Judgments of Phoneme Errors Under Four Modes of Audio-visual Presentation

Jugements sur les erreurs phonémiques dans quatre modes de présentation audio-visuelle

James C. McNutt
School of Human Communication Disorders
McGill University, Montréal, Québec

Linda Wicki
Terry Fox School
Montréal, Québec

Judith Paulsen
Porcupine General Hospital
Timmins, Ontario

Key Words: Phoneme errors, phonetic transcription, variance, reliability, audio, video, soundfield, headphones

Abstract

The purpose of this investigation was to examine the reliability of judgments of phoneme errors between two modes of recorded stimulus presentation (audio and video) under two conditions (soundfield and headphones). The number of judged phoneme errors were found to be related to the phoneme itself as well as to the audio-visual mode of presentation. Overall, the use of video presentations resulted in fewer judged phoneme errors and less variance, while the use of the audio mode resulted in a greater number of judged phoneme errors and greater variance. No significant differences were found between the headphone and soundfield conditions for either audio or video presentations.

Résumé

Cette enquête a pour objet d'étudier la fiabilité des jugements sur des erreurs phonémiques faites à cause de deux modes d'enregistrement (champ sonore et avec des casques à écouteurs). Le nombre d'erreurs phonémiques jugées a été relié tant au phonème lui-même qu'au mode de présentation audiovisuelle. Dans l'ensemble, l'utilisation de présentations vidéos a suscité moins d'erreurs phonémiques jugées et moins d'écarts, tandis que l'utilisation du mode audio a provoqué un plus grand nombre d'erreurs phonémiques jugées et un plus grand écart. On a constaté aucune différence significative entre la condition du casque à écouteurs et le champ sonore pour les présentations audios et vidéos.

Introduction

Perceptual judgments of phoneme errors in children's speech samples are frequently the basic data for research, assessment, and remediation of speech-sound errors. However, there are variables that may affect the reliability of such phoneme-error judgments. "Reliability," as defined by McCauley (1989), "is the degree to which a measure will be consistent when the time of measurement, the person doing the measuring, or some other variable changes." (p.29) Variables affecting the reliability of phoneme-error judgments can be divided into two broad categories: intrinsic factors, or those relating to the child's consistency in performance; and extrinsic factors, or those related to the examiner's consistency in judgement.

Intrinsic factors are concerned with the variability of the child's speech found in repeated samples. Children vary in their productions for a variety of reasons. Research in this area has focussed on determining the nature of this variability (Costely & Broen, 1976; Curtis & Hardy, 1959; Diedrich & Irwin, 1970; Stephens & Daniloff, 1977).

Extrinsic factors relate to the three following general areas. First, sampling factors such as the phonetic and linguistic context may result in varying productions and appear to influence the reliability of the transcription (Faircloth & Faircloth, 1970; Kent, 1982; McDonald, 1964). Second, personal characteristics of the examiner such as attentiveness, dialect, and relationship with the child have all been shown to affect the examiner's transcriptions of the speech of the child (Shriberg & Kent, 1982; Paynter & Edwards, 1977). Finally, the audio-visual conditions under which the stimuli are presented for transcription may have a differential effect on transcriber judgments. This later factor may relate to the acoustic nature of the target phoneme (Stephens & Daniloff, 1977) as well as to other perceptual aspects of examiner performance (Scharf, 1968; Shriberg, 1972; Irwin, 1977; Hoffman, 1976; Shriberg & Kent, 1982). The present investigation will focus on the extrinsic effects of audio-visual recordings on the listener's judgement of phoneme errors.

Frequently, limitations prevent phonetic transcription or judgement of phoneme errors to be made from speech samples of subjects as the stimuli are spoken. As a result, recording these stimuli with audiotape or videotape for later transcription or error judgement has become standard practice. However, the use of audio or video recordings may also have an effect on the judgement of listeners for a number of reasons. Stimuli may be degraded by electronic recording

and presentation. Similarly, visual cues which are present in live or videotaped speech may not be available when presented through audio tape only. Alternatively, the use of headphones may focus attention on the acoustic signal and mask ambient noise.

