8

Speech Perception without Speaker Normalization

An Exemplar Model

KEITH JOHNSON

8.1 INTRODUCTION

Speaker normalization is a hypothesized perceptual process in which differences between speakers are reduced prior to identification of linguistic categories. There are a variety of conceptions of speaker normalization, all of which in one way or another involve a mapping from a speaker-specific representation to a relatively speaker-neutral abstraction, which is presumably an appropriate probe to linguistic memory.

For example, in Gerstman's (1968) range normalization, formant values are expressed relative to the speaker's range of produced formants. So, if a vowel has a first formant (F1) value of 500 Hz and the speaker's F1 in other vowels ranges from 300 Hz to 700 Hz, the normalized value of F1 is 0.5 because 500 is halfway between 300 and 700.

In the Joos (1948)/Potter and Steinberg (1950) approach, the formant values of a vowel are considered relative to each other. So, if a vowel has a second formant (F2) value of 1500 Hz and F1 of 500, one dimension of the normalized representation is the difference between them. In Syrdal and Gopal's (1986) and Traunmüller's (1981) implementations the differences are calculated after transforming the formant values to the Bark scale. J. D. Miller (1989) and Nearey (1978, 1989) calculated the differences of the log formant values. Bladon, Henton,

Talker Variability in Speech Processing

and Pickering (1984) implemented this approach by sliding auditory spectra up or down (depending on the sex of the speaker) on the frequency scale.

What all of these models share is the basic property that the auditory representation of the speech signal must be modified in some way prior to recognition.

Gesture recovery models (Fowler, 1986; Liberman & Mattingly, 1985) appear at first to be very different from these acoustic normalization models. Although gesture recovery is principally concerned with the problem of contextual variation in speech (the flip side of speaker variation in the lack of invariance problem, see Perkell & Klatt, 1986), it can also be seen as a type of speaker normalization. This is because gestures are abstract speaking intentions, not actual articulator movements. With an understanding of gesture as a relatively speaker-neutral abstraction, the process of gesture recovery is by definition a speaker normalization process of the same sort as those mentioned in the preceding paragraphs.[1]

In this chapter I will outline a model of speech perception that unlike these approaches includes no speaker normalization process. The next section describes the model in rough conceptual terms. In the third section I discuss some elaborations of the basic model that address points of plausibility and implementation. The fourth section is a description of a small working exemplar model of vowel recognition and results of some simulations.

8.2 PERCEPTION BY EXEMPLARS

In exemplar models of perception (Estes, 1993; Hintzman, 1986; Nosofsky, 1986, 1988, 1991; Nosofsky, Kruschke, & McKinley, 1992) a perceptual category is defined as the set of all experienced instances of the category. That is, no abstract category prototypes are posited. The process of categorization then involves comparing the to-be-categorized item with each of the remembered instances of each category, and categorization is based on sums of similarity over each category. Hintzman (1986) demonstrated that a model of this sort behaves as if categorization is based upon category prototypes, although category abstraction is produced at decision time rather than during acquisition.

A pure exemplar model is obviously impossible because it is necessary to assume that the perceiver remembers too much (Neal Johnson, personal communication, calls this the "head-filling-up problem"). This will be discussed further in the next section, but at this point it should be mentioned that this apparent need

[1] I am not arguing against Fowler's (1986) characterization of speech perception as "direct" gesture recovery. She emphasizes that her Gibsonian brand of gesture recovery is nontranslational, in the sense that the gestural intentions of the speaker are directly perceived by the listener (or better, perceiver). The mechanism involved in deriving gestures from an acoustic signal, although a huge black-box in the theory, is not at issue here. I am suggesting that recovery of linguistic intentions from speech articulation involves some abstraction because gestural intentions vary from speaker to speaker for the same linguistic entities (Johnson, Ladefoged, & Lindau, 1993).

for unlimited memory is matched by the apparent availability of unlimited memory in picture recognition. Standing, Conezio, and Haber (1970) found that people could recognize thousands of previously seen pictures with surprising accuracy and over surprisingly long times. Goldinger (Chap. 3, this volume; 1992) also found that implicit memory for (instances of) words is strong and long-lasting (see also Palmeri, Goldinger, & Pisoni, 1993; Schacter & Church, in press). So, although an exemplar model seems to need unrealistic amounts of memory, people have a surprising ability to remember instances.

Figure 1 illustrates categorization in an exemplar model of speech perception. In this illustration, an exemplar is an association between a set of auditory properties and a set of category labels. The auditory properties are output from the peripheral auditory system, and the set of category labels includes any classification that may be important to the perceiver, and which was available at the time that the exemplar was stored—for example, the linguistic value of the exemplar, the gender of the speaker, the name of the speaker, and so on. The association between sound and category is indicated in the figure as an oval labeled "exemplar," with lines stretching off in one direction toward a vector of auditory properties and in another direction toward a vector of category labels. The stack of ovals stands for the set of all exemplars. Given an item to be categorized, its auditory properties are compared with each exemplar's auditory properties, and the similarity between the item and each exemplar determines the activation level of the exemplar. If the match is good, the activation level of the exemplar is high. The sum of activations over all of the exemplars of a category is taken as evidence that the unknown sound should be categorized as an instance of that category. This is true for each of the types of categories, and so in this way, the model performs speech and speaker recognition simultaneously, as do humans (Remez, Fellowes, & Rubin, 1995).

This type of model can be implemented as a set of formulas (following Nosofsky, 1988). Auditory similarity s_{ij} between exemplars i and j is calculated by (1), where the auditory property m of exemplar j is written x_{jm}, the Euclidian distance between exemplar j and item i is written d_{ij}, w_m is an attention weight given to property m, and c is a sensitivity constant.

$$d_{ij} = [\Sigma w_m (x_{im} - x_{jm})^2]^{1/2} \tag{1a}$$

$$s_{ij} = \exp(-c d_{ij}) \tag{1b}$$

The degree to which item i activates exemplar j in memory is calculated from similarity using formula (2) by assuming that each exemplar has a base activation level (N_j) and optionally that Gaussian noise is added.

$$a_{ij} = N_j s_{ij} + e_j \tag{2}$$

Evidence for category C_1 given item i is then the sum of the activations of exemplars j of C_1 as shown in formula (3).

$$E_{1,i} = \Sigma a_{ij}, j \in C_1 \tag{3}$$

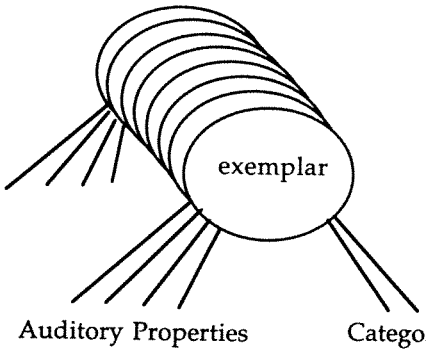

FIGURE 1 A set of exemplars relating auditory properties to category labels.

