BOOK REVIEWS

Speech Perception, Production and Linguistic Structure. Edited by Yoh'ichi Tohkura, Eric Vatikiotis-Bateson, and Yoshinori Sagisaka. Tokyo: Ohmsha, and Amsterdam: IOS Press, 1992. 463 pp. $115

Reviewed by
BRUNO H. REPP*
Haskins Laboratories

This attractive volume is the result of a workshop held at ATR (Advanced Telecommunications Research) Laboratories, Kyoto, in November 1990, immediately preceding the International Conference on Spoken Language Processing in Kobe. The contributions are arranged into two major sections (Speech Perception; Speech Production and Linguistic Structure), each of which has three subsections. Each of these contains a number of articles followed (with one exception) by several shorter commentaries; the exact distribution is: 5+2, 2+2, 5+4; 2, 3+3, 5+2. Thus there are 22 articles and 13 commentaries in all. Nearly all authors are well-established researchers, roughly one third of them Japanese, the majority from the United States, plus a few European representatives.

The first subsection deals with the rather specific topic of "Contextual Effects in Vowel Perception", with papers by Sumi Shigeno, Robert Allen Fox, Caroline B. Huang, and Masato Akagi, followed by comments from Dominic W. Massaro and Sieb G. Nooteboom. Context effects are not only interesting from a general psychophysical perspective but constitute one of the central problems faced by automatic speech recognition. The contributions by Huang and Akagi represent this latter perspective, while Shigeno and Fox are primarily concerned with modelling human perception. **Shigeno**'s study (a slightly condensed version of Shigeno, 1991) represents a continuation of her earlier careful work on contextual effects in the perception of isolated vowels as a function of category membership, spectral distance, and temporal proximity. The most salient result is that successive vowels show contrastive effects as long as they belong to different categories, but assimilative effects when they belong to the same category. In agreement with dual-process models of categorical perception, Shigeno suggests that assimilation occurs in auditory memory, whereas contrast arises from categorical representations. **Fox** reports four experiments aimed at demonstrating that phonotactic regularities of the language constrain vowel perception in syllabic contexts. He compares the identification of vowels in open syllables and in the context of a following /r/, which

* I am grateful for the support of Haskins Laboratories during the preparation of this review, and to Yoh'ichi Tohkura for helpful comments on a first draft.

Address correspondence to Bruno H. Repp, Haskins Laboratories, 270 Crown Street,
Hew Haven, CT 06511–6695.

neutralizes the tense–lax contrast in English. The first experiment, using a synthetic /ɪ–ɛ–æ/ continuum in the two contexts, shows a weakening of the perceptual /ɛ–æ/ distinction preceding /r/. The second experiment, a multidimensional scaling analysis of similarity judgments on a larger set of naturally produced syllables, supposedly also shows a reduction of the perceptual distance between tense and lax vowels in the /r/ context, although I have difficulty seeing this; the /r/ context just seems to rotate the perceptual space, but not to affect the distances. Experiment 3, however, confirms a deleterious effect of following /r/ on the identification of natural vowels, and Experiment 4 shows that a following /l/ (which permits tense–lax contrasts) does not have a similar effect. Fox argues that the observed phonotactic effect is perceptual rather than a response bias, though the basis for that claim remains to be elucidated. For one thing, perceptual effects can take the form of a bias (e.g., Repp, Frost, and Zsiga, 1992).

Huang is concerned with the information conveyed by the dynamic formant trajectories within (monophthongal) vowels in CVC contexts. She first shows that an automatic classification algorithm gains little if the trajectories are used to infer single formant target values; however, a substantial improvement results when three points from each vowel's trajectories are presented to the algorithm. Further improvements result when information about duration and consonantal context is added. These data are then compared to those of human listeners who identified the excised vowel nuclei or the full CVC syllables. Though it is not quite clear how degree of agreement between machine and human performance was determined, it seems to be highest for the three-point condition, which confirms the perceptual importance of formant trajectories. Of course, this is hardly a new insight (see, e.g., Nearey and Assmann, 1986), but it is good to see it put to use in the context of automatic speech recognition. **Akagi**, in the subsequent paper, presents a detailed and elegant study of contextual effects among successive vowel-like stimuli, based on a dual-process model of interaction, both at the level of individual spectral peaks and at the level of phoneme boundaries. The model seems similar to Shigeno's, though Akagi makes no reference to it or to any other related work. The results, obtained with vowels preceded by either single-formant stimuli or full vowels at various temporal intervals, provide a detailed map of assimilation and contrast effects. The results for full vowel contexts resemble those of Shigeno, showing assimilation when spectral distance is small and contrast when it is large, with little effect of temporal separation. For single-formant stimuli, however, temporal factors seem to be most important, with assimilation at separations of less than 70 msec, and contrast beyond. Akagi's study is very clever and the data are informative, but it must be said that they were obtained by presenting two female subjects with more than 100,000 (!) stimuli, resulting from a large factorial design, each stimulus to be identified as either /u/ or /a/. The necessity and indeed the humanity of such an excessive design must be seriously questioned.

