Running head: MODIFIED LPC RESYNTHESIS

Modified LPC resynthesis for increased speech stimulus discriminability[a)]

Athanassios Protopapas[b)]

Scientific Learning Corporation

Berkeley, CA

Bruce McCandliss

Center for the Neural Basis of Cognition

Carnegie Mellon University

[a)]The technique described herein constitutes proprietary technology (patent pending) of Scientific Learning Corp., Berkeley, CA. It was first presented at the 136th Meeting of the Acoustical Society of America held in Norfolk, VA, October 1998 [J. Acoust. Soc. Am. **104**,1855 (1998)].

[b)]Correspondence address: Department of Educational Technology, Institute for Language and Speech Processing, Epidavrou & Artemidos 6, Marousi, GR-151 25 ATHENS, Greece; Phone: +30 210 6875409; Fax: +30 210 6854270; e-mail: protopap@ilsp.gr

Abstract

Efficient phonetic training for second language learning and for language impairment remediation requires increased salience of distinctive acoustic differences. Linear predictive coding (LPC) analysis of speech has been sometimes used for constructing phonetically ambiguous stimuli via parameter interpolation. 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. The method is extended to extrapolation from these coefficients, resulting in resynthesized pairs of stimuli that are acoustically and perceptually more distinct than the original speech signal pair. These ‘‘exaggerated’’ stimuli can be used to gradually train nonnative or impaired listeners to make the corresponding phonetic distinctions.

PACS numbers: 43.72.Ew, 43.71.Es, 43.71.Hw

INTRODUCTION

Recent studies with specifically langauge impaired (SLI) children have demonstrated that some of the difficulties these childeren have in phonetic perception can be rapidly improved through perceptual training procedures (Tallal et al., 1996; Merzenich et al., 1996). One hallmark of the perceptual training procedures used in these studies was that speech tokens that were initially difficult for these subjects to differentiate were presented in an accoustically modified form aimed at rendering them highly discriminable. As training progressed, the degree of modification was gradually "faded" until subjects were able to successfully differentiate the unmodified tokens. These training procedures have essentially found a way to apply well established psychological training principles (Terrace, 1963) to the domain of phonetic perception. Similar techniques have proven successful in improving adults' abilities to differentiate non-native phonemic distinctions, as in the case of French Canadians learning to discriminate /ð/ from /θ/ (Jamieson & Morosan, 1986; Morosan & Jamieson, 1989).

Thus far, the acoustic modifications employed in these training studies have been restricted to rather simple dimensions, such as temporal stretching of segments that contain fast formant transitions, selective amplification of certain segments, or both applied together. While these methods of creating overly discriminable stimuli captalize on dimensions that can be easily and generically manipulated and also easily faded in a linear fashion during training, it is possible that these forms of manipulation effectively enhance only a small subset of the possible phonetic contrasts that could benefit from this training paradigm. For

example, these manipulations do little to alter gross spectral characteristics (e.g., formant frequencies) that are critical to many phonetic distinctions of vowels, fricatives, and liquids.

We submit that the effectiveness of these training studies could be improved and expaned if the methods that researchers used to produce the gradients of overly discriminable stimuli incorporated the specific acoustic dimensions along which the phonemes under training differ, including, in particular, distinctive spectral information. By exaggerating the specific spectral differences inherent in the contrast between two phonemes, training efforts could be enhanced in two ways: (a) these training techniques could be applied to a wider range of phonemic contrasts, including those not rendered more discriminable by selective amplification or temporal stretching; and (b) creating a continuum of training stimuli (from overly discriminable to unmodified) that systematically vary along the same spectral dimensions that are critical to the phonemic contrast being trained may provide an additional basis for enhancing learning.