The weight parameters w_m allow the model to ignore variation on certain stimulus dimensions (and hence are called *attention weights*). As we will see in section 8.4, certain speaker normalization effects can be modeled with changes in the attention weights.

The sensitivity parameter c serves to limit the impact of distant exemplars. Because the function relating distance to similarity is exponential, the impact of distant neighbors on the calculation of total activation E can be reduced to almost nothing. So, the similarity function provides a sort of K nearest-neighbors classification, in which only nearby neighbors are considered.

The base activation levels N_j may vary as a function of some experimental manipulations. For example, a forced-choice identification task can be simulated by forcing the base activation levels of the available responses to be 1 and those of all other categories to be 0. This move forces the identification response to be one of the selected responses because the activations for the other categories are 0 (or, if e_j is used, hover around 0).

8.2.1 Compensation for Speaker Variability in an Exemplar Model

In this model of speech perception, the auditory properties that distinguish speakers are retained in the exemplars. The result of this is that the exemplars that are most similar to a to-be-categorized item are those exemplars that were spoken by the same or a similar speaker. By retaining speaker-specific information in the long-term category representation (the set of exemplars), the model makes it possible to categorize new items by reference to appropriate prior examples—a subset of exemplars that resemble the to-be-recognized item on speaker-specific dimensions.

In Johnson (1990a) I argued for an indirect model of speaker normalization in which cues for vowel identity are evaluated relative to the perceived identity of the speaker. The type of model that I envisioned was one in which the perceived identity of the speaker established (or guided the selection of) a frame of reference for the evaluation of linguistic cues (see also Nearey's [1978] sliding template

model of vowel perception, and Whalen and Sheffert's [Chap. 7, this volume] discussion of speaker model construction). The exemplar model outlined here is an indirect model because categorization takes place by reference to items in memory that retain speaker information. That is, the frame of reference (a model of the speaker) is inherent in the set of exemplars, and the similarity calculation (formula 1, above) limits the comparison to items in memory that are sufficiently close to the to-be-categorized item.

8.2.2 Attention Weights

The model parameters w_m shrink or expand the perceptual space along each of the auditory dimensions. These parameters are called *attention weights* (Nosofsky, 1986) because they control the degree to which the categorization process is sensitive to particular auditory properties. For example, if a particular experimental task calls upon the listener to categorize a set of vowel stimuli (to take a relevant example) primarily upon the basis of their first formant frequencies, it is reasonable to expect that the listener will tune in to F1. This can be modeled by letting the weight for F1 (or the weights for the critical bands in the F1 region) be larger than the weights for other auditory properties.[2]

Of course, attention weights can be set so that speaker cues like fundamental frequency are ignored. In vowel perception this would lead to the vowel confusions that you might expect if perception depended solely on location in the F1/F2 plane (for example). However, in modeling the behavior of listeners, I have not found any situations in which this happens. That is, although the structure of the exemplar model makes it possible to ignore a speaker-specific dimension like F0 by setting its weight to 0, in practice this does not happen.

8.2.3 Base Activation Level N_j

Ganong (1980) observed that phonetic categorization is influenced by word frequency. The effect that he observed is reminiscent of the perceptual magnet effect (Kuhl, 1991) in that high-frequency words tended to attract listener's responses.[3] One way to account for this behavior (if the raw number of exemplars does not; see McQueen, 1991) is to assume that the activation parameters N_j tend

[2] I am being purposefully vague about several issues in this chapter. One of them is the definition of *auditory property*. At some points I consider formant values to be auditory properties and at other points I consider auditory properties to be critical-band activation levels, or even vector quantized spectral templates. In the simulations presented in section 8.4 I use measured formants, F0, and durations, but in earlier work (Johnson, 1990b) I used auditory-based spectra. For vowels, these seem to be interchangeable (Fant, 1960), but a more general model of speech perception will probably have to be based on spectra.

[3] Thinking about similarities between the Ganong effect and the perceptual magnet effect suggests an exemplar-based explanation of the perceptual magnet effect (assuming, as argued in the text, that word-frequency effects are related to exemplar coding). This topic is beyond the scope of this chapter.

to be larger for higher frequency words. Nosofsky et al. (1992) proposed that base activation level is subject to a time function such that recent exemplars have higher base activation than do past exemplars. On average then, more frequent words will have a greater number of recent exemplars and therefore higher aggregate base activation levels.

The base activation level parameter may also be a route for allowing higher level processing to have an influence on speech perception. If syntactic or semantic context leads to the prediction that a particular set of words is likely to occur, the base activation levels of the exemplars of these words could be increased. Manipulation of base activation in this way changes the recognition process directly, as opposed to the use of a language model as a filter on the output of recognition, as is done in automatic speech recognition. Whether this is a desirable property is open to debate (see the discussion in Norris, 1994).

In some of the simulations described below, I assumed that in the forced-choice paradigm listeners increase the base activations of exemplars of the permissible responses and decrease (to zero) the base activations of all other exemplars. Obviously, real listeners do not absolutely rule out impermissible alternatives, but this simplification made the models easier to implement.

Base activation level may also be manipulated to simulate a couple of types of context effects that occur in speech perception experiments. Contrast effects observed in selective adaptation and anchoring paradigms that seem to arise from "modifications of internal perceptual referents" (Fox, 1985, p. 1552) can be modeled as changes in base activation level. Fox (1985) proposed that contrast effects occur after speaker normalization. In the model proposed here, which has no speaker normalization stage, the contrast can be implemented as a base-level adjustment such that the base level for all exemplars of the contrasting category is increased by some amount.

