A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus–response alignment

Lynne E. Bernstein
Center for Auditory and Speech Sciences, Gallaudet University, Washington, DC 20002-3695

Marilyn E. Demorest
Department of Psychology, University of Maryland Baltimore County, Baltimore, Maryland 21228-5398

Silvio P. Eberhardt
Department of Engineering, Swarthmore College, Swarthmore, Pennsylvania 19081-1397

(Received 18 February 1993; accepted for publication 4 February 1994)

A solution to the following problem is presented: Obtain a principled approach to studying error patterns in sentence-length responses obtained from subjects who were instructed to simply report what a talker had said. The solution is a sequence comparator that performs phoneme-to-phoneme alignment on transcribed stimulus and response sentences. Data for developing and testing the sequence comparator were obtained from 139 normal-hearing subjects who lipread (speechread) 100 sentences and from 15 different subjects who identified nonsense syllables by lipreading. Development of the sequence comparator involved testing two different costs metrics (visemes versus Euclidean distances) and two related comparison algorithms. After alignments with face validity were achieved, a validation experiment was conducted for which measures from random versus true stimulus–response sentence pairs were compared. Measures of phonemes correct and substitution uncertainty were found to be sensitive to the nature of the sentence pairs. In particular, correct phoneme matches were extremely rare in random pairings in comparison with true pairs. Also, an information-theoretic measure of uncertainty for substitutions in true versus random pairings showed that uncertainty was always higher for random than for true pairs.

PACS numbers: 43.71.Es, 43.71.Gv, 43.71.Ma

INTRODUCTION

Traditionally, speech perception research has used phoneme confusion data obtained with nonsense syllable stimuli to infer perceptual processes or characteristics (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973). The problem of interest to us was to develop a method for studying phoneme errors in perception of connected speech. In order to accomplish this, a principled method was needed for aligning stimulus and response on a phoneme-by-phoneme basis.

Example 1 is typical of the data we were interested in, which were obtained from a study of lipreading (speechreading):

Stimulus: Why should I get up so early in the morning?
Response: Watch what I'm doing in the morning! (1)

In the example, three words are correct (*in the morning*), but the incorrect words do not appear to be selected at random. By inspection, we see that incorrect words contain both potentially correct as well as perceptually similar phonemes to those in the stimulus words. For example, the /w/ in *watch* may correspond perceptually to the /w/ in *why*. Some stimulus and response phonemes are merely similar, for example, the /S/[1] in *should* and the /C/ in *watch*. In this paper, we describe the development of a system to generate phoneme-to-phoneme alignments for stimulus-response pairs such as example 1, based on sequence comparison methods (Kruskal and Sankoff, 1983). We refer to our system for aligning stimulus-response pairs as a *sequence comparator*.

I. PRELIMINARY SEQUENCE COMPARATOR

A. Description of the database used for development

1. Subjects

The data used in developing the comparator were obtained from adult subjects who were recruited to be in experiments reported in Bernstein *et al.* (1989), Eberhardt *et al.* (1990), and Demorest and Bernstein (1992). Subjects who lipread sentences were 139 young adults who reported having normal hearing and normal or corrected vision and who were native speakers of English. Seventy subjects were men. Four men and eleven women visually identified nonsense syllables (Eberhardt *et al.*, 1990); their data were used to generate costs based on a Euclidean metric (see below). All subjects were paid for their participation.

2. Stimuli

Sentences were the 100 CID Everyday Sentences (Davis and Silverman, 1970), recorded on video laserdisc (Bernstein and Eberhardt, 1986). Sentences were spoken by a male and a female talker. Consonant-vowel (CV)

 0001-4966/94/95(6)/3617/6/$6.00

nonsense syllables used for costs estimates (see below) were spoken by the same male and female talkers. Two tokens from each talker were obtained for the 22 consonants, /p,b,m,f,v,T,D,w,r,C,J,S,Z,t,d,s,z,k,g,n,l,h/, combined with the vowel /ɑ/. The isolated vowel /ɑ/ was also recorded. All stimuli were displayed on a color video monitor.