Consider, for example, the case in which the training goal is to improve distinctive categorization of the English liquid consonants /r/ and /l/. These are differentiated primarily on the basis of an initial steady state and subsequent transition of the third formant frequency (Miyawaki et al., 1975), and are thus unlikely to be made more distinct by either temporal stretching or selective amplification of critical regions. Difficulties in learning to discriminate these two phonemes are well documented in adult listeners whose native language neutralizes this distinction, such as Japanese or Korean (Miyawaki et al., 1975). One class

of explanation for such difficulties in adult second language learning relies on a form of biological critical period in which the youthful facility for learning new phoneme categories is reduced in adulthood (Lenneberg, 1967; Scovel, 1988). However, a host of laboratory training studies with Japanese natives have demonstrated at least some improvements in /r/--/l/ discrimination over multiple weeks of intervention (Strange & Dittman, 1984; Lively, Logan, & Pisoni, 1993; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; Logan, Lively, & Pisoni, 1991), suggesting that although these adults still maintain some plasticity in this domain, it might not be typically elicited by years of natural exposure to English as a second language.

Focusing more directly on the specific role of native language (L1) interference in perceiving non-native phonemic distinctions may provide a more explicit account of why Japanese speaking adults might have persistent difficulty learning to discriminate /r/ and /l/ based on natural exposure to English, and why certain training methods can help to overcome these difficulties. For example, Japanese-speaking adults may perceive acoustic inputs corresponding to the English /r/ and /l/ phonemes as belonging to a single Japanese phonological category (Jones, 1967; Takagi, 1993; Best & Strange, 1992), leading the listener to form indistinguishable percepts for these two acoustic inputs.

Recent work with connectionist models provide an explanation for how the top-down effect that perceptual categories have on perception of acoustic information can be self-reinforcing, and how this tendency can impede learning in certain circumstances (McClelland, Thomas, McCandliss, & Fiez, in press). For example, if a range of acoustic inputs lead to the

activation of a single perceptual category representation, then the link between that range of acoustic patterns and the representation of the perceptual category will be reinforced through Hebbian learning. The prepotent tendency of Japanese natives to perceive acoustic inputs that would be labeled /r/ and /l/ by English listeners as members of the same native perceptual category is reinforced each time an acoustic input in this range is heard, thus creating a self-perpetuating L2 learning difficulty.

The modeling work suggests that it is possible to overcome this self-reinforcing tendency by exaggerating the distinction between the acoustic inputs, and thus avoiding the activation of the native perceptual category (McClelland et al., in press). That is, modeling suggests that it is possible to form two separate categories for /r/ and /l/ in place of the single Japanese category by training with [r] and [l] stimuli artificially modified to be more distinct, i.e., perceptually more different from each other. Gradually, as the degree of exaggeration is faded, it should become possible to discriminate unmodified [r] and [l] tokens.

Our approach to creating overly discriminable speech contrasts can be conceived as an extension of constructing phonetically *ambiguous* stimuli based on a pair of natural (recorded) speech. The creation of acoustic continua between two speech sounds has a considerable history in linguistic research. In the context of research into the effects of acoustic characteristics on phonetic perception and lexical access, as well as in sentence processing and phonetic or lexical effects thereon, phonetically ambiguous stimuli are often needed to neutralize or delay phonetic perceptual decisions thus bringing so-called higher levels of processing in

a more prominent, and thus observable, position. When absolute control of individual acoustic features is not critical (hence the cue-impoverished and unnatural-sounding output of formant synthesizers can be avoided), the methods used to create ‘‘intermediate’’ stimuli have included 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/). Whether the acoustic properties of stimuli thus created could be produced from any possible vocal tract configuration is questionable, therefore this method may have limited utility in applied settings (e.g., phonetic training).

Recently, some researchers have used digital signal processing algorithms for phonetic continua not amenable to the substitution and averaging methods. For example, a ‘‘computer program’’ was reported to have been 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 (Atal & Hanauer, 1971; Markel & Gray, 1976) is a much studied and used method, and processing code is available in a great variety of development environments.

Such LPC-resynthesis techniques may hold advantages as a general acoustic modification approach that can be applied to a range of stimulus contrasts to create intermediate versions of two speech stimuli without

making any explicit assumptions about the nature of those differences. Furthermore, since this method produces a multidimensional vector of coefficients that captures critical spectral difference between two stimuli, by extrapolating along this vector it should be possible to create a range of exaggerated versions of the stimuli. When such stimuli are used in perceptual fading training paradigms, the modifications that are faded would represent different points along the same multimentional vector that defines the spectral difference between the two stimuli being trained.

