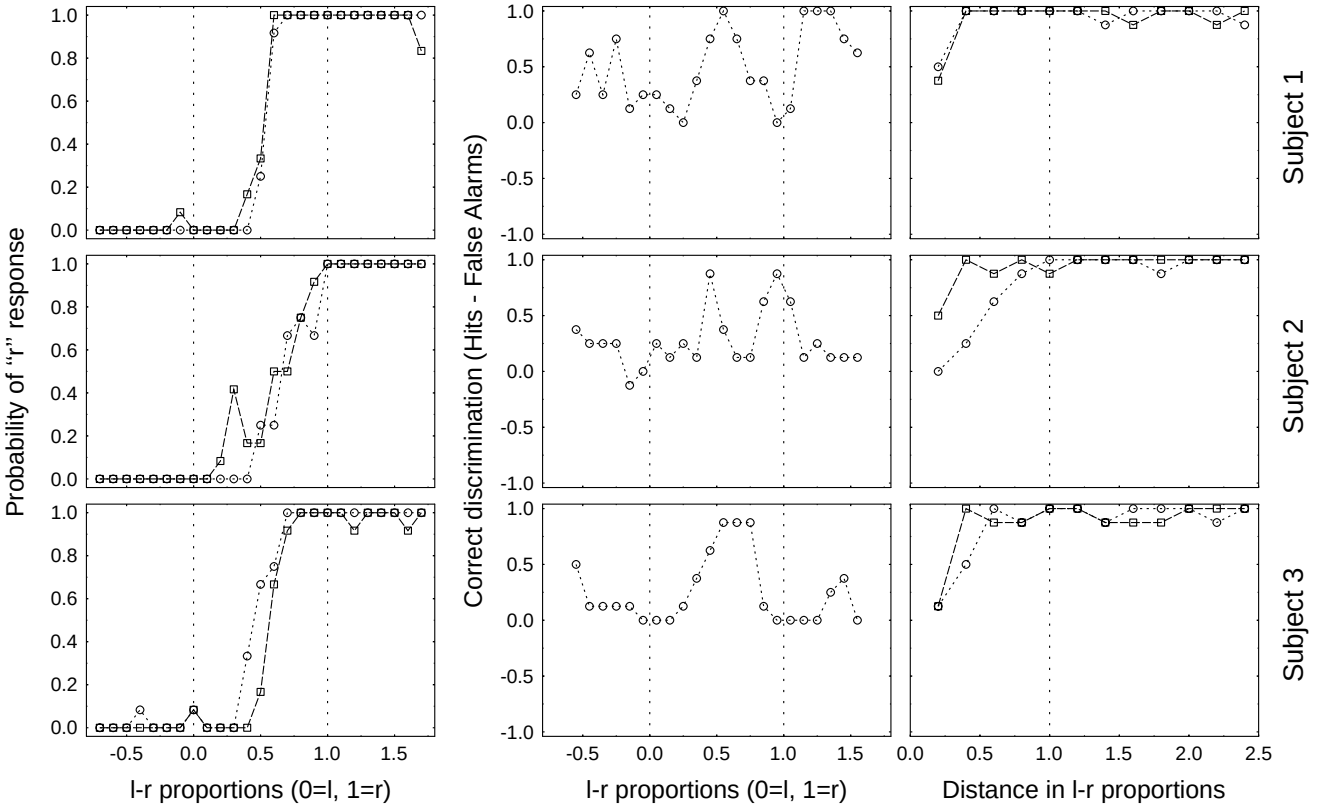

Figure 4. Identification and discrimination curves for 3 native American English speakers with the resynthesized stimuli along the [ra]–[la] continuum for two stimulus voices (male: squares on dashes; female: circles on dots). Each row shows data from a single subject. Left column: Identification (labeling) performance on resynthesized stimuli for *r* (see Equation 3 between -0.7 and 1.7 in 0.1 steps. Middle column: Discrimination performance on stimulus pairs 0.3 units apart (in *r* units) along the continuum. Right column: Discrimination performance on stimulus pairs symmetric with respect to *r*=0.5 for increasing values of *r* distance.

efficient dimension and time may aid in reducing some of the artifacts, thus speeding up the procedure which, overall, produces stimuli that can be very useful in speech research. Moreover, splicing only the critical portion of the ambiguous resynthesized signal onto the natural (original) remaining utterance improves the naturalness of the entire stimulus. To be successful (and undetectable), such splicing must be done at an appropriate point in the waveform, such as a zero crossing, preserving the fundamental period across the juncture between the resynthesized and the natural segments.