3. Procedures

a. Sentences. Subjects were instructed to type whatever they thought the talker had said, including sentence and word fragments, and they were encouraged to guess. Testing took 20–30 min to complete. Demorest and Bernstein (1992) reported in an analysis of data from 104 of the 139 subjects that, overall, subjects obtained an average of 1.55 words correct per sentence, which represents 20.8% of the 749 stimulus words. For individual subjects, the performance range was from 0.01 to 4.02 words correct per sentence (0.3%–53.7%). For individual sentences it was from 0.21 to 5.67 words correct (2.6%–91.3% of the sentence).

b. Nonsense syllables. Subjects performed in a 23-alternative forced-choice paradigm (Eberhardt *et al.*, 1990). For each talker, there were 46 tokens (two tokens each of 22 CV syllables plus the isolated vowel /ɑ/). These tokens were presented in random order across ten blocks of trials in each of two sessions.

4. Response preparation

Errors in responses to sentence stimuli were corrected when unambiguously attributable to spelling or typing. Software was written that transmitted each response to a text-to-speech synthesizer (DECtalk DTC01, Version 2.0) that produces a quasiphonemic transcription as one of its stages in the text-to-speech process (Educational Services Department, Digital Equipment Corporation, 1984). The phonemic transcription was stored line-by-line in a new data file. Transcription errors were hand corrected. The transcribed responses formed the database for sequence comparator development. Because the 139 subjects did not always provide a response to each of the 100 sentences, the number of responses was 12 291.

B. Preliminary algorithm and costs assignment

Sankoff and Kruskal (1983) provide theoretical and detailed discussions of sequence comparison techniques, and the reader who wishes to implement these algorithms is urged to consult their chapters. Sequence comparison algorithms belong to the class of dynamic programming procedures. In conceptual terms, sequence comparison rests on (i) recurrence equations that are used to minimize *distance* between two sequences of elements and (ii) *costs* assigned to the possible alignments between the elements that are to be aligned. A main feature of sequence comparison methods is that they take into account the possibility that the two sequences to be compared may differ in length because of the occurrence of elements in the stimulus (target) but not in the response (source) or vice versa.

TABLE I. Cost of seven types of elementary alignments.

Type	Example	Cost
Exact match	a,a	0
Substitution within a viseme group	b,p,m	1
Substitution within consonants, but across consonantal visemes	b,g	2
Substitution within vowels, but across vocalic visemes	a,i	2
Substitution of consonants for vowels and vice versa	a,b	3
Insertion of a vowel or consonant in the response		1
Deletion of a vowel or consonant in the stimulus		1

Initially, the simplest recurrence equation was implemented (Kruskal, 1983, pp. 25–26), which calculates the minimum distance among all possible alignments of strings (without permuting order). The distance metric that was initially adopted was based on previous research by Fisher (1968), Walden *et al.* (1977), and Wozniak and Jackson (1979) on visual confusability of phonemes in nonsense syllables. In those studies, groups of phonemes that are mutually confused at levels substantially above chance (e.g., 75% mutual confusability) are referred to as *visemes* (e.g., /g,k,n,t,d,y/ is a viseme group obtained by Walden *et al.*). The viseme was temporarily adopted as the basis for assigning phoneme-to-phoneme costs.

Table I summarizes the costs of the various types of elementary alignments. Elements in the stimulus not found in the response are referred to as *deletions*, and elements in the response not found in the stimulus are referred to as *insertions*. The term *substitution* refers to alignment of elements that are different from each other. When an alignment is constructed, characters referred to as *indels* (here '-') mark the occurrence of an insertion or deletion. Indel costs were assigned so as to prevent alignments between an indel and a phoneme, when a substitution was more plausible (in terms of what is known about visual phonetic ambiguity). For example, the alignment of stimulus /b/ with response /p/ would cost 1, but alignment of stimulus /b/ with response indel plus stimulus indel with response /p/ would cost 2. Since previous research had not derived a viseme classification for every English phoneme, some cost assignments were based on analogy to phonemes with similar place of articulation. The cost for aligning a consonant with a vowel was made relatively high based on the rationale that consonants are associated with articulatory closures, and vowels are associated with mouth openings of relatively long duration. Also, visual vowel identification is quite accurate (Montgomery and Jackson, 1983).

C. Results

The basic sequence comparison algorithm in combination with the viseme-based costs was evaluated with the responses to the first 50 CID Everyday Sentences, lipread by 67 subjects, 29 of whom were randomly assigned to the female talker. The combination of the basic sequence comparison algorithm and the viseme costs estimates was judged inadequate, because numerous equal-distance align-

ments were generated. Example 2 shows two of 11 alternate alignments obtained for a particular stimulus–response pair (Stimulus: *Here's a nice quiet place to rest.* Response: *That's the way it goes*).