Preliminary perceptual data are presented to demonstrate in principle the applicability of the proposed method for the test case of making /r/ and /l/ more discriminable for Japanese-native adults who learned English as a second language. Subsequent work is underway that extends this work in training studies based on the perceptual fading paradigm (McCandliss, Fiez, Conway, Protopapas, & McClelland, 1998).

I. PROCESSING METHOD

Insert Figure 1 about here

A linear predictive coding (LPC) model of the speech signal can be formulated to be equivalent to a lossless tube ‘‘vocal tract model’’ and thus one can use LPC analysis to derive an equivalent ‘‘vocal tract shape’’ for values derived from the LPC analysis technique. Consider a lossless tube equivalent model of the vocal tract (after Rabiner & Schafer, 1978, pp. 82ff) comprising p tubes of equal length l/p and fixed cross-sectional areas A_i (Figure 1). Traveling waves in each tube are subject to pressure

and volume velocity continuity at the boundaries between adjacent tubes, where mismatched impedance (due to cross-sectional area differences) results in wave reflection at the junctions. Each junction can be characterized by the amount of backward-traveling wave reflected, called 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

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

Setting the radiation load at the ''lips'' (e.g., $A_{p+1} = \infty$ for the completely lossless case), the transfer function of this system can be shown to be of the form

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

which is the same as the steady-state system function of a slowly time-varying digital filter obtained by linear prediction analysis of order p. Moreover, the partial correlation (PARCOR) coefficients k_i derived in the course of solving the LPC equations to compute the predictor coefficients α_i turn out to be related to the reflection coefficients of the lossless tube model simply as

$$r_i = -k_i. \tag{3}$$

This simple relationship between LPC coeffients and parameters in an equivalent vocal tract model demonstrates how LPC analysis can be used to derive an equivalent ''vocal tract shape.''

In practice, the shape of the model is defined by the log ratios between adjacent areas, called log area ratio coefficients g_i, which are

derived from the speech waveform using the formula (from Rabiner & Schafer, 1978, p. 444):

$$g_i = \log\left(\frac{A_{i+1}}{A_i}\right) = \log\left(\frac{1-k_i}{1+k_i}\right), \quad \text{for } 1 \le i \le p. \tag{4}$$

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. That is, 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.

Insert Figure 2 about here

Consider, for example, the syllables [da] and [ga], recorded by a male speaker, the spectrograms of which are shown in Figure 2. 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 4, 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 δ_i between corresponding coefficients as

$$\delta_i = g_i^{[da]} - g_i^{[ga]}, \quad 1 \le i \le p. \tag{5}$$

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 $g_i^{[ga]}$ would define the log area ratio coefficient set for a vocal tract model in between those corresponding to the original [da] and [ga]. Specifically, for $\lambda \in [0,1]$ one can define

$$g_i^{\lambda} = g_i^{[ga]} + \lambda \delta_i, \quad 1 \le i \le p \tag{6}$$

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

$$k_i = \frac{1 - e^{g_i}}{1 + e^{g_i}}, \quad 1 \le i \le p. \tag{7}$$

to be used for LPC resynthesis of a signal with ‘‘[da]--[ga] proportions’’ of λ:(1-λ).

Insert Figure 3 about here

Figure 3 shows the spectrograms of the resulting resynthesized signals for values of λ 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, ten 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 λ changes, so that there is always the same number of formants of the appropriate prominence.

However, nothing restricts application of this method to $0 \leq \lambda \leq 1$. Using values of λ 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 (spectral) distinction will be obtained by distorting the recorded syllables away from their natural acoustic properties.

Insert Figure 4 about here

Figure 4 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 λ 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 $\lambda = 0.0$ (corresponding to the original [l]) and $\lambda = 1.0$ (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], as intended.

Informal listening of the stimuli thus created indicated that they can sound quite natural for $0 \leq r \leq 1$ for a reasonable range of processing parameters (LPC order, processing window length, frame rate, sampling rate). The [da]--[ga] and [rak]--[lak] examples above demonstrate the applicability of the method for LPC orders 14 to 24. The resynthesized stimuli become progressively less natural sounding as λ moves away from the $[0,1]$ interval, necessitating some additional fine-tuning, possibly including setting the LPC order by trying several values, imposing an amplitude envelope on the resynthesized signal to avoid extreme fluctuations, and smoothing the reflection coefficients in time. Splicing only the critical (distinctive) 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.