Identification and discrimination testing of the resynthesized stimuli is necessary to ensure that their perceptual characteristics are indeed as desired, i.e., that the stimuli can be identified as one of the two intended phonemes in the expected (categorical) manner. Discrimination between stimuli should increase with increased difference in *r* values; most importantly, discrimination should be easiest when a stimulus pair straddles the "boundary" between [r] and [l] labels and most difficult when the two stimuli to be discriminated are assigned the same phonetic category.

Figure 4 shows the identification and discrimination performance of 3 adult native English speakers using the stimuli from two [ra]–[la] continua (one with a male and one with a female voice) for *r* between -0.7 and 1.7 in steps of 0.1. The relatively abrupt perceptual transition between [r] and [l] labeling and the peak in discrimination roughly coinciding with the perceptual boundary between [r] and [l] indicate that these resynthesized stimuli are perceived in a manner comparable to the synthetic speech stimuli used in previous experiments. Note also that the exaggerated stimuli are consistently labeled as exemplars of their respective (exaggerated) category (left column, points outside the $[0,1]$ range), and that stimulus pairs separated by at least the natural [r]-[l] distance (i.e., 1.0 or more in the right column) are perfectly discriminable for native English speakers, as expected. The increased discrimination for some stimulus pairs 0.3 *r*-units apart outside the $[0,1]$ range (middle column) is partly due to unwanted artifacts introduced during extrapolation processing and in part because stimulus exaggeration sometimes causes phonetic distortion (here especially on the [r] side). This is only to be expected since the purpose of the processing is to push phonetic exemplars away from their natural position and thus possibly to the fringes or entirely outside their respective phonetic category; what is important is that the acoustic differences between stimuli thus created are of the same kind as the differences between the natural tokens.