Alignment 1

```
Stimulus:  hIrs-xnAskwA|tplestUrEst
Response:  d@tsDx----weItg-oz------
```
(2)

Alignment 2

```
Stimulus:  hIrs-xnAskwA|tplestUrEst
Response:  D@tsDx----w-----eIt-goz-
```

For some stimulus–response pairs, alternate alignments numbered in the thousands. In addition to the problem of multiple alignments, many examples were observed in which incorrect response words were fragmented so as to align across several stimulus words (e.g., /we/, *way*, in Alignment 2 above). Phoneme-to-phoneme similarity had been maximized at the expense of putatively realistic temporal correspondences.

II. MODIFIED SEQUENCE COMPARATOR

The sequence comparator was modified in two steps, reflecting two possible sources of its unsatisfactory performance. First, an augmented algorithm was implemented, then the costs assignments were modified to better represent perceptual dissimilarities among phonemes.

A. The augmented algorithm

The augmented algorithm (Kruskal and Sankoff, 1983, pp. 297–299) assigns an additional cost to an indel when it begins and/or ends a sequence of indels. The effect of this algorithm is to reduce the fragmentation of words caused by the use of indels. The augmented algorithm and original viseme-based costs assignments were evaluated with the same data submitted to the original comparator. Although the number of alternate equal-distance alignments could be reduced by employing a sufficiently high cost for beginning a string of insertions or deletions, the mean number of alignments per response was still an unacceptably high 7.6.

B. Modifications to the costs assignments

Viseme groupings, because of the manner in which they are obtained, obscure more subtle patterns of dissimilarities among phonemes (see, e.g., Benguerel and Pichora-Fuller, 1982; Scheinberg, 1988). If the costs assignments are regarded as a representation of perceptual dissimilarity space, multidimensional scaling suggests itself as a source for cost estimates. Having obtained a scaling solution, it is straightforward to use the coordinates of each of the scaled entities to obtain Euclidean distances among them. Therefore, in order to obtain a more refined set of costs assignments, scaled distances among phonemes were derived from data on visual identification of nonsense syllables.

Consonant data from nonsense syllable identification confusions were submitted to multidimensional scaling using the ALSCAL procedure (SPSS, Inc., 1986). A four-dimensional solution was accepted with S-stress=0.084, R^2=0.96. The coordinates of the 23 stimuli were submitted to the SPSS^x PROXIMITIES procedure to obtain a matrix of scaled distances. To obtain distances among vowels, the confusion matrix obtained by Montgomery and Jackson (1983, Table I, p. 2136) was analyzed using the ALSCAL procedure. In conformity with them, a two-dimensional solution was accepted. The resulting value of S-stress was 0.264 and R^2 was 0.77. Intervowel costs were obtained via the same procedures used with the consonants. Eight symbols used by DECtalk received cost assignments based on theoretical visual similarities, because perceptual judgments were not available. Those symbols were /x,|,y,Y,L,N,G,M/. For example, the syllabic /N/, as in /b ^ tN/(*button*), was defined to be visually identical to /n/. Substitution costs between vowels and consonants, and indel costs were chosen as a result of informal but extensive observations of the comparator's performance.

The augmented algorithm and the costs assignments based on Euclidean distances were submitted to an initial evaluation incorporating two criteria: Was the number of unique solutions (i.e., stimulus–response pairs for which only one alignment was obtained) adequately high? Were the alignments plausible on *a priori* grounds related to knowledge about the visual confusability of phonemes, that is, did they have face validity? This evaluation made use of the responses to the entire database.

C. Properties of equal-distance alignments

The augmented algorithm required selection of a value to be charged for initiating a string of indels. All the data were run through the sequence comparator several times with a range of initiation values. Above a particular value, the number of alternate equal-distance alignments remained fairly constant. A value was chosen that resulted in a mean of 1.32 alternate alignments per stimulus–response pair. This resulted in 9 540 responses uniquely aligned and 2 129 with dual equal-distance alignments. The characteristics of the 2 129 dual alignments were investigated: 46.5% (990 responses) were the result of one indeterminate phoneme alignment; 37.7% (803 responses) occurred because a string of response phonemes was aligned with a string of insertions placed in two alternate locations in the stimulus; and the remaining 15.8% (336) with alternate alignments involved multiple substitution differences. Therefore, most dual alignments may be considered trivially different. That is, they tend to differ by the alignment of indel characters or one phoneme.