The reliability of listener judgments has been examined in some clinical populations (Darley, Aronson, & Brown, 1969a,b; Kearns & Simmons, 1988). However, investigations of the effect of different modes of audio-visual presentation appear limited to overall judgments of phoneme errors. Comparisons of past investigations to determine the effects of audio versus video presentation or the effects of the use of soundfield versus headphones are difficult for numerous reasons. In some investigations, judgments have been made from the soundfield presentation of stimuli (Hoffman & Shuckers, 1978), while in others the presentation has been through headsets (Shriberg, 1972; Stephens & Daniloff, 1977). Additionally the linguistic unit for presentation has varied and has included the use of nonsense syllables (Wright, 1954), words (Shriberg, 1972), and sentence length materials (Stephens & Daniloff, 1977; Hoffman & Shuckers 1978). In many of these investigations, transcriptions of "live" speech have been used for comparison despite the problems of intrinsic variability (Elbert, Shelton, & Arndt, 1967; Hoffman & Schuckers, 1978; Shriberg, 1972; Stephens & Daniloff, 1977; Wright, 1954). No single investigation has included both audio and video presentations of stimuli under both soundfield and headphone conditions.

The phoneme in question as well as the mode of presentation may affect the consistency of phoneme-error judgments. Judgments made from tapes of /r/ are typically more reliable than tapes of /s/ (Diedrich & Bangert, 1980). When comparisons between live and audiotaped presentation are made, judgments of /s/ are significantly less reliable under audiotaped conditions, while judgments of /r/ are approximately the same under both conditions (Brungard, 1961; Elbert, Shelton, & Arndt, 1967; Hoffman & Schuckers, 1978; Shriberg, 1972; Stephens & Daniloff, 1977). Although the reliability of error judgments under different conditions may be related to the phoneme in question, there are no data available for this effect other than for /s/ and /r/ (Shriberg & Kent, 1982). No investigation has examined the effect of different audio-visual conditions across a variety of phonemes.

In evaluating the effects of different modes of stimulus presentation there are several measures that may be considered. The first and most common measure is to compare the number of judged errors under the differing conditions. The second method used to evaluate the effect of differing modes of stimulus presentation is to examine the variance in judgments of errors under different conditions. If little variance is present, then the listeners are in agreement on their judgments. (See Shriberg, Kwiatkowski, & Hoffman [1984] for an excellent discussion of validity and reliability concerns in phonetic transcriptions.) The importance of considering variance in connection with reliability has also been noted by Stevens & Daniloff (1977).

The present investigation was designed to determine whether differences exist in either the number or the variances of judged phoneme errors phonetically transcribed under four different listening conditions. The four listening conditions were: (1) audiotape-with-headphones; (2) audiotape-in-soundfield; (3) video-with-headphones; and (4) video-in-soundfield. Judgments of correct and error productions were made from broad phonetic transcriptions of five different phonemes - /r/, /l/, /s/, /z/, and /Θ/.

Method

Subjects

The judges/listeners were 14 female and 2 male graduate students (mean age: 24.7 years) majoring in human communication disorders. Although their backgrounds differed, all were completing a graduate course in phonetic transcription that included both disordered and normal speech. Those 16 judges with the highest intra-judge reliability in the phonetics course were selected for this investigation. The judges selected exhibited mean point-to-point, intra-judge reliability of 89.8%, as determined previously from two audiotaped, soundfield recordings of a *Screening Deep Test of Articulation* (SDTA), (McDonald, 1976) in which both errors and correct tokens were presented for all 9 phonemes. This figure agreed favorably with the 90% figure for acceptable intrajudge reliability suggested by Bernthal and Bankson (1988). Each listener passed a hearing screening at 15 dB HL (American National Standard Institute, 1969) at octave frequencies from 500-4000 Hz.

Stimuli

The speech stimuli were taken from 4 children (aged 4,9; 4,3; 4,3; and 5,1) with normal phoneme errors for their ages. The stimuli used in the four listening conditions were obtained from the children repeating, after the examiner's live model, 18 wordpairs that contained 10 different samples for each of the five target phonemes /s/, /z/, /Θ/, /l/ and /r/ (listed in Appendix A). These five phonemes were chosen because they are a common source of error for children; additionally, /s/ and /r/ were chosen because they have been examined in previous research of this sort. Target phonemes were repeated as word pairs (CVC-CVC) with each phoneme occurring an equal number of times in word initial/prevocalic and word

final/postvocalic positions. Wordpairs were used to more closely approximate continuous speech and to permit a number of errors in each presentation (McDonald, 1964). All children displayed normal speech production errors for their ages and, although there were variations in degree of articulatory status, the speech of the children was highly intelligible. Each of the 16 listeners judged the 200 stimuli (5 phonemes x 10 samples each x 4 children) under each of the four listening/viewing conditions. As a result 12,800 judgments were examined in the present investigation with 3,200 judgments under each listening condition.