The two commentaries on the preceding four articles are brief. **Massaro** predictably takes this opportunity to recite the canons of his "fuzzy logical model" of speech perception. Moreover, he admonishes the authors to "keep the big picture in mind in their day-to-day struggle with the wonders of speech perception" and to design their experiments according to the tenets of his own "paradigm for speech perception research", even though this factorial paradigm has rarely yielded data of the richness of any of the four studies commented on. **Nooteboom**'s ideas are more constructive and to the point. He argues that categorical perception explains the assimilation/contrast findings of Shigeno and Akagi, and that temporal proximity is likely to enhance these effects by pulling speech stimuli into a single stream of sounds. Within-stream contrast serves to enhance phoneme boundaries.

The second subsection is devoted to "Perceptual Normalization of Talker Differences" and offers articles by Tatsuya Hirahara and Hiroaki Kato, by Howard Nusbaum and Todd Morin, and by Kazuhiko Kakehi, followed by a commentary by David Pisoni. The topic, like that of the preceding subsection, is one of signal importance to both human and machine speech recognition, and the work reported is very interesting. **Hirahara and Kato** provide another example of the meticulous and wide-ranging parametric studies that Japanese researchers seem to excel in. By presenting an array of 200 isolated vowels varying in formant frequencies and, independently, in fundamental frequency (F_0) for identification to 24 listeners, they mapped out the perceptual space and demonstrated shifts in category boundaries in F_1–F_2 space caused by F_0. These shifts were much reduced, however, when F_0 (in Bark) was subtracted from each coordinate of the vowel space, which provided an effective normalizing procedure. In other words, vowel quality remained constant as long as F_0 was shifted along with the other formants on a Bark scale. The authors argue that low harmonics of a relatively high F_0 may act as a "perceptual formant" because they are not subject to auditory integration; this would account for changes in perceived vowel quality when F_0 alone is changed. An intriguing difference in the Japanese /a/–/o/ boundary between male and female listeners is also reported.

Nusbaum and Morin, in the subsequent paper, describe five experiments using a common paradigm: the comparison between single (blocked) and multiple (mixed) talker conditions. For several types of stimuli (isolated vowels, consonants and vowels in CV syllables, monosyllabic words) they find slightly reduced identification accuracy and particularly longer reaction times in the multiple-talker condition, which they take as evidence of a normalization process. In their most interesting manipulation, they added a secondary task (either one or three numbers to hold in memory) and showed that memory load further slows reaction times in the mixed, but not in the blocked condition. This indicates that the normalization process requires cognitive resources that the memory task competes for. Nusbaum and Morin further show that eliminating F_0 by using whispered stimuli reduces accuracy in the mixed condition only, and that mixing similar talkers (of the same sex) has little detrimental effect (though it is not clear whether the talkers could actually be discriminated). The authors distinguish two kinds of normalization processes: "contextual tuning" and "structural estimation"; the first seems to operate in the blocked, and the second in the mixed, condition. This is an interesting set of experiments, although the tasks are somewhat artificial, some methodological information is missing, and at least one important earlier study (Summerfield and Haggard, 1975) is not mentioned.

Kakehi's paper is rather brief and evidently a summary of a study published previously in Japan. Using a large set of Japanese syllables, he examined the time it takes to adapt to a single speaker (about four trials), as well as the time to lose that adaptation (about seven trials, though there is a significant irregularity in the data that the author glosses over). There were also talker-specific variations. **Pisoni** says little about the preceding papers but rather reports some relevant research from his own laboratory, showing effects of talker variation on reaction time and memory performance. His important conclusion is that talker-specific information is not "stripped off" but is retained in memory along with the phonetic structure of an utterance. This observation raises the question (in my mind, at least) of whether the normalization process referred to by Nusbaum and Morin and many others is really a process at all, in the sense that it changes mental representations of speech, or whether it is simply a reflection of cognitive uncertainty caused by stimulus variability.