D. Face validity

Table II gives examples of alignments for one stimulus sentence. Close scrutiny of many such alignments suggested that this particular sequence comparator had achieved the initial goal, and hence further evaluation was appropriate.

TABLE II. Selected alignments for sentence 76, "She'll only be gone a few minutes."

Response	Alignment
1 She won't let for a minute	#Si L#on- li#- bi#gcn#x#fY#- mIn¦ts# #Si#w ont#lE t#-- --- - fR#x#mIn¦t-#
2 You know you'll be gone for months	#Si L#o --n li#bi#gcn#x#fY#mIn¦ts# #-Y#n o#yuL#-- bi#gcn#- fR#m^n-Ts#
3 Don't be for me	#SiL#onl i#bi#gcn#x#fY#mIn¦ts# #--d ont#- bi#--- - fR#mi----#
4 Don't look out for me	#SiL#on- li#bi#gcn#x#- fY#mIn¦ts# #--d ont#lU -- --k#W t#fR#mi----#
5 You know who my offer please	#Si L#o nli#bi#gcn#x#fY#mIn¦ts# #-Y#n o#h-u#ma#--- c fr#p-li-z#
6 You don't even belong in this	#Si L#onl i#--- bi#gcn#x#fY#mIn -¦ts# #-Y#d ont#i vxn#bx lcG#- -- -In#DI-s#

Note: The symbol "#" represents a word boundary.

III. EXPERIMENT: RANDOMLY ASSIGNED VERSUS TRUE STIMULUS–RESPONSE PAIRS

The comparator's sensitivity to characteristics of the data supplied to it was evaluated. An initial question was: Would it produce numerous phoneme-to-phoneme matches regardless of the nature of the stimulus–response pairs provided as input, for example, if there were no perceptual basis for the relationship between stimulus and response strings? To evaluate this question, the number of exact phoneme matches obtained in alignments from true versus randomly assigned stimulus–response pairs was compared. Phonemes correct in randomly assigned pairs should reflect general constraints (phonetic, lexical, and syntactic) across the entire set of stimuli and responses. Matches in true pairs should reflect perceptual processes, in addition to the same general constraints.

However, because the *purpose* of the sequence comparator was primarily to study error patterns in sentences, it was necessary to determine that phoneme substitution patterns varied as a function of the data. If substitution patterns are the same regardless of how stimulus and response strings are paired, then clearly the alignments reflect only the properties of the comparator and not of perceptual processes. To evaluate this question, uncertainty measured in bits was calculated for phoneme-to-phoneme substitutions in true versus randomly assigned stimulus-response pairs. That is, the alignment procedure was held constant, but the responses were submitted to the comparator with true or randomly assigned stimulus sentences. Uncertainty (or entropy) was calculated per stimulus phoneme as

$$-\sum_{k=1}^{n} p_k \log_2 p_k,$$

where p_k is the proportion of phoneme substitutions for the kth response phoneme and k is an index of summation that represents each possible substitution error in turn. Uncertainty is high when a stimulus phoneme is paired with a variety of relatively, equally probable responses and low when it is paired with a few relatively, highly probable responses. Since the possible substitutions are a function of the costs assignment, even the random pairs will have a certain quantity of systematicity. Nevertheless, substitution uncertainty should be lower for phonemes in true pairs, because perceivers should generate even more systematic elementary alignments than would a random pairing. One reason why this might occur is that perceivers may constrain their responses in relation to preceding sentential semantic constraints.

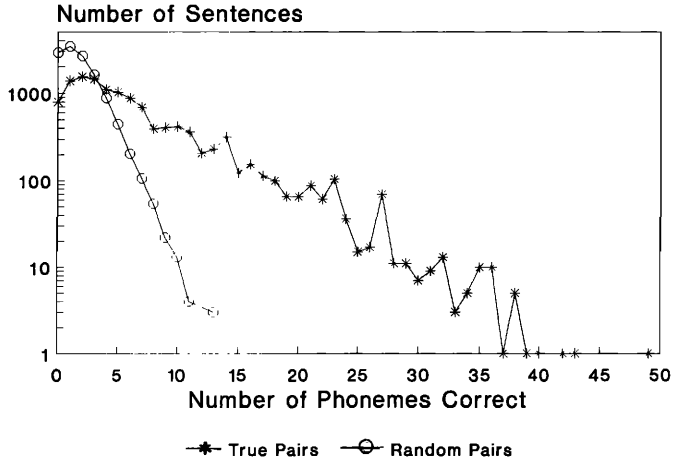

FIG. 1. Distribution of sentences as a function of the number of correct phonemes in true versus random pairings of stimulus and response. $N=11\,468$ random pairs, $N=12\,291$ true pairs.