Instrumentation

Stimuli for both audio and video presentations were prepared simultaneously from live speech samples recorded in a quiet room in a day-care centre. Stimuli for audio conditions were prepared with a Yamaha TC 800D cassette tape recorder and an Aiwa DM-85 dynamic microphone. Stimuli were recorded on TDK SA audio tapes. Stimuli for video conditions were prepared using a Panasonic WV-3320 camera with a microphone to record stimuli on 1/2 inch Panasonic VHS video cassette tapes. The microphone distances were approximately six inches from the speaker's mouth. Conditions were similar to those frequently used by clinicians when recording speech samples.

Judgement Procedures

The phonetic transcriptions of errors were performed in 4 one-hour blocks, each of which was separated by a period of 7 days. Conditions and individual children were presented in a counter-balanced order to the four groups of listeners. Listeners were instructed to transcribe phonetically only those sounds within brackets on the scoring form (see Appendix A). Transcription was done using conventional IPA symbols for the target phonemes. Only one stimulus presentation was given. Following the transcription, the investigators totalled the number of phonemes transcribed as errors. An error was defined as any substitution, distortion, or omission of the target phoneme.

Stimuli were presented to listeners under the four conditions in the following manner. Listeners were divided into 4 groups. In the headphone condition for both audio and video recordings a splitter box was used with four Koss HUX dynamic stereo headphones (listed range: 75 Hz to 35 kHz) to ensure comparable presentation to each of the four listeners. The stimuli were presented at a comfortable listening level. Specific procedures are detailed for the following four audio-visual conditions.

In the first condition, audio-with-headphones, stimuli were presented to listeners using a Yamaha TC 800D cassette tape player with headphone procedures and splitter box as described above. In the second condition, audio-soundfield, stimuli were presented with a Yamaha TC 800D cassette tape player coupled to an Ampex 620 amplifier speaker (117 v.a.c. 5 amps). In the third condition, video-with-headphones, a Panasonic omnivision VHS System and Sony Trinitron (24 inch screen) were used to present the video recordings of the children during testing along with the headphones and splitter box procedures as described for the audiotaped presentation. In the fourth condition, video-soundfield, video equipment (the same used in the video-with-headphones condition) was used with the audio portion presented through the video speaker of the Sony Trinitron. In both soundfield conditions the listeners were approximately four feet from the speakers.

Data was grouped by condition and by phoneme. Differences in error scores were evaluated by analysis of variance and the Newman-Keul's post-hoc procedure where appropriate (SAS Institute, 1989). Differences in variance were evaluated by F-Max tests (Glass & Stanley, 1970).

Results

Differences in Mean Scores

Table 1 lists the mean and standard deviation of judged errors for the 16 listeners for each phoneme under each of the four listening conditions. A repeated measures analysis of variance with the three factors of listening condition x child x phoneme revealed significant main effects for listening condition [F (3,79)=14.17, p <0.0001]; child [F (3,79)=98.69, p <0.0001], and phoneme [F (4,79)=283.59, p <0.0001] as well as significant interactions between child and listening condition [F (9,79)=2.44, p <0.0095], condition and phoneme [F (12,79)=2.84, p <0.0007], and child and phoneme [F (12,49) = 49.14, p <0.0001]. No significant child x phoneme x condition interaction was found.

Post-hoc analysis by Newman-Keul's procedures determined a significantly greater number of judged errors (p <0.01) in audio-with-headphones versus video-soundfield conditions, audio-soundfield versus video-soundfield conditions, audio-soundfield versus video-with-headphones conditions, and audio-with-headphones versus video-with-headphones conditions. No significant differences were discovered between video-with-headphones and video-soundfield conditions or between audio-with-headphones and audio-soundfield conditions.

Differences between conditions for individual phonemes were determined to be significant (p <0.01) by the Newman-Keul's procedure only for the phoneme /z/. A significantly

Table 1. The mean number of errors and standard deviations (*SD*) determined for each phoneme (and total) by 16 listeners under each of the four listening conditions.