The third and most substantial subsection deals with the "Perception and Learning of Non-Native Language". It includes articles by Reiko A. Yamada and Yoh'ichi Tohkura, by Scott E. Lively, David B. Pisoni, and John S. Logan, by Winifred Strange, by Jacques Mehler and Anne Christophe, and by Patricia K. Kuhl, followed by commentaries by Howard C. Nusbaum and Lisa Lee, by Anne Cutler, by Shigeru Kiritani, Fumi Katoh, Akiko Hayashi, and Toshisada Deguchi, and by Morio Kohno. The first two papers, and the third in part, deal with the narrow problem of Japanese speakers' difficulties in distinguishing the American English phonemes /ɪ/ and /l/, which has attracted an unusual amount of research effort, much of it using synthetic /ɪ/–/l/ continua of the kind developed at Haskins Laboratories. **Yamada and Tohkura**, in their presentation (which overlaps substantially with Yamada and Tohkura, 1992), report data showing that Japanese listeners perceive /w/ in the boundary region of /ɪ/–/l/ continua, that their identification accuracy for synthetic and natural stimuli is correlated, that they are sensitive to stimulus range, and that (unlike native speakers of English) they seem to use second-formant transition information to distinguish English /ɪ/ and /l/. The most interesting and extensive part of their study compared 120 Japanese subjects who had never lived abroad with 122 Japanese who had lived in the U.S. for some time. Not only did the second group outperform the first in /ɪ/–/l/ identification accuracy, but there was a clear relationship between the age at which subjects had moved to the U.S. and their accuracy: Virtually all the high performers had moved before the age of 11. Yamada and Tohkura's work, by the way, shows something that American studies tend to downplay: There are quite a few speakers of Japanese who can discriminate English /ɪ/ and /l/ perfectly well.

Lively, Pisoni, and Logan summarize various attempts to train Japanese speakers in the laboratory to improve their discrimination of /ɪ/ and /l/. After criticizing earlier approaches that used synthetic speech and discrimination tasks, they report the results of their own studies using natural speech, a variety of utterances, and multiple speakers in a categorization task (partially reported also in Logan, Lively, and Pisoni, 1991). Over 15 sessions, the subjects' performance improved significantly but not impressively (between 6% and 9%). In view of considerable variation in accuracy for different speakers' tokens in the training phase, the results of post-training generalization tests, in which utterances from a single novel speaker were presented, seem uninterpretable. The test should have included multiple novel talkers, or at least the single novel talker should have been rotated with the talkers used in training in a counterbalanced design. Still, the training methods of Lively *et al.* seem intuitively reasonable, and their data provide useful information about contextual variation and speaker differences in /ɪ/–/l/

discriminability. They also relate their approach to work on exemplar-based storage models of memory (e.g., Hintzman, 1986). **Strange**, in the following article, reviews more broadly the methodological variables relevant to training studies and points to a crucial factor (surprisingly neglected in the two preceding papers), viz., the relation of the non-native categories to be discriminated to the phonetic categories of the native language. After reporting some (previously unpublished) evidence that the nature of the training task makes little difference in Japanese listeners' discrimination of the /r/–/l/ distinction (though all training was done with synthetic speech), she surveys several experiments that involved not only /r/ and /l/, but a number of other phonetic distinctions, including some that are difficult to discriminate by native speakers of English (a welcome relief from the heavy focus on what "others" cannot do). These data provide some interesting glimpses of effects of phonetic context and of individual differences among subjects with regard to trainability – a dimension almost totally ignored in this type of research, but highly relevant to the natural task of second-language acquisition. It is fair to conclude, however, that there is so far remarkably little success in training subjects in the laboratory to discriminate non-native categories, a skill that seems to be very difficult for most adult subjects to acquire. It would be interesting to conduct such training studies with children who should be more malleable in that regard, though not necessarily more responsive to boring laboratory training procedures. Perhaps, if the training procedures were placed in a motivating social context, they would meet ith greater success.