Johnson (1991) found a type of speaker continuity effect, where it seemed that listeners expected the voice of the speaker to remain constant across time (and a silent gap of 4 s broke this expectation). To model this effect we can adjust the base activation level for all exemplars of a particular speaker's voice, or the set of similar-sounding speakers, so that future categorizations are more heavily influenced by exemplars from the same speaker. This type of speaker continuity mechanism might explain listeners' decreased word-recognition speed and accuracy for multiple-speaker listening as opposed to single-speaker listening (Mullennix, Pisoni, & Martin, 1989).

8.3 ELABORATIONS AND IMPLEMENTATION

This section explores some elaborations that must be considered if the basic exemplar model discussed in the previous section is to be taken seriously as a model of human speech perception. In addition, I discuss some issues that must be addressed if the model is to be implemented and tested. This section is more

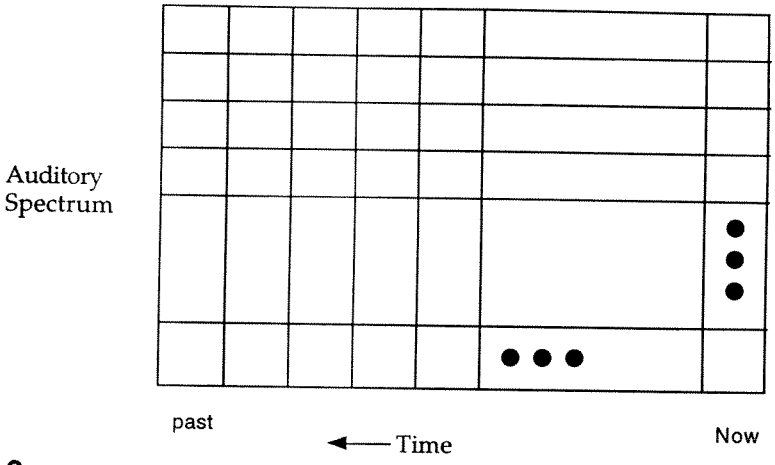

FIGURE 2 An auditory buffer for use in an exemplar model. A quantized auditory spectrum is represented in the vertical dimension with each row representing a frequency region, and time is represented on the horizontal dimension with time passing from right to left.

speculative than the last because I have not yet put together a working implementation of the elaborated model.

8.3.1 Incorporating Time in an Exemplar Model

Previous research on exemplar models has focused on the perception of simple novel visual figures. Consequently, perception of time-varying stimuli has not been considered. One way to incorporate time into an exemplar model is to add a buffer that retains auditory parameters over some interval of time. The model sketched in Figure 1 had a vector of auditory parameters representing the output of the auditory system at one instant in time. This can be extended by incorporating a short-term memory for auditory parameters (Figure 2).

The matrix in Figure 2 is a buffer for incoming speech signals, each column stands for a brief interval of time (on the order of 10 ms), and each row stands for an auditory property. As time passes, the columns shift to the left with the "now" column being filled with the newest set of auditory properties and the columns to the left constituting a veridical short-term memory of the signal (Crowder, 1981). Figure 1 showed connections between each exemplar and only four auditory properties. In this more complicated model each cell in the matrix in Figure 2 is connected to each exemplar.

Another complicating factor is that the similarity between the incoming matrix and the auditory patterns stored with each exemplar must be evaluated each time a new column is added to the matrix. This is obviously not an efficient way to process a speech signal in a serial computing architecture, in which similarity between exemplars and the incoming signal must be evaluated one after the other. But, this point is less problematic if similarity is calculated in parallel, assuming that each exemplar is an independent agent. Still, by including time in the model,

with its consequent need for continuous evaluation of the matrix, we introduce a locus for implementing various approaches for temporal selective attention and segmentation strategies.

For example, a general segmentation routine could be implemented by a surprise detector. In this approach, the matrix is evaluated if the auditory vector added to the "now" column is quite different from the immediately preceding column. This is analogous to saying that attention is drawn to the matrix when something happens there. So, sudden acoustic changes such as those that occur at segment boundaries would trigger a recognition attempt.

Word-based segmentation is also possible. In this approach, if the contents of the matrix have been identified as a word, the system will delay further evaluations of the matrix until the "now" column has scrolled some distance into the past.

Rhythmic attention to the matrix (e.g., stress-based or syllable-based segmentation, Cutler & Norris, 1988) involves probing the matrix cyclically at a rate determined by the intervals between previous recognitions, or previous signal events. Note here that stress and syllable cycles correspond to intervals that might be established by the surprise detector, because stress locations and syllable boundaries are associated with signal events. So the choice of a language-specific segmentation strategy might be guided by attention to certain types of auditory changes.

8.3.2 The Head-Filling-Up Problem

The head-filling-up problem is one of the main weaknesses of an exemplar model of speech perception. It is simply not possible that each experienced auditory pattern is stored at a separate location in the brain. The patterns are too complicated (as suggested by the elaboration in Figure 2), and there are too many of them. So, in order to seriously consider an exemplar model we have to have a way to implement it without storing each exemplar.

Kruschke's (1992) connectionist exemplar model is appealing in this regard because a covering map takes the place of exemplars. Figure 3 illustrates a covering map representing a space of exemplars given two auditory properties and two categories (note that only a few of the many interconnections are illustrated). Each location in the map corresponds to a vector of possible auditory properties. Attention weights, as before, govern the mapping from each property to the map (one weight for each property), and association weights govern the mapping from locations in the map to category nodes (one weight between each location in the map and each category). In a richer representation with more auditory properties and categories, the covering map and hence the number of association weights becomes quite large, but unlike a literal exemplar model its size is bounded.

In Kruschke's approach, the weights, both attention weights and association weights, are learned (by a gradient descent learning procedure) by feedback on correct categorization. Consequently, exemplars are encoded in the model as weight modifications rather than through explicit storage.