In the following section preliminary data are presented on the perception of such resynthesized stimuli by native and nonnative listeners. The resulting identification and discrimination curves confirm the expected discriminability manipulation, thus suggesting the feasibility of the proposed applications.

II. PERCEPTUAL EVALUATION

Insert Figure 5 about here

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. An identification test is conducted by presenting participants with a single stimulus in each trial (e.g., 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 then plotted against position in the acoustic continuum; typically the resulting curve is flat around both endpoints with a very abrupt transition from one to the other response category at some point close to the acoustic ‘‘middle.’’ This phenomenon of abrupt perceptual transition given gradual acoustic change is termed ‘‘categorical perception’’ and is often thought to constitute the hallmark of phonetic perception.

In order to ascertain that there is indeed an abrupt perceptual transition and not merely an artifact of having only two response categories, a discrimination test is required to assess the participants’ ability to discriminate between pairs of stimuli drawn from the same continuum. One way to conduct such a test is with two stimuli being presented in each trial and the participant judging them to be ‘‘same’’ or ‘‘different.’’ In the method used here, each pair of stimuli were synthesized with λ values at a fixed distance of 0.3. For example, a data

point plotted for discrimination at 0.55 shows the subject's ability to discriminate a stimulus synthesized with λ=0.4 from one with λ=0.7. Typically, when the two stimuli to be discriminated belong to the same ''perceptual category,'' i.e., are given the same phonetic label in the identification test, they are difficult to discriminate, whereas stimuli drawn from different categories, i.e., from opposite sides on the transition boundary from the identification test, they are easy to discriminate.

In the context of the present method, there is an additional discrimination test of interest. Specifically, discriminability between stimuli should increase with increased difference in λ values. That is, in a task where pairs of stimuli are judged ''same'' or ''different,'' performance should be better for, e.g., stimuli 2.0 λ-units apart (i.e., one synthesized with λ=−0.5 and the other with λ=1.5) than for stimuli 0.5 λ-units apart (i.e., one with λ=0.25 and the other with λ=0.75).

Insert Figure 6 about here

Figure 5 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 λ 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 λ-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.

According to our hypothesis, given 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 6 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 who had lived in the U.S. for several months, 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%.

In contrast to the ‘‘categorical’’ identification curves obtained from the native English speakers, note the U-shaped identification curves for all three Japanese 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 5). 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 6, right column). Clearly, discrimination between the naturally spaced resynthesized tokens (λ values of zero and one, corresponding to natural [l] and [r], respectively) is very poor. 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 λ 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.

III. 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.

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 trainee is most deficient. It is thus possible to tailor training schedules to the specific areas of weakness for each individual. In addition, stimulus specificity means that each syllable is not generically ''enhanced'' but is specifically acoustically moved away from the one with which it is most confusable. This is because of the coefficient extrapolation along the difference vector between particular stimuli, thus also increasing training specificity and efficiency. This modification specificity is likely to maximize the utility of the modification itself because salience of the discrimination is affected directly at the most relevant acoustic feature and not indirectly (as, for example, with selective amplification). The proposed method is also simple in that a single parametrically varying vector representing the acoustic characterization of a phonetic contrast captures both the ''natural'' and the ''exaggerated'' speech forms as well as those in between.

The implications of this demonstration for training nonnative phonetic contrasts are very significant because standard perceptual training practice 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 [r] from [l] in a variety of contexts. In addition, the proposed method opens up new research possibilities for second-language phonetic learning because the specificity of modification raises the

question of transfer to untrained phonetic contrasts that differ along a similar dimension. It remains to be investigated whether increasing the salience of a contrast by affecting directly the relevant acoustic properties has the effect of generalizing to other contrasts with greater efficiency than previous methods have achieved.

References

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, *50*, 637--655.

Best, C., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *J. Phonetics*, *20*, 305-330.

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

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. *J. Acoust. Soc. Am.*, *30*, 122-126.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Percept. Psychophys.*, *40*, 205-215.