The remaining two articles in this section focus not on the perception of non-native categories (except tangentially) but on the acquisition of the native phonology and phonetics; therefore, the subsection should really have been given the heading "Acquisition and Perception of Native and Non-Native Language". **Mehler and Christophe** focus on the question of the units of speech perception. They first review recent research by Mehler (in collaboration with Cutler, Norris, and Segui) who used a syllable monitoring task to demonstrate that speakers of French employ a syllabic segmentation strategy, whereas speakers of English do not. A study in which the two languages were interchanged and another study with true bilinguals suggested that speakers have only one processing strategy that they apply to native (or dominant) and non-native (or non-dominant) languages alike, though apparently a strategy can be "inhibited" as well. Recent work is cited which extends the paradigm to Spanish and Catalan, with intermediate results: These speakers seem to have a syllabification strategy at their disposal, but seem to employ it only in certain contexts or under certain conditions. The authors conclude that "speech processing procedures depend on the maternal language of the speaker", and I find it surprising that they had seriously considered an alternative hypothesis. In the second part of their paper, Mehler and Christophe discuss the important problem of speech unit acquisition in young infants, on which Mehler and his group have conducted pioneering research with neonates. They cite several studies in progress, which suggest that "infants do not perceive speech as a string of phonemes, but in terms of some higher-order unit(s)". One recent study seems to suggest that infants are sensitive to subtle word boundary cues in fluent speech. Interesting and important as this research is, I cannot avoid the feeling that Mehler and his colleagues are conceptually enslaved by the idea of (phonological) units. A more fluid model of perceptual differentiation might be appropriate, in which the units are not pre-ordained by phonological theories but emerge autonomously from patterns of repetition and statistical frequency in the input. Such a view, incidentally, also entails that "processing" strategies will be language-specific, to the degree that languages are different from each other.

Kuhl, in the final article in this subsection, summarizes in somewhat schoolmasterly fashion her research on vowel category "prototypes" in human adults, six-months old infants, and macaque monkeys (see also Kuhl, 1991). Both variants of the human species are shown to have a prototype or best exemplar for the /i/ category, whereas monkeys do not. The evidence comes from a discrimination task which shows that, if a prototype exists, discrimination is more difficult in its vicinity than at some distance from it in the acoustic vowel space. This is interesting research, but it is still rather limited in its restriction to a single category of isolated vowels. More recent cross-linguistic work on English and Swedish infants is, unfortunately, only mentioned very briefly; unlike Mehler and Christophe, Kuhl seems unwilling to share preliminary results of work in progress. The most interesting part of the results is that for the infants who, by the age of six months, seem to have acquired a notion of what a good /i/ sounds like. Kuhl seriously considers the possibility that some prototypes might be innate (a hypothesis that seems absurd to me) but ultimately favors an explanation in terms of exposure to speech. Monkeys, of course, are at a serious disadvantage because they have not been exposed to human speech in the same way, nor do they presumably care much about human vowels. They may well have a prototype for some conspecific vocalization that has communicative significance for them.

Of the four commentaries in this subsection, only two turn out to be relevant to the topic. **Nusbaum and Lee** go on for too long about too little, ending up with the suggestion that perceptual learning can be understood as a reshaping of the distribution of attention over the speech signal, which seems little more than a restatement of the problem. **Cutler**'s very brief commentary reiterates the unsurprising conclusion (*cf.* Mehler and Christophe) that native phonology constrains native as well as non-native language processing, and then mentions a recent paradigm-bound finding of longer phoneme monitoring times for vowels than for consonants. The remaining two "commentaries" are actually short reports of research that seem to have found shelter in the wrong place. The paper by **Kiritani** *et al.* deals with perceptual normalization of vowels in children and infants and clearly belongs in the preceding subsection. The paper by **Kohno**, while quite intriguing, does not really fit anywhere in this volume. Following up an important but rarely cited study by Hibi (1983), he presents further evidence of a qualitative change in the processing of rhythmic sequences at a rate of about three per second, which presumably reflects a shift from a nonintegrative to an integrative processing of successive auditory events.

Part II of the book begins with two articles by John J. Ohala and Hiroya Fujisaki, respectively, that are assigned to a separate subsection entitled "Introduction", though neither serves a truly introductory function with respect to the following papers. **Ohala**, in characteristically enlightening fashion, discusses the difficulty of distinguishing phonetic variation due to active causes (style of speech, coarticulation, etc.) from similar variation that, though it may originally have been caused by the same factors, has become phonologized and part of the language norm. He summarizes the methodology of historical linguistics and points out its limitations, which include the inability to establish the causal basis of sound change. He goes on to cite some empirical research that begins to address this question. To cite just one example, Solé (1992) – originally in collaboration with Ohala – has presented rather strong evidence that anticipatory nasalization of vowels is part of the language norm (i.e., a sound change) in English, but not in Spanish and several other languages, where it is merely a passive coarticulation effect. **Fujisaki** presents the latest version of his model for generating Japanese F_0 contours, which has been influential for over two decades, even though it was originally described (like almost all of Fujisaki's brilliant work) only in technical reports and conference proceedings. The model has two basic components: phrase commands that generate impulse-like changes in F_0, and accent commands that generate stepwise changes. A model of the underlying physiological mechanism is also briefly discussed, which postulates the strain of the vocal cord as the principal variable being controlled.