A. Method

1. Data preparation

An SPSS^x program randomly assigned stimulus numbers to each of the 12 291 responses such that the true stimulus and response were never paired. Then the true and random stimulus-response pairs were separately submitted to the sequence comparator. A response distribution was obtained for *each* of the 2 124 stimulus phonemes in the 100 CID Sentences. Substitution uncertainty was then calculated on substitution errors in the response distributions for individual phonemes for true versus random pairings. Only unique alignments and one of the two from dual alignments were used in the analyses.

2. Results

Figure 1 shows the number of stimulus-response pairs that resulted in from 0 to 49 phonemes correct for true versus random pairs. The figure shows that when sentences were randomly assigned, the likelihood of obtaining six or more phonemes "correct" (i.e., exact matches) was extremely low. In fact, 96.7% percent of the random pairs (11 892) resulted in five or fewer exact phoneme matches. In contrast, 49% of the true pairs resulted in five or more exact phoneme matches. Thus true responses can be distinguished from randomly assigned responses.

The low number of phoneme matches for random pairings needs to be considered in relation to the content of the stimulus set as a whole. The CID Everyday Sentences have relatively few lexical items. The type-token ratio was 0.42 (i.e., 313 lexical items and 749 tokens). Two hundred four lexical items were used only once, but 545 were repeated, with *the* providing the highest contribution at 35 occurrences. Furthermore, a concordance analysis showed that many words occurred in similar contexts. For example, *how do you*, occurred three times. Random stimulus–response pairs would be expected to produce a certain quantity of phoneme matches simply on the basis of the restricted range of words and phrases in the stimuli. That this occurred rarely shows that the comparator is not free to optimize isolated portions of alignments at the expense of the total alignment, as guaranteed by the recurrence algorithm.

Figure 2 shows average substitution uncertainty, holding type of pair (true or random) constant. Each data point was obtained by averaging the substitution uncertainty for all occurrences of a particular phoneme. The figure shows that random pairs consistently produced higher phoneme substitution uncertainty. Thus although alignments are constrained by the costs assignment, this constraint is not so great as to conceal perceptually generated error proportions. This result supports the hypothesis that errors in true responses reflect perceptual processes that operate similarly across subjects, producing systematic error patterns.

3. Discussion

The heuristic nature of the sequence comparator deserves comment. It was assumed that with appropriate additional tools (such as statistics and software to extract patterns in alignments, which we have written) systematic performance patterns could be captured within an experimental paradigm that requires subjects to perform a relatively unconstrained task. This method was applied to results from a comparative study of normal-hearing and profoundly hearing-impaired young adults (Bernstein, 1993). Among results was the finding that substitution uncertainty was higher for normal-hearing lipreaders. Other analyses of comparator output have been shown to be effective in illuminating individual differences among normal-hearing lipreaders (Demorest and Bernstein, 1991).

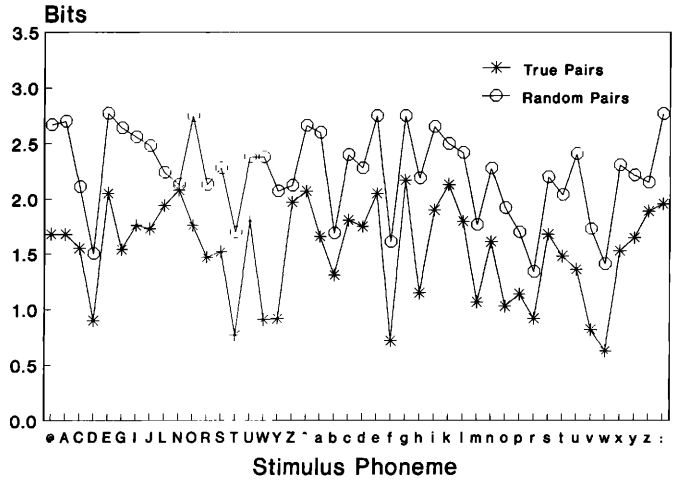

FIG. 2. Average substitution uncertainty for each stimulus phoneme as a function of the type of data: true versus random pairings of stimulus and response. $N=11\,468$ random pairs, $N=11\,666$ true pairs.