Jones, D. (1967). *The phoneme: Its neture and use.* Cambridge, UK: Cambridge Univ. Press.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, *67*, 971-995.

Lenneberg, E. (1967). *Biological foundations of language.* New York: Wiley.

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. *J. Acoust. Soc. Am.*, *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. *J. Acoust. Soc. Am.*, *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. *J. Acoust. Soc. Am.*, *89*, 874-886.

Markel, J. D., & Gray, A. H., Jr. (1976). *Linear prediction of speech.* Berlin: Springer-Verlag.

McCandliss, B. D., Fiez, J. A., Conway, M., Protopapas, A., & McClelland, J. L. (1998). Eliciting adult plasticity: Both adaptive and non-adaptive training improves Japanese adults' identification of English /r/ and /l/. In *Soc. Neurosci. Abs.* (Vol. 24, p. 1898). Los Angeles, CA.

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

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *J. Exp. Psychol. Hum. Percept. Perform.*, *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.

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. *Percept. Psychophys.*, *18*, 331-340.

Morosan, D. E., & Jamieson, D. G. (1989). Evaluation of a technique for training new speech contrasts: Generalizations across voices, but not word-position or task. *Journal of Speech and Hearing Research*, *32*, 501-511.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phonetic identification task. *J. Exp. Psychol. Hum. Percept. Perform.*, *19*(4), 699-725.

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

Scovel, T. (1988). *A time to speak: A psycholinguistic inquiry into the critical period for human speech.* New York: Newbury House/Harper & Row.

Smits, R., ten Bosch, L., & Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. *J. Acoust. Soc. Am.*, *100*, 3852-3864.

Strange, W., & Dittman, S. (1984). Effects of discrimination training on

the perception of /r-l/ by Japanese adults learning English. *Percept. Psychophys.*, *36*, 131-145.

Takagi, N. (1993). *Perception of american english /r/ and /l/ by adult japanese learners of english: A unified view.* Unpublished doctoral dissertation.

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.

Terrace, H. S. (1963). Discrimination learning with and without ‘‘errors’’. *J. Exp. Anal. Behav.*, *6*, 1-27.

Author Note

We thank Srikantan Nagarajan for very useful comments on the manuscript.

Figure Captions

Figure 1. A lossless tube model comprising p concatenated tubes, each of constant cross sectional area A_i and length l/p.

Figure 2. 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.

Figure 3. Spectrograms of the resynthesized syllables along a continuum from [ga] (λ=0.00) to [da] (λ=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 2. The displayed frequency range is 0--5.5 kHz and each stimulus is 260 ms long.