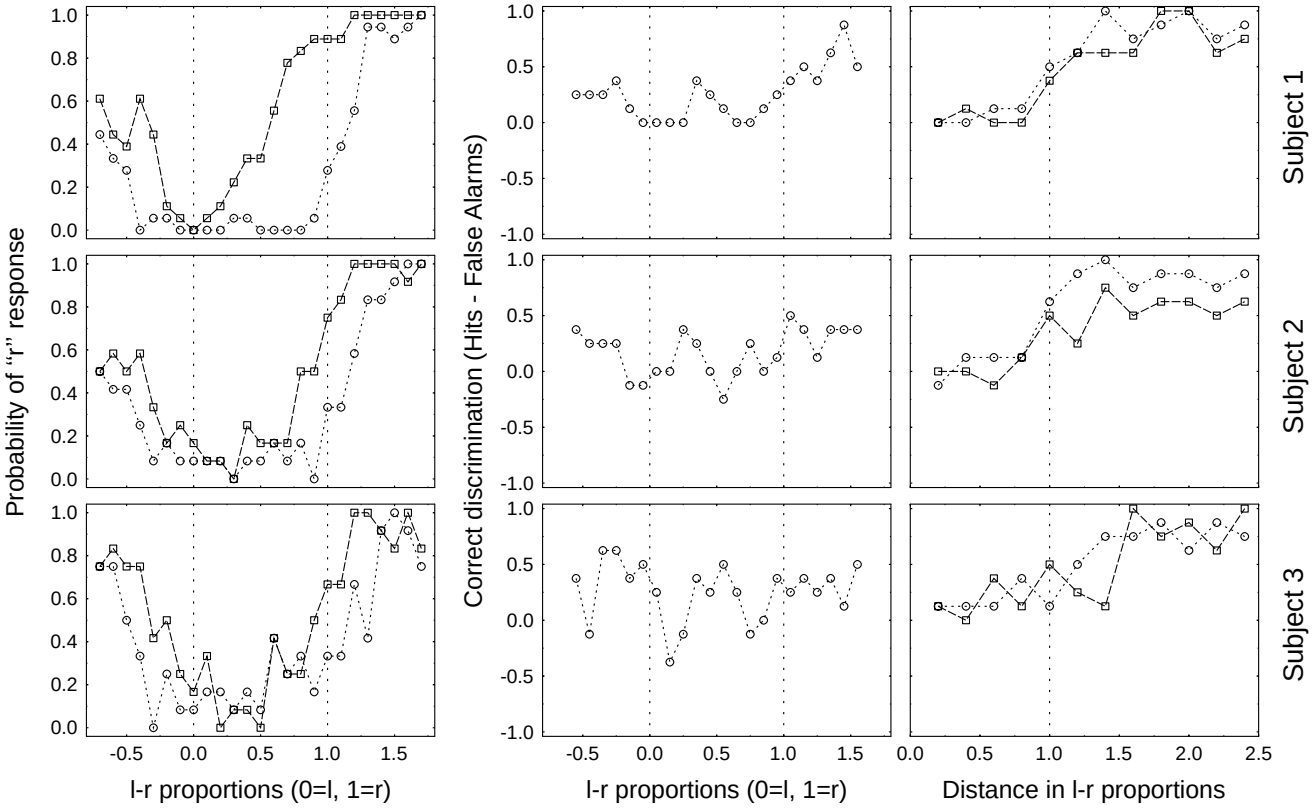

Figure 5. Identification and discrimination curves for 3 Japanese speakers with the resynthesized stimuli along the [ra]–[la] continuum for two stimulus voices (male: squares on dashes; female: circles on dots). Each row shows data from a single subject. Left column: Identification (labeling) performance on resynthesized stimuli for *r* (see Equation 3) between -0.7 and 1.7 in 0.1 steps. Middle column: Discrimination performance on stimulus pairs 0.3 units apart (in *r* units) along the continuum. Right column: Discrimination performance on stimulus pairs symmetric with respect to *r*=0.5 for increasing values of *r* distance.

It has been hypothesized that with sufficient exaggeration, listeners unable to discriminate the natural stimuli would be able to make accurate disctinctions of the processed stimuli. To illustrate this point, Figure 5 shows the performance of three Japanese listeners on the identification and discrimination of the resynthesized [ra]–[la] stimuli. The subjects were one male and two female students in their twenties, recruited at Berkeley through a newspaper ad and paid for their participation. Testing was done in a quiet room at the offices of Scientific Learning Corp. All three subjects were informally judged to be very inaccurate in [r]–[l] production; their performance in identifying words beginning with a singleton [r] or [l] consonant ranged between 60 and 70%.

Note the unusual U-shaped identification curves for all three subjects, with most stimuli in the natural (i.e., $[0,1]$) range identified as "L" and with stimuli from one voice (the male in this case) rated as "R" more often than stimuli from the other (the female) voice. The discrimination curves of these subjects also attest to their very poor performance, never exceeding 0.5 (proportion of hits minus false alarms) in the natural and ambiguous range, in striking contrast to the natives' performance (Figure 4). It must be noted that at least two of these Japanese listeners (subjects 2 and 3) seem to have been unable to use the slight artifacts and distortions present in the stimuli in making their discrimination judgments, so their performance with pairs in the "exaggerated [r]" range is also very low. This is further evidence of their lack of an appropriate phonetic category relative to which some stimuli may be judged to be worse exemplars (as by the native English speakers).