The following subsection, entitled "Articulatory Studies", contains articles by Kevin G. Munhall, J. Randall Flanagan, and David J. Ostry, by Eric Vatikiotis-Bateson and Janet Fletcher, and by Mary E. Beckman and Jan Edwards, followed by three brief commentaries, respectively by Osamu Fujimura, René Collier, and Kiyoshi Honda. To anticipate, two of these commentaries are in fact on Fujisaki's presentation: **Collier** argues in favor of modelling the perceptually significant aspects of intonation, rather than the raw F_0 contour of the acoustic signal. However, to the extent that a model actually succeeds in closely approximating the F_0 contour (as Fujisaki's model seems to do), Collier's proposal seems superfluous. **Honda** adds some interesting physiological observations on the active control of F_0 lowering. **Fujimura**'s comments, unfortunately, provide only an obscure promise of some global model to come. Also, his contribution evidently was not edited or updated since the 1990 workshop, as his introductory paragraph does not jibe with the contents of the present volume.

The three articulatory studies, largely by-passed by the commentators, are actually quite interesting, though they suffer from a problem endemic to speech production research: insufficient numbers of subjects and large inter-subject variability. **Munhall** *et al.* largely escape that criticism by restricting themselves to mere examples of data in the context of theoretical observations. They sketch their approach to articulatory modelling, which is based on the notion of a kinematic space whose coordinates are joint angles or displacements, rather than effector movements in Cartesian space. Preliminary observations suggest simple linear relationships among the coordinate variables, which represent built-in constraints on articulatory trajectory formation. This elegant and promising approach contrasts with the somewhat undisciplined presentation by **Vatikiotis-Bateson and Fletcher**, who review a profusion of extremely complex and variable data that are difficult to make head or tail of. Their claim is very interesting and important: that local changes in a phrase may affect articulatory patterns throughout the whole utterance. However, it is not clear whether they have solid evidence to support their claim, or whether they are dealing with variability pure and simple. **Beckman and Edwards** tell a more coherent story, although they seem guilty of rather casual and selective reporting of data, possibly focusing on those most favorable to their conclusions. Their theoretical framework is the task-dynamic model developed at Haskins Laboratories, and they report results (in part already presented by Edwards, Beckman, and Fletcher, 1991) that illustrate how the dynamic control variables of stiffness, amplitude, and phasing are differentially employed to produce different types of stress and lengthening phenomena in (painfully stilted) laboratory-style utterances. It is to be hoped that the extremely promising theoretical ideas of these authors and those of the two preceding papers (each group including a former associate of Haskins Laboratories) will soon be supported by sufficiently extensive data.

The last subsection, "Acoustic Studies", includes papers by Nobuyoshi Kaiki and Yoshinori Sagisaka, by Nick Campbell, by Anne Cutler, by Jacques Terken and René Collier, and by Sieb G. Nooteboom and Wieke Eefting, with commentaries by Yoshinori Sagisaka and by Mary E. Beckman. **Kaiki and Sagisaka** report a statistical analysis of segment durations in a sizable corpus of Japanese speech, produced by a single speaker (a professional announcer). Somewhat confusingly, the technique is described as "categorical factor analysis", though in fact it seems to be a version of linear regression analysis. Many independent variables were considered, including position in utterance groups of various sizes, length of utterance group, part of speech, etc. Still, the full regression model derived from one half of the data did not predict the segment durations in the other half as accurately as one would like. One of the more surprising detailed results is a substantial shortening of vowels in sentence-final position. **Campbell**'s subsequent discussion, based on the same data base, reveals this effect as being due entirely to the prevalence of the Japanese past-tense particle /-ta/, which always occurs in sentence-final position. Apart from arguing convincingly for the necessity of detailed linguistic analysis of speech corpora, Campbell mainly demonstrates the utility of z-scores as a way to eliminate differences in intrinsic segmental duration from cross-segment comparisons. **Cutler**, in the following paper, briefly summarizes some of her research on word boundary cues, referring the reader to papers published elsewhere for details. Her work suggests that speakers of English expect words to start with strong (unreduced) syllables, in agreement with the statistical predominance of such words in the language. When induced to speak very clearly, speakers may introduce subtle durational cues to word boundaries, mainly pre-boundary lengthening or pausing. One limitation of this research (which it shares with Mehler's) is its adherence to binary theoretical concepts such as strong/weak, where in fact syllables probably vary in "strength" depending on segmental composition and context. Also, her (admittedly abbreviated) discussion does not convey the richness of possible lexical entries (ranging from individual letter names to words, compounds, familiar phrases, and memorized texts) and the somewhat arbitrary distinctions imposed by the orthography between what counts as two words and what counts as one. The "word boundary problem" may in fact be that of prosodic grouping in general.