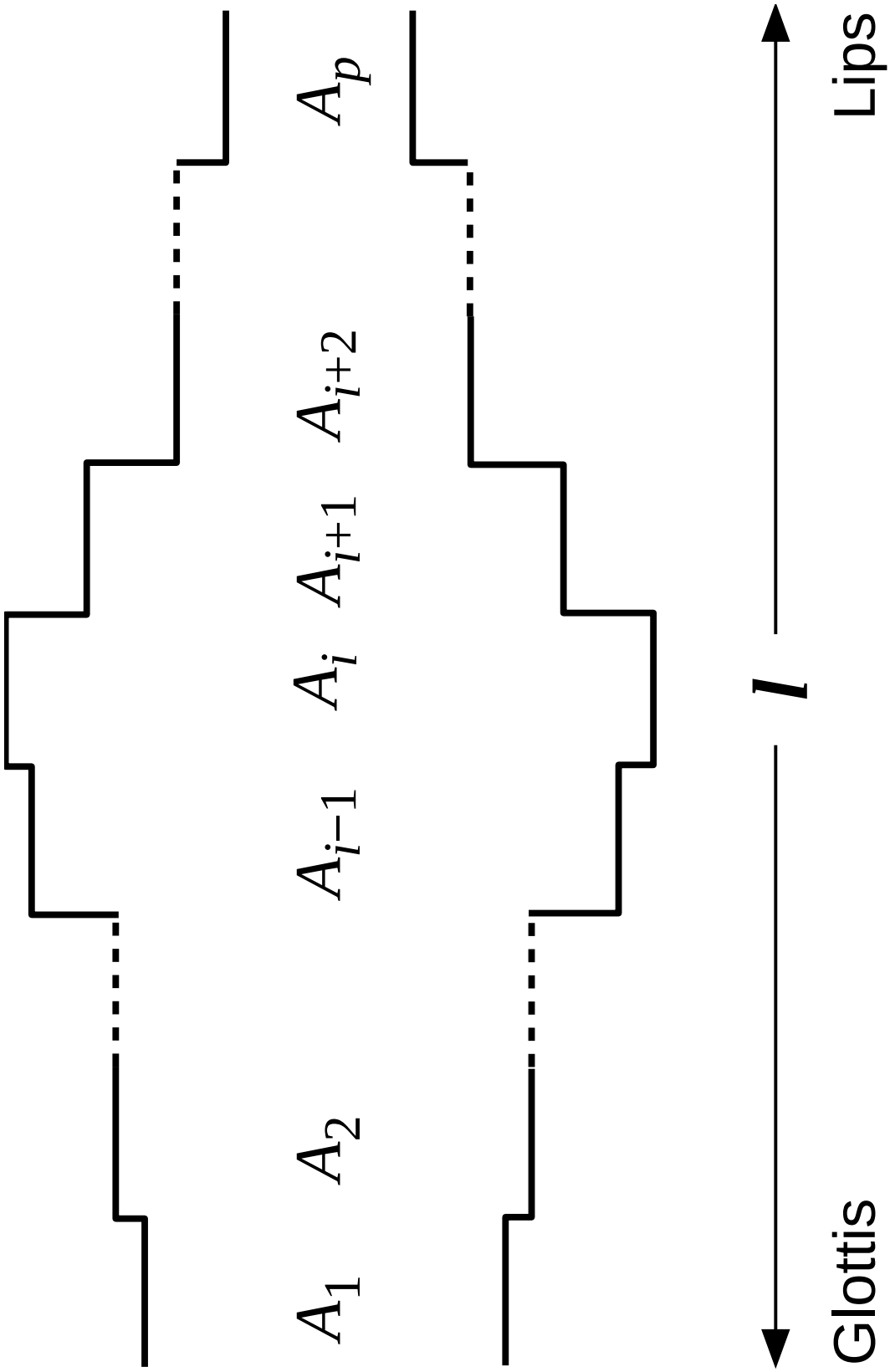
Figure 4. Spectrograms of the resynthesized syllables along a continuum on the line defined by the vector of log area ratio coefficients from [lak] to [rak]. The indicated values of λ were used with Equation 6 to interpolate and extrapolate from the two sets of LPC-derived log area reflection coefficients. The displayed frequency range is 0-5.5 kHz and each stimulus is 265 ms long.

Figure 5. 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 λ (see Equation 6 between -0.7 and 1.7 in 0.1 steps. Middle column: Discrimination performance (i.e., hits−false alarms over total number of

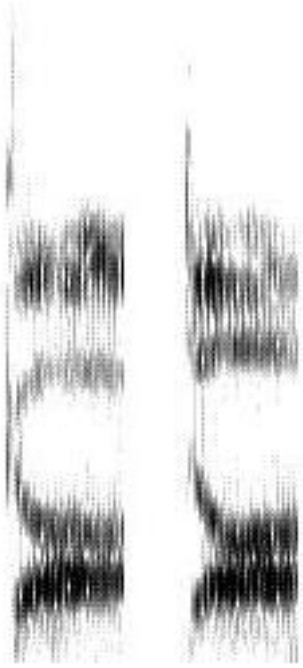
trials) on stimulus pairs 0.3 units apart (in λ units) along the continuum. Right column: Discrimination performance on stimulus pairs symmetric with respect to λ=0.5 for increasing values of λ distance.