Phoneme	AH	A	VH	V
	MEAN (*SD*)	MEAN (*SD*)	MEAN (*SD*)	MEAN (*SD*)
/s/	5.6 (3.2)	5.8 (2.7)	5.8 (3.4)	5.3 (3.5)
/z/	18.8 (7.3)	17.8 (7.9)	12.3 (4.9)	12.4 (5.4)
/r/	1.5 (1.5)	1.1 (1.8)	0.4 (0.5)	1.1 (1.4)
/l/	2.6 (2.2)	1.9 (1.1)	1.9 (1.8)	1.3 (1.7)
/Θ/	15.7 (4.5)	15.3 (3.8)	33.4 (8.1)	12.6 (3.2)
Total	44.2 (12.5)	41.9 (10.2)	33.4 (8.1)	32.9 (7.5)

AH Audio-with-headphones
A Audio-soundfield
VH Video-with-headphones
V Video-soundfield

greater number of errors for /z/ occurred in the audio-with-headphones condition than in the video-soundfield or the video-with-headphones conditions. Significantly more errors also were judged for /z/ in the audio-soundfield condition than in the video-with-headphones or the video-soundfield conditions. Significant differences were not found between headphone and soundfield conditions for either the audiotaped or videotaped recordings.

Differences in Variance

The results of *F*-tests that compare variances of judged error scores across the four audio-visual conditions indicated that when scores for all phoneme judgments are grouped the only significant contrast (F=2.77, p <0.05) showed greater variance under the audio-with-headphones condition than under the video-soundfield condition. Differences in variance between other conditions when all phonemes were grouped were not significant.

Significant differences (p <0.01) in variances between conditions were found for judged errors of /l/ and /r/. Significant differences in variance were observed for /l/ between the audio-with-headphones and the audio-soundfield conditions (F=4.09, p <0.01), and between the video-with-headphones and the audio-soundfield conditions (F=2.54, p <0.05). Significant differences (p <0.01) in variance for transcriptions of /r/ were found between the video-with-headphones condition and each of the other three conditions (audio-soundfield, F=8.48; audio with-headphones, F=5.73; video-soundfield, F=4.79).

Discussion

In selecting the children and the phonemes that comprised the speech sample, our intention was to use children who differed in the number and type of phoneme errors. These selection differences are reflected in the main effects of (1) child and (2) phoneme and in the interaction between child and phoneme. Significant differences in these areas showed that the samples used in the present investigation did indeed have differences in the number and type of phoneme errors.

The purpose of the present investigation was to evaluate the effects of four listening conditions on listener judgments of phoneme errors. Results have shown that the mode of audio-visual presentation had a significant effect upon the judgement of phoneme errors and that these effects varied with the phoneme examined.

A surprising finding was the nature of the differences that were dependent upon the mode of presentation. A greater number and variance of judged errors were found to exist under the audio conditions, while fewer errors and less variance were observed under the video conditions. These findings could be taken to indicate that the audio condition is more sensitive (as opposed to video) because a higher number of errors were identified under this condition. However, a high error score may indicate that the audio condition perceptually distorted some correct productions, resulting in an increase in error judgments. For opposite reasons, the lower error scores for the video condition may have resulted from some masking of the real errors in production.

Similar precautions apply to the interpretation of variance measures. The audio-video condition under which the transcription took place could increase or decrease the difficulty in arriving at a decision. Stephens and Daniloff (1977) have commented "that tape recordings can alter the perceptual qualities of misarticulated /s/ in such a way that reliability is reduced." (p. 219) The altered perceptual qualities of /s/ and other sounds may effect the results in two different ways. If the condition increased the difficulty of arriving at a decision, greater variance would be anticipated under these conditions. If the condition decreased the difficulty (possibly masking the error), the variance would be reduced. There is no way to determine the true errors in this situation unless one arbitrarily (or collectively through a process of consensus) determines that judgments under one condition (including a live condition) are "real." The findings in the present investigation underline the importance of examining both the number of judged errors and the variances when considering the reliability of listener judgments.

Overall (with combined phonemes) the listening conditions had the same effect on variance as on the number of

perceived errors. As the number of judged errors increased, so did the variance of these judged errors. Such relationships have been noted previously in relation to differences between live and audiotaped conditions. Lower reliability values and a wider range of scores (greater variance) were found for judgments made under audio conditions than under live conditions (Stephens & Daniloff, 1977).