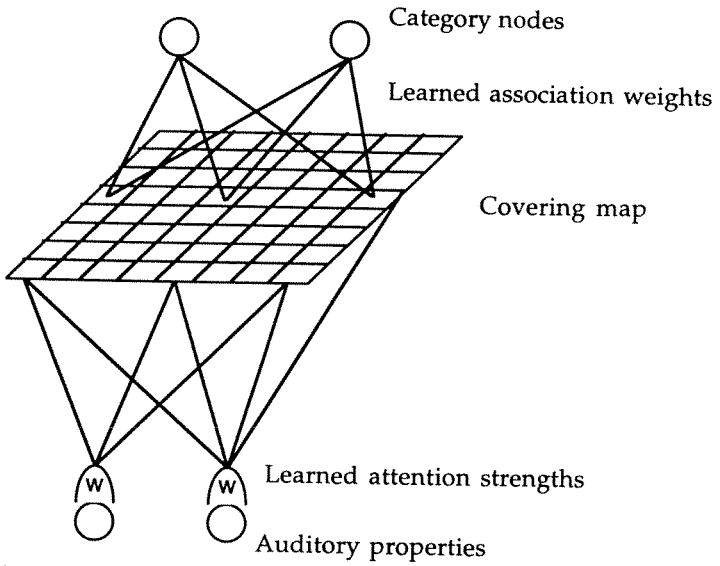

FIGURE 3 A covering map exemplar model of perception. Each input auditory property is connected to each location in the map of possible exemplars, and each location in the map is connected to a set of category nodes.

The covering map as illustrated is essentially a quantized perceptual space. Given the fact that there are limits on the perceivability of incremental changes in duration, pitch, and timbre (just noticeable differences), this assumption seems reasonable. For the sake of a working implementation we may want to take this a step further and use vector quantization (Linde, Buzo, & Gray, 1980).

8.3.3 The Production–Perception Link

Some speech experiences are of one's own speech, which presumably code not only auditory properties and categorical labels, but also articulatory properties. Therefore, an exemplar model can, in principle, also be used to give an account of the production–perception link.[4] This observation leads to some predictions.

In discussing the perception of speech produced by a glossectomee (a person whose tongue was surgically removed), Fowler (1990) invokes the notion *gestural mirage.* She says,

> For the deviant speech of the glossectomized speaker to be perceived as speech at all, the speaker must create acoustic signals that mimic those produced by normal articulations. When he succeeds, listeners hear the normal vocal-tract actions not the compensatory articulations that, in fact, occurred. That is, they hear a mirage. (p. 533)

[4] McGowan (Chap. 11, this volume) discusses some ways of characterizing the articulatory information that might be included in these exemplars.

In extreme cases, such as speech produced without a tongue, or speech produced by a mynah bird, articulatory impressions (the movements a listener would make to produce similar-sounding speech) are obviously imaginary. However, less extreme examples of gestural perception, such as speech produced by people who have slightly different articulatory strategies (Johnson et al., 1993), are no less imaginary. If, as I am suggesting here, the production–perception link is based on one's own speech, then the gestural knowledge derived or generated while listening to others is based on ego exemplars. Gestural mirages are the norm, not the exception.

Consider the perception of a different sort of intention in speech communication. When a person speaks he or she clearly intends to convey some meaning to the listener, but just as clearly we often understand each other only approximately. Our differing experiences of the conventions that link language to the world give us all somewhat different frames of reference for interpreting the utterances we hear. Using my own semantic–pragmatic frame of reference I construct mirages that encode "what I think you said."

An exemplar model that encodes the production–perception link in ego exemplars (but not exclusively so; McGurk & McDonald, 1976) provides a basis for gestural perception and for the imitation of another's speech, but these gestural mirages are not central to perception.

8.4 SOME MODELING RESULTS

This section returns to the unelaborated model set out in section 8.2, fills in some details, and reports the results of some simulations. These results demonstrate in a general way some of the properties discussed in section 8.2, as well as the use of an exemplar model to account for the presentation type effect reported in Johnson (1990b).

8.4.1 A Corpus of Vowel Exemplars

Thirty-nine native speakers of English (14 men and 25 women) read the words *aid, awed, had, head, heed, hid, hood, hud, odd, owed,* and *who'd* five times each (in random order). Five acoustic parameters were derived from each token—fundamental frequency, first, second, and third formant, and vowel duration. The frequency measurements were taken from the midpoint of the vowel.

These exemplars are obviously much simpler than the time-varying auditory spectra envisioned in section 3, but previous research has shown that these acoustic dimensions correlate well with perceptual dimensions for vowels. For example, multidimensional scaling (MDS) studies (Fox, 1981; Shepard, 1972; Singh & Woods, 1970; Terbeek, 1977) have found that vowel formant values and F0 correlate with derived dimensions of the perceptual vowel space. Strange, Jenkins,

TABLE I Model Parameters and Errors for Simulations of Vowel Identification, Sex Identification, and for Fits to the Blocked and Mixed Conditions in Johnson (1990b)[a]

	Vowel identification	Sex identification	Blocked	Mixed
c	0.105	0.36	0.137	0.255
w_{F0},	0.25	0.95	0.2	0.714
w_{F1},	0.25	0.048	0.226	0.012
w_{F2},	0.12	0.0	0.039	0.041
w_{F3},	0.27	0.003	0.415	0.17
w_{dur}	0.1	0.0	0.12	0.078
Error	20%	0.02%	0.047	0.069

[a] Error values for the identification simulations are given in percent incorrectly identified, and error values for the experiment simulations are given in root mean square deviation from the actual results.

and Johnson (1983; see also Strange, 1989) also found that vowel duration is used by English listeners to disambiguate spectrally similar vowels.

In addition to these acoustic properties, each exemplar was encoded with categorical labels indicating (a) the intended word, (b) the sex of the speaker, and (c) the identity of the speaker.

8.4.2 Vowel Identification

Each token in this corpus of exemplars was in turn removed from the corpus and treated as an unknown token to be identified using the remaining exemplars in formulas (1–3) (section 8.2). Base activation levels for all of the exemplars was fixed at 1, and no random noise e_j was added in the calculation of similarity.

The attention weights (w_{F0}, w_{F1}, w_{F2}, w_{F3}, w_{dur}) and the sensitivity parameter c were adjusted using a simulated annealing algorithm (Masters, 1995) to maximize percent correct vowel identification.

Overall percent correct achieved by the best-fitting model was 80%. The model parameters are shown in Table I. The values of the attention weights indicate that each of the five dimensions was important for this exemplar-based vowel classification task.[5] This level of vowel identification accuracy is comparable to human listeners' ability to identify vowels synthesized from midpoint formant values (Lehiste & Meltzer, 1973; Ryalls & Lieberman, 1982).