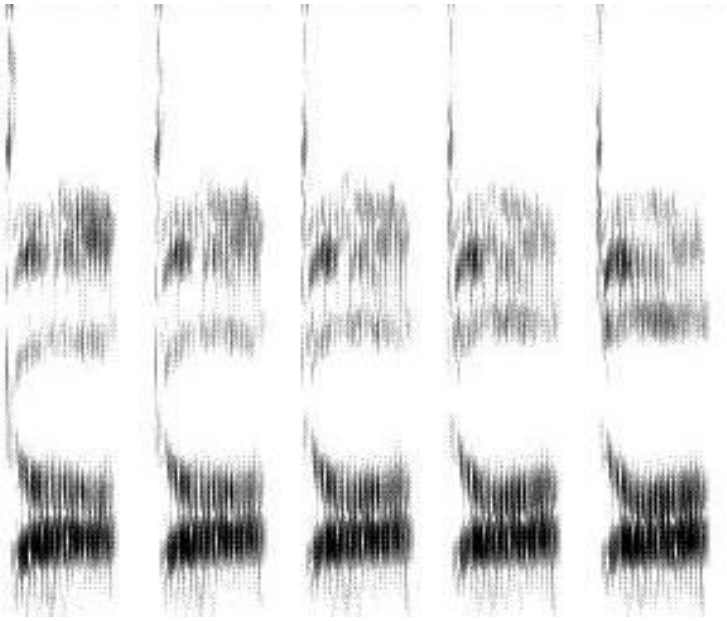
Figure 6. Identification and discrimination curves for 3 Japanese speakers with the resynthesized stimuli along the [ra]--[la] continuum for two stimulus voices (male: squares on dashes; female: circles on dots). Each row shows data from a single subject. Left column: Identification (labeling) performance on resynthesized stimuli for λ (see Equation 6) between -0.7 and 1.7 in 0.1 steps. Middle column: Discrimination performance (i.e., hits−false alarms over total number of trials) on stimulus pairs 0.3 units apart (in λ units) along the continuum. Right column: Discrimination performance on stimulus pairs symmetric with respect to λ=0.5 for increasing values of λ distance.