We are currently preparing to substitute DECtalk with an online lexicon [Seitz (in preparation)] with which phonetic detail in the transcriptions can be controlled, including information regarding lexical stress. We are also investigating visual speech perception of phonemes in the major phonetic contexts of English. Depending on the outcomes of this study, the comparator will be modified as well as the level of phonetic detail in the transcriptions.

ACKNOWLEDGMENTS

This research was supported by the following NIH Grants: DC00023, NS22183, and DC00695. We wish to acknowledge the helpful critiques provided by Drs. Philip F. Seitz, Edward T. Auer, Louis D. Braida, Kenneth W. Grant, and Brian E. Walden.

[1]In this paper, phonological notation corresponds to that used by the DECtalk text-to-speech system, which was used to obtain all of the transcriptions.

Benguerel, A. P., and Pichora-Fuller, M. K. (1982). "Coarticulation effects in lipreading," J. Speech Hear. Res. **25**, 600–607.

Bernstein, L. E. (1993). "Sequence comparison techniques can be used to study speech perception," presented for the Committee on Hearing, Bioacoustics, and Biomechanics in a symposium, *Speech Communication Metrics and Human Performance*, National Academy of Sciences Auditorium, Washington, DC, 3–4 June.

Bernstein, L. E., and Demorest, M. E. (**1993**). "A general theory of speech perception must account for speech perception without audition," in Program 34th Annual Meeting The Psychonomic Society, p. 645.

Bernstein, L. E., and Eberhardt, S. P. (**1986**). Johns Hopkins Lipreading Corpus I–II: Disc 1 (Johns Hopkins University, Baltimore, MD).

Bernstein, L. E., Eberhardt, S. P., and Demorest, M. E. (**1989**). "Single-channel vibrotactile supplements to visual perception of intonation and stress," J. Acoust. Soc. Am. **85**, 397–405.

Davis, H., and Silverman, S. R. (**1970**). *Hearing and Deafness* (Holt, Rinehart and Winston, New York).

Demorest, M. E., and Bernstein, L. E. (**1991**). "Computational explorations of speechreading," J. Acad. Rehab. Aud. **24**, 97–111.

Demorest, M. E., and Bernstein, L. E. (**1992**). "Sources of variability in speechreading sentences: A generalizablity analysis," J. Speech Hear. Res. **35**, 876–891.

Eberhardt, S. P., Bernstein, L. E., Demorest, M. E., and Goldstein, M. H., Jr. (**1990**). "Lipreading sentences with single-channel vibrotactile transformations of voice fundamental frequency," J. Acoust. Soc. Am. **88**, 1274–1285.

Educational Services Department, Digital Equipment Corporation. (**1984**). *DECtalk DTC01 Programmer Reference Manual* (Digital Equipment Corporation, Maynard, MA).

Fisher, C. G. (**1968**). "Confusions among visually perceived consonants," J. Speech Hear. Res. **11**, 796–804.

Kruskal, J. B. (**1983**). "An overview of sequence comparison," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. B. Kruskal (Addison-Wesley, Reading, MA).

Kruskal, J. B., and Sankoff, D. (**1983**). "An anthology of algorithms and concepts for sequence comparison," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. B. Kruskal (Addison-Wesley, Reading, MA).

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–52.

Montgomery, A. A., and Jackson, P. L. (**1983**). "Physical characteristics of the lips underlying vowel lipreading performance," J. Acoust. Soc. Am. **73**, 2134–2144.

Sankoff, D., and Kruskal, J. B. (**1983**). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, Reading, MA).

Scheinberg, J. C. S. (**1988**). "An analysis of /p/, /b/, and /m/ in the speechreading signal," Ph.D. dissertation, City University of New York.

Seitz, P. F. (**in preparation**). "A polylectal lexicon for research on North American English auditory word recognition."

SPSS, Inc. (**1986**). SPSS^{x} *User's Guide* (McGraw-Hill, New York).

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J. (**1977**). "Effects of training on the visual recognition of consonants," J. Speech Hear. Res. **20**, 130–145.

Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.

Wozniak, V. D., and Jackson, P. L. (**1979**). "Visual vowel and diphthong perception from two horizontal viewing angles," J. Speech Hear. Res. **22**, 354–365.