The vowel confusion matrix produced by the model is shown in Table II. This matrix is significantly correlated ($r = 0.988$), with the confusion matrix reported by Peterson and Barney (1952) for listeners' identifications of naturally

[5] One caution about interpreting these weights. They scale differences in Hz (or ms in the case of vowel duration) and variance on the dimensions is correlated with the mean. So, although the weights for F0 and F1 are the same, the contribution of F1 to vowel identification is larger because the variance of F1 is larger.

TABLE II Vowel Confusion Matrix Produced by an Exemplar Model of Vowel Perception[a]

	Response										
	aid	**awed**	**had**	**head**	**heed**	**hid**	**hood**	**hud**	**odd**	**owed**	**who'd**
aid	**85.34**	.52	.00	.00	10.99	2.09	.52	.00	.00	.52	.00
awed	.52	**68.06**	.52	.00	.00	.00	.00	1.57	22.51	6.81	.00
had	.00	1.06	**84.66**	9.52	.00	.00	.53	1.06	1.06	.53	1.59
head	.53	.00	3.68	**83.16**	.00	7.89	2.11	1.58	.53	.53	.00
heed	11.52	.00	.00	.00	**85.86**	.00	1.05	.00	.00	.00	1.5
hid	4.40	.00	.55	6.59	2.75	**84.07**	.55	.00	.00	.55	.55
hood	.00	.00	.00	.00	.53	.00	**78.42**	8.42	.53	10.00	2.11
hud	.53	2.14	4.28	2.67	.53	.53	11.76	**74.87**	.53	1.60	.53
odd	.00	25.93	1.59	.53	.00	.00	.53	1.59	**68.25**	1.59	.00
owed	.00	5.79	.00	.00	.00	.00	9.47	1.05	.00	**78.95**	4.74
who'd	.52	.00	.00	.00	.52	1.56	5.21	1.04	.00	3.13	**88.02**

[a]The speaker's intended vowel is listed down the left column and the model's identification response is listed in the first row. Percent correct identifications averaged across speakers are printed in boldface and confusions are indicated in the off-diagonals.

produced vowels (in a list comparable to the one used in this study), and the confusions—that is, the matrix without the diagonal—are also significantly correlated ($r = 0.716$) with the confusions in Peterson and Barney's study. This indicates that the acoustic measures used in this study are adequate for further model studies of vowel perception (i.e., without the elaborations mentioned in section 8.3). In addition, these results show that it is possible, using an exemplar model, to simulate human vowel perception without normalization.

8.4.3 Sex Identification

In a simulation that was analogous to the one just reported, the model was tuned to maximize correct sex identification. By giving greatest attention to F0 (the best-fitting parameters are shown in Table I), the model was able to achieve 99.8% correct sex identification for these tokens. But even with the model parameters that resulted in the best vowel identification (previous section) the sex of the speaker was correctly identified 96% of the time.

8.4.4 Twiddling Some Knobs

Formant values for the *hood–hud* continua that were used in Johnson (1990b) are shown in Table III. The vowel portions of those stimuli were 190-ms long and two continua were synthesized, one with F0 at 120 Hz and one with F0 at 240 Hz.

TABLE III Formant Values of the *Hood–Hud* Continuum Used in Johnson (1990b) and the Simulations Reported in this Chapter

Token #	1	2	3	4	5	6	7
F1	474	491	509	526	543	561	578
F2	1111	1124	1137	1150	1163	1176	1189
F3	2416	2424	2432	2440	2448	2456	2464

The vowel identification model described in section 8.4.2 (the set of model parameters that gave the best vowel identification) was used to identify these tokens in a forced-choice experiment. As mentioned earlier, I assumed that in a forced-choice experiment only exemplars of the permitted categories are used to evaluate the stimuli. The results from an open-class version of this simulation were comparable to the results reported here, but some of the stimuli in the middle of the high F0 continuum were identified as "owed."

In order to compare the results of this simulation with listeners' identification performance, I calculated the probability of a *hood* response for each token using Luce's (1963) choice rule—the probability of a *hood* response is equal to the similarity of the token to the *hood* category divided by the sum of the similarities of the token to the *hood* and the *hud* categories, where "similarity" is defined by formula (2) above.

Figure 4 shows probability of a *hood* response as a function of token number for both the high F0 and low F0 continua. These results show a clear speaker normalization effect. That is, as has been found with human listeners (Johnson, 1990b; R. L. Miller, 1953), tokens with high F0 were more likely to be identified as *hood.* Later paragraphs will discuss some ways that the exemplar model's parameters can be adjusted to quantitatively simulate listener performance. For now I want to emphasize the fact that the results shown in Figure 4 show that an exemplar model optimized for correct vowel identification over a corpus of naturally produced vowels shows a speaker normalization effect without any adjustment of the model parameters.

In section 8.2.2 I discussed the possibility that base activation levels of exemplars (N_j in formula 2) might be sensitive to word frequency, giving rise to the Ganong (1980) effect. Figure 5 shows a simulated Ganong effect produced by manipulating base activation level (Figure 5 shows results for the low F0 continuum; results for the high F0 continuum were comparable). As before, probability of a *hood* response to tokens from the *hood–hud* continuum were calculated using Luce's choice rule. In one simulation *hood* was treated as a more frequent word than *hud* by setting N_{hood} to 0.6 and N_{hud} to 0.4 (where N_{hood} indicates the N_j values associated with all of the exemplars of *hood*). Results of this simulation, using the vowel-identification model parameters and forced-choice identification, are labeled "greater *hood* base activation" in Figure 5. In a second simulation, *hud* was