Modified LPC resynthesis, Figure 1

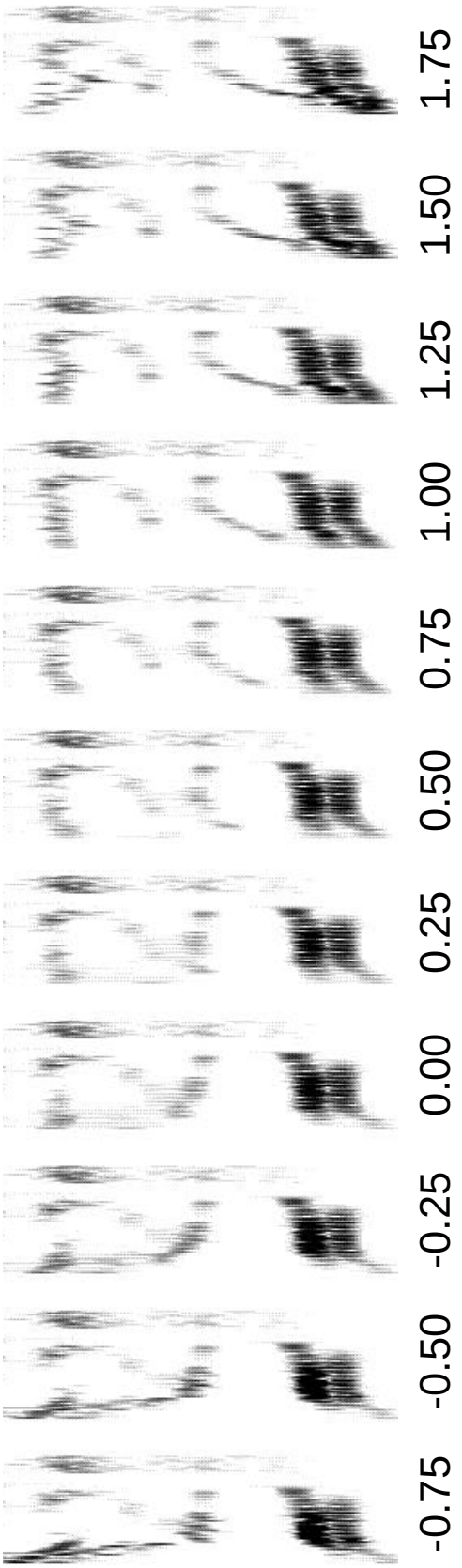


Modified LPC resynthesis, Figure 2

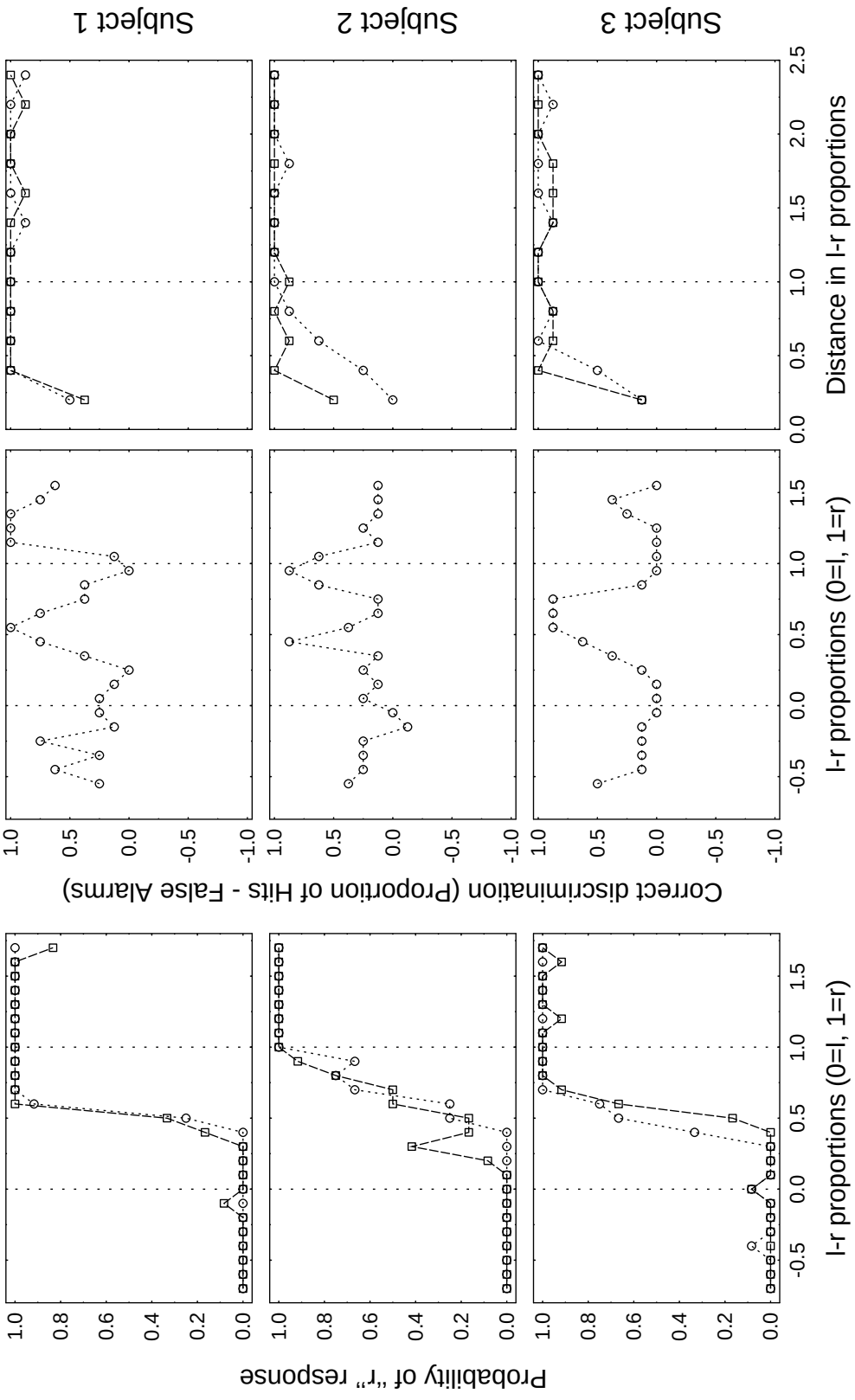



0.00
0.25
0.50
0.75
1.00

Modified LPC resynthesis, Figure 4



Modified LPC resynthesis, Figure 5



Modified LPC resynthesis, Figure 6