Terken and Collier examine the influence of syntactic structure on prosody in a Dutch corpus obtained (like the Japanese corpus mentioned above) from a single professional speaker. They find that a major syntactic boundary (NP–VP) is marked by pitch inflection, pausing, and lengthening, whereas lesser syntactic boundaries are marked by pitch inflections only, if they are marked at all. The material is limited in structural variety, however, and the authors appropriately characterize their work as a pilot study within a general effort to improve the naturalness of a text-to-speech system. In the final article, **Nooteboom and Eefting** examine how a speaker adjusts prosody to meet a listener's needs. They report a study that demonstrates that the lengthening of "new" relative to "given" lexical items is a concomitant of the presence *vs.* absence of pitch accent (in Dutch). They also show (in a study reported in more detail in Eefting, 1992) that the judged naturalness of speech suffers when the durational characteristics are not in agreement with accentuation. **Sagisaka**, in the first of the two commentaries, surprisingly does not respond to Campbell's critical examination of his own results but instead comments on long range control of timing and on differences in prosodic organization between Japanese and English. **Beckman** concludes with some perceptive comments on rhythm in different languages, but surprisingly refers to the "three papers" in this section and to a paper by Fant *et al.*, which is not contained in this volume. Apparently, this commentary was not updated to reflect the final contents of the book.

In summary, despite some peculiarities of organization, this is a generally well-edited and consistently interesting collection of articles, some of which present original findings not available elsewhere. The "commentaries" are less successful, on the whole, and could have been omitted without much damage. The book would be a worthwhile addition to the library of anyone interested in the contemporary speech scene, but many potential buyers will find the price tag prohibitive.

(Received August 4, 1992)

REFERENCES

EDWARDS, J., BECKMAN, M.E., and FLETCHER, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, **89**, 369–382.

EEFTING, W. (1992). The effect of accentuation and word duration on the naturalness of speech. *Journal of the Acoustical Society of America*, **91**, 411–420.

HIBI, S. (1983). Rhythm perception in repetitive sound sequence. *Journal of the Acoustical Society of Japan*, **4**, 83–95.

HINTZMAN, D.L. (1986). "Schema abstraction" in a multiple trace memory model. *Psychological Review*, **93**, 411–428.

KUHL, P.K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, **50**, 93–107.

LOGAN, J.S., LIVELY, S.E., and PISONI, D.B. (1991). Training Japanese listeners to identify /r/ and /l/. *Journal of the Acoustical Society of America*, **89**, 874–886.

NEAREY, T., and ASSMANN, P. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, **80**, 1297–1308.

REPP, B.H., FROST, R., and ZSIGA, E. (1992). Lexical mediation between sight and sound in speech-reading. *Quarterly Journal of Experimental Psychology*, **45A**, 1–20.

SHIGENO, S. (1991). Assimilation and contrast in the phonetic perception of vowels. *Journal of the Acoustical Society of America*, **90**, 103–111.

SOLÉ, M.-J. (1992). Phonetic and phonological processes: The case of nasalization. *Language and Speech*, **35**, 29–43.

SUMMERFIELD, A.Q., and HAGGARD, M.P. (1975). Vocal tract normalization as demonstrated by reaction times. In G. Fant and M.A.A. Tatham (eds.), *Auditory Analysis and Perception of Speech* (pp. 115–142). London: Academic Press.

YAMADA, R.A., and TOHKURA, Y. (1992). The effects of experimental variables on the perception of American English /r, l/ by Japanese listeners. *Perception & Psychophysics*, **52**, 376–392.