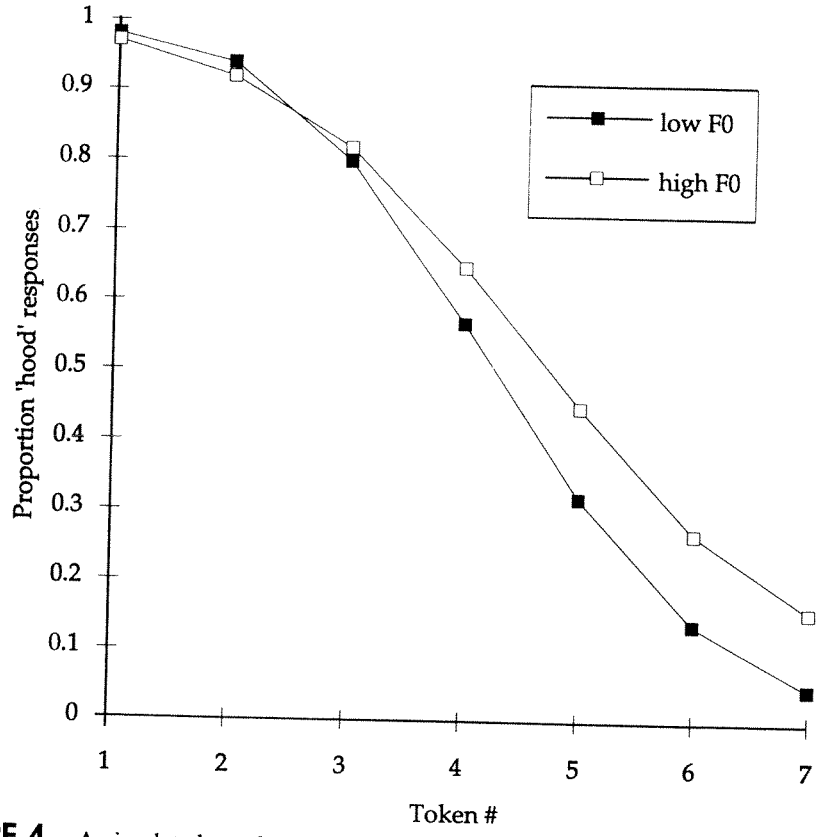

FIGURE 4 A simulated speaker normalization effect. The response is of a vowel identification exemplar model to a continuum ranging from *hood* to *hud.* One version of the continuum has an F0 value of 120 Hz (filled symbols) and another version has an F0 value of 240 Hz (open symbols).

treated as the more frequent word by reversing the N_j values. Results of this simulation are labeled "greater *hud* base activation" in Figure 5.

The boundary shift seen in this simulation is comparable to the boundary shift found by Ganong (1980) when word frequencies of the response alternatives were manipulated. He found a tendency for high-frequency words to be perceived more often (requiring less convincing acoustic cues) than low-frequency words. Similarly, in the exemplar model, when base activation of the exemplars of a category is high a stimulus is more likely to be categorized as belonging to that category than it would have been otherwise.

As mentioned above, the effect of syntactic and semantic expectations on speech perception can also be modeled by manipulating base activation. I could have said that Figure 5 shows a simulation of the perception of the *hood–hud* continuum in two semantic contexts, one that leads to the expectation that the word will be *hood* and one that leads to the expectation that the word will be *hud.* In this interpretation the N_j values reflect semantic priming rather than word frequency.

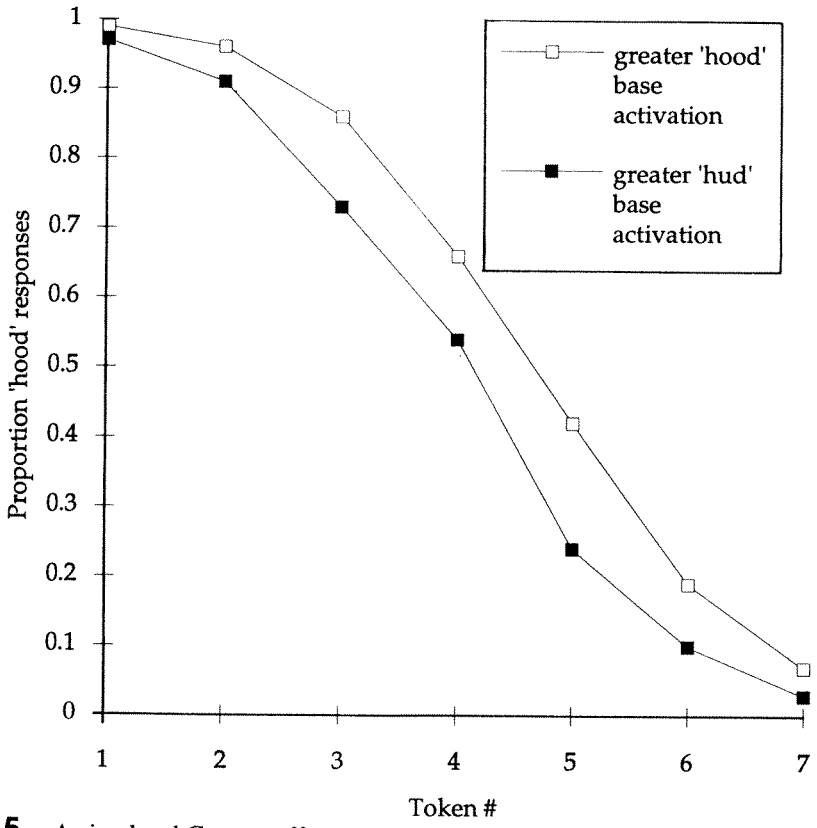

FIGURE 5 A simulated Ganong effect. The figure shows the response of a vowel identification exemplar model to a continuum ranging from *hood* to *hud* (low-F0 stimuli only). When the base activation levels of all *hood* exemplars were elevated there were more "hood" responses (open symbols), and when the base activation levels of all *hud* exemplars were elevated there were more "hud" responses (filled symbols).

8.4.5 Simulating the "Presentation Type" Effect

In Johnson (1990b) I reported that a speaker normalization effect (for the *hood–hud* continuum that we have been discussing) occurs when the stimuli are randomly intermixed, but disappears when the stimuli are blocked by F0. Analogous effects have been found in vowel identification (see Nearey, 1989, Tables 1 & 2) and auditory word recognition (Mullennix, et al., 1989). These studies have found that perceptual accuracy is reduced when words produced by different speakers are randomly intermixed, as compared to perception of the same stimuli blocked by speaker.

Nusbaum and Morin (1992) presented data suggesting that mixed-speaker presentation causes listeners to shift their attention to acoustic properties that are relevant for speaker identification (F0 and higher formants in their experiment). It makes sense to assume that listeners will tend to focus their attention to stimulus dimensions that vary from trial to trial, and thus that in a mixed-speaker presen-

tation listeners will attend more to F0 than they do when stimuli are blocked by speaker. This predicts that in simulating the results of Johnson (1990b), the attention weight for F0 will be larger in the mixed condition than in the blocked condition.