Most importantly, let us turn our attention to the discrimination performance of the three Japanese subjects on pairs of stimuli taken symmetrically around the acoustic [ra]–[la] midpoint (Figure 5, right column). Clearly, discrimination between the naturally spaced resynthesized tokens (*r* values of zero and one, corresponding to natural [l] and [r], respectively) is very poor, as expected. However, discrimination of stimuli spaced further apart is increasingly improved, approaching or attaining perfect performance for distances around 1.5 and higher (i.e., for the pair of stimuli with *r* values of -0.25 and 1.25). Thus the data are consistent with our hypothesis 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 well on the basis of this acoustic cue (or set of cues) if it is sufficiently exaggerated to become salient.

The implications of this demonstration for training nonnative phonetic contrasts is potentially very significant, because standard perceptual training practice since the days of Skinner dictates that initiation of training from an easily discriminable stimulus condition enables or at least greatly enhances learning when combined with a gradual modification of the training stimuli through increasingly difficult conditions towards the desired target stimuli. This prediction is currently being tested in training Japanese listeners to discriminate English [ra] from [la].

Summary and Conclusion

A method based on LPC analysis has been presented for resynthesizing speech stimuli based on a pair of natural recorded tokens. The LPC-based vocal tract equivalent model coefficients are interpolated to generate stimuli perceptually ambiguous between the two original tokens. Extrapolation outside the range defined by the natural tokens along the line connecting them in model coefficient space results in "exaggerated" stimuli that differ spectrally in the same way the original natural pair did but more so.

Perceptual testing has confirmed the expected perfomance pattern for native English speakers with both the ambiguous and the exaggerated stimuli. Furthermore, it was shown that the exaggerated stimuli are more discriminable than those synthesized with parameter values corresponding to the natural tokens. Japanese speakers who were demonstrably unable to discriminate between natural [r] and [l] tokens were able to discriminate between pairs of stimuli exaggerated according to the method proposed here. It is expected that listeners from diverse native linguistic backgrounds or with an acoustically-based language learning impairment that hinders their phonetic perception (and possibly production) ability may be succesfully trained using such exaggerated stimuli to accurately make the appropriate phonetic distinctions.

References

Bishop, D. V. M. (1992). The underlying nature of specific language impairment. *Journal of Child Psychology and Psychiatry*, *33*, 3-66.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, *27*, 143-165.

Farmer, M. E., & Klein, R. M. (1995). The evidence for a temporal processing deficit linked to dyslexia: A review. *Psychonomic Bulletin and Review*, *2*, 460-493.

Gathercole, S. E., & Baddeley, A. S. (1993). *Working memory and language.* Hillsdale, NJ: Erlbaum.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, *100*, 1111-1121.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Efect of third-formant transitions on the perception of the voiced stop consonants. *The Journal of the Acoustical Society of America*, *30*, 122-126.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, *67*, 971-995.

Lively, S. E., Logan, J. S., & 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. *The Journal of the Acoustical Society of America*, *94*, 1242-1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, *96*, 2076-2087.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*, 874-886.

Masterson, J., Hazan, V., & Wijayatilake, L. (1995). Phonemic processing problems in developmental phonological dyslexia. *Cognitive Neuropsychology*, *12*, 233-259.

McClelland, J. L. (1998, April). *Reopening the critical period: A Hebbian account of interventions that induce change in language perception.* Presented at the 1998 Annual Meeting of the Cognitive Neuroscience Society in San Francisco, CA.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 433-443.

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science*, *271*, 77-81.

Merzenich, M. M., Schreiner, C., Jenkins, W., & Wang, X. (1993). Neural mechanisms underlying temporal integration, segmentation, and input sequence representation: Some implications for the origin of learning disabilities. *Annals of the New York Academy of Sciences*, *682*, 1-22.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics*, *18*, 331-340.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phonetic identification task. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 699-725.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals.* Englewood Cliffs, NJ: Prentice-Hall.

Smits, R., Bosch, L. ten, & Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. *The Journal of the Acoustical Society of America*, *100*, 3852-3864.

Strange, W., & Dittman, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics*, *36*, 131-145.

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., Schreiner, C., Jenkins, W. M., & Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, *271*, 81-84.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, *101*, 192-212.