Figure 6 shows the perceptual data from Johnson (1990b) plotted with light lines, and the results of an exemplar model simulation of the experiment plotted with dark lines. Figure 6A shows the blocked-speaker condition, and Figure 6B shows the mixed-speaker condition. Responses to stimuli with high F0 (240 Hz) are plotted with open squares and responses to stimuli with low F0 (120 Hz) are plotted with filled squares. The separation of the response functions for low and high F0 tokens in the mixed condition is similar to the "speaker normalization effect" simulated in section 8.4.4 (see Figure 4). However, the magnitude of the separation is much larger in the listeners' performance.[6]

To simulate this experiment the simulated annealing procedure was used to find model parameters that gave the best fit to the actual data in the blocked condition (assuming that the N_j for *hood* and *hud* exemplars were 1 and those for all other vowel categories were 0, and using Luce's choice rule as before). Then model parameters that provided the best fit to the actual data in the mixed condition were found. The model parameters and root mean square (RMS) error are shown in Table I.

The fits to the data are quite good; the predicted identification results are different from the actual data by only about 5% for each token on average. Also, as predicted by Nusbaum and Morin (1992), the attention weight for F0 was larger in the mixed condition than in the blocked condition. Note that the model parameters that gave the best fit in the blocked condition are very similar to the parameters that were found in the vowel identification task, and that the model parameters that gave the best fit in the mixed condition were similar to the parameters found to provide maximally accurate sex identification. This finding is consistent with the hypothesis that listeners focus their attention on the changing identity of the speaker in the mixed condition. Interestingly, this pattern of attention allocation may also be detrimental for vowel identification.

I tested this speculation by running the vowel identification simulation that was described in section 8.4.2, with the model parameters found in this simulation for the mixed F0 condition. Vowel identification accuracy was reduced from 80% (the best the exemplar model could do) to 72%. This result (as well as the experimental findings reported by Nusbaum & Morin, 1992) suggests that the talker variability effect found by Mullennix et al. may have been caused by a tendency for listeners' attention in the mixed condition to be drawn to acoustic properties that are more relevant for speaker identification than for word recognition.

[6]The lack of boundary shift in blocked condition may have resulted from the fact that the synthetic stimuli were more similar to male exemplars, or may have been the result of a Parducci (1975)-style response bias.

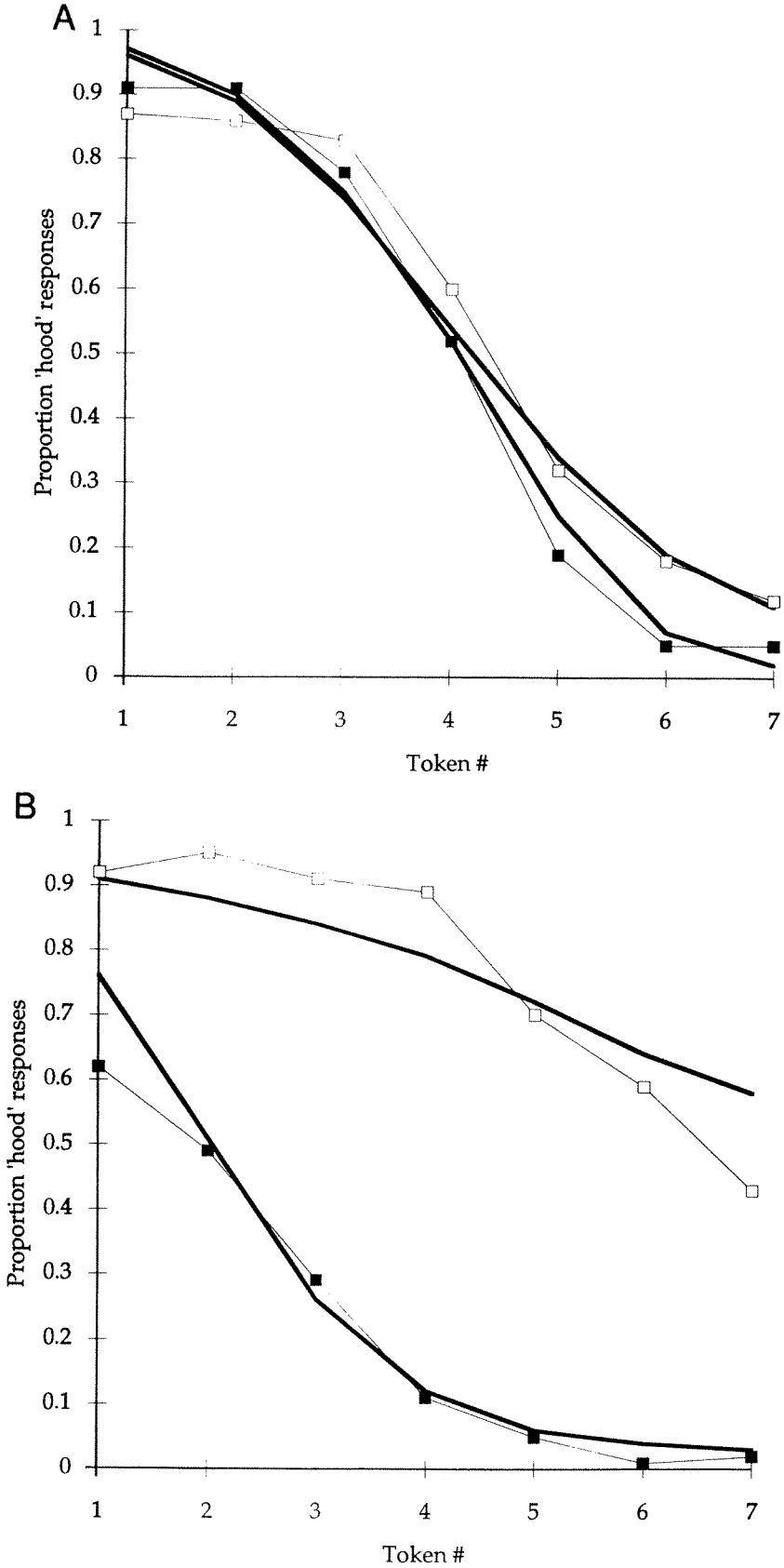

FIGURE 6 Model fits to the Johnson (1990b) presentation-type experiment. The actual data are shown with light lines and symbols (open for high F0 tokens, and filled for low F0 tokens). Model fits are shown with darker lines. (A) the data and model fits for the blocked condition; (B) the data and model fits for the mixed condition.

8.5 CONCLUSION

In this chapter I have outlined an exemplar-based model of speech perception, indicated some ways that the basic model can be elaborated, and presented results of several simulations showing that this model of speech perception mimics some aspects of human vowel perception performance and response to talker variability.

The picture of vowel normalization that emerges from this study is radically different from the traditional view. In this approach, speaker normalization behavior (both the ability to recognize vowels produced by different speakers and the presence of a boundary shift in the *hood–hud* continuum) is not caused by a representation changing process. Instead, these patterns of behavior emerge from the complex internal structure of linguistic categories. Because the model retains the variability encountered in speech it is able to cope with the variability that it encounters in new tokens. That is, the model *uses* talker variability in speech processing.

The title of this chapter, "Speech perception without speaker normalization" implies that I view speech perception as a passive process (see Nusbaum & Magnuson, Chap. 6, this volume). This is not the case. I have explicitly assumed that listeners may selectively shift their attention to different acoustic aspects of the speech signal, may cyclically focus attention on the signal, and may use top-down information to increase sensitivity to selected perceptual categories. All of these sources of flexibility in listener's behavior involve active response to the listening situation. However, I do not include among the active processes of speech perception representation-changing normalization routines of any sort. The model outlined here is an active, flexible model of speech perception without speaker normalization.

I have focused on one particular source of variability in the speech signal—acoustic differences between talkers. Here I would like to point out that I expect that the elaborated model outlined in section 8.3 will also be able to handle other sources of variability as well. For example, the model would compensate for dialect variation by using the dialect variability inherent in remembered exemplars. Some effects of dialect familiarity, for instance, would emerge naturally from this model without having to suppose that a dialect "normalization" rule is learned. Variation in the speech signal caused by changes in speaking rate would be handled in the same way (including vowel reduction and even resyllabification and extensive gestural reorganization). So, although I have focused on talker variability in this chapter, I am aiming for a general model that uses the same mechanism to handle many different sources of variability in the speech signal.

Although much remains to be specified, an exemplar-based approach offers an important alternative to the traditional view of speech perception. Future research needs to be devoted to fleshing out the elaborated model, and exploring control mechanisms that might be used to manage the flexibility of an exemplar model.

ACKNOWLEDGMENTS

This research was supported by the National Institute on Deafness and Other Communication Disorders under Grant No. 7 R29 DC01645-04. Elizabeth Strand recorded the speech database, and David Pimental made the acoustic measurements. Discussions with Rob Fox, Neal Johnson, John Mullennix, and the members of the Ohio State University Linguistics Lab were helpful as I worked on this chapter.

REFERENCES

Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication, 4,* 59–69.

Crowder, R. (1981). The role of auditory memory in speech perception and discrimination. In T. Meyers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 167–179). New York: North-Holland.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 113–121.

Estes, W. K. (1993). Concepts, categories, and psychological science. *Psychological Science, 4,* 143–153.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3–28.

Fowler, C. A. (1990). Calling a mirage a mirage: Direct perception of speech produced without a tongue. *Journal of Phonetics, 18,* 529–541.

Fox, R. A. (1981). *Perceptual structure of English monophthongs and diphthongs.* Paper presented at the 17th regional meeting of the Chicago Linguistic Society, Chicago.

Fox, R. A. (1985). Auditory contrast and speaker quality variation in vowel perception. *Journal of the Acoustical Society of America, 77,* 1552–1559.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110–125.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE, AU-16,* 78–80.

Goldinger, S. D. (1992). *Words and voices: Implicit and explicit memory for spoken words. Research on speech perception* (Tech. Rep. No. 7). Bloomington: Indiana University Speech Research Laboratory, Department of Psychology.

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

Johnson, K. (1990a). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88,* 642–654.

Johnson, K. (1990b). Contrast and normalization in vowel perception. *Journal of Phonetics, 18,* 229–254.

Johnson, K. (1991). Differential effects of speaker and vowel variability on fricative perception. *Language and Speech, 34,* 265–279.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America, 94,* 701–714.

Joos, M. (1948). Acoustic phonetics. *Language, 24,* (Suppl. 2), 1–136.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

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.

Lehiste, I., & Meltzer, D. (1973). Vowel and speaker identification in natural and synthetic speech. *Language and Speech, 16,* 356–364.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communication, COM-28,* 84–95.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Masters, T. (1995). *Advanced algorithms for neural networks: A C++ sourcebook.* New York: Wiley.

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature (London), 264,* 746–748.

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

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114–2134.

Miller, R. L. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America, 25,* 114–121.

Mullennix, J. M., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365–378.

Nearey, T. M. (1978). *Phonetic feature systems for vowels.* Bloomington, IN: IU Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–2113.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 113–134). Tokyo: Ohmsha Publishing.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1–20.

Parducci, A. (1975). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press.

Perkell, J. S., & Klatt, D. H. (Eds.) (1986). *Invariance and Variability in Speech Processes.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the identification of vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Potter, R., & Steinberg, J. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America, 22,* 807–820.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1995, September). *Talker identification is based on phonetic information* (Tech. Rep.). New York: Barnard College, Speech Perception Laboratory.

Ryalls, J., & Lieberman, P. (1982). Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America, 72,* 1631–1634.

Schacter, D. L., & Church, B. A. (in press). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David, & P. B. Denes (Eds.), *Human communication: A unified view.* New York: McGraw-Hill.

Singh, S., & Woods, G. (1970). Perceptual structure of 12 American vowels. *Journal of the Acoustical Society of America, 49,* 1861–1866.

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2560 visual stimuli. *Psychonomic Science, 19,* 73–74.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America, 85,* 2135–2153.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America, 74,* 695–705.

Syrdal, A., & Gopal, H. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086–1100.

Terbeek, D. (1977). A cross-language multidimensional scaling study of vowel perception. *UCLA Working Papers in Phonetics, 37,* 1–271.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America, 69,* 1465–1475.