The differences among conditions for individual phonemes were significant only for /z/ (for number of errors) and /r/ and /l/ (for variance). However, examination of Table 1 shows that 4 of the 5 phonemes under examination followed a similar trend. The phoneme /s/ was the only phoneme that did not reflect this trend. Previous investigations have found that the reliability of /s/ varies with the condition of presentation (Stephens & Daniloff, 1977), while that of /r/ does not (Brungard, 1961; Hoffman & Schuckers, 1978). Previous investigations of /s/ or /r/ error judgments also suggest that there is a lower reliability for /s/ than for /r/ (Elbert, Shelton, & Arndt, 1967; Hoffman & Schuckers, 1978; Stephens & Daniloff, 1977). Data from this investigation does not readily permit a comparison within conditions, however, the present measures related to /s/ appear stable, while measures of /r/ show significant variability. The relative reliability of measures of /s/ and /r/ warrant continued investigation considering their high incidence of error in phonologically disordered children.

Another surprising finding for the combined phoneme data relates to headphone use. Significant differences were not found in the number or variances of judged errors between audio-with-headphones and audio-soundfield conditions or between video-with-headphones and video-soundfield conditions. This finding runs counter to the supposition that the use of headphones results in more reliable (i.e., less variable) judgments of phoneme errors. The use of headphones was expected not only to provide a higher quality signal than soundfield, but also to reduce the signal-to-noise ratio during playback of the stimuli. As a result the headphone conditions were expected to yield less variance than the soundfield conditions. However, no such effect was found.

Judgments of phoneme error were used in this investigation as the basis for measurement. Such two-way judgments (Shriberg & Kent, 1982) result in the smallest number of categories to be differentiated. The use of such judgments should result in the highest agreement both within and between judges. The results of this investigation provide a framework and data base for comparison with future research. Further investigation is needed to examine the effects of varying modes of audio-visual presentation upon the judgments of smaller categories such as those from narrow phonetic transcription.

Acknowledgments

We wish to express our appreciation to Donald Doehring and Gloria Waters, School of Human Communication Disorders, McGill University, for their help in this investigation and to Michael Walsh, McGill Computer Centre, for his assistance with statistics and analysis.

Address all correspondence to: James C. McNutt, Ph.D.
School of Human Communication Disorders, Beatty Hall, McGill University, 1266 Pine Avenue West, Montréal, Québec H3G 1A8

References

American National Standard Institute (1969). *Specifications for audiometers* (ANSI S3.6-1969). New York: ANSI.

Brungard, M.(1961). Effect of consistency of articulation of /r/ and /s/ on gains made with and without therapy. Unpublished Ph.D. dissertation, Pennsylvania State University, 1961. Cited in P.R. Hoffman & G.H. Schuckers, (1978). Audio-recording effects upon judgement reliability of children's /r/ misarticulation. *Perceptual and Motor Skills, 47*, 455.

Bernthal, J.E., & Bankson, N.W., (1988). *Articulation and Phonological Disorders, 2nd edition*. Englewood Cliffs, NJ: Prentice-Hall.

Costely, M., & Broen, P. (1976). The nature of listener disagreement in judging misarticulated speech. Paper presented at the American Speech and Hearing Association National Convention, Houston, Texas (November 1976). Cited in L. Shriberg and R. Kent, (1981) *Clinical Phonetics*. New York: John Wiley and Sons, Inc., 164.

Curtis, J., & Hardy, J. (1959). A phonetic study of misarticulations of /r/. *Journal of Speech and Hearing Research, 2*, 244-257.

Darley, F.L., Aronson, A., & Brown, J.R. (1969a). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research, 12*, 462-497.

Darley, F.L., Aronson, A., & Brown, J.R. (1969b). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*, 246-269.

Diedrich, W., & Bangert, J. (1980). *Articulation Learning*. Houston, Texas: College-Hill Press.

Diedrich, W., & Irwin, J. (1970). Training speech clinicians in the recording and analysis of articulatory behaviour. Summary Report, Year 1, U.S. Office of Education Grant, November.

Elbert, M., Shelton, R.L., & Arndt, W.B. (1967). A task for evaluation of articulation change: 1. development of methodology. *Journal of Speech and Hearing Research, 10*, 281- 288.

Faircloth, M.A., & Faircloth, S.R. (1970). An analysis of the articulatory behaviour of a speech defective child in connected speech and in isolated-word responses. *Journal of Speech and Hearing Disorders, 35*, 51-61.

Glass, G.V., & Stanley, J.C., (1970). *Statistical Methods in Education and Psychology*. Englewood Cliffs, NJ: Prentice Hall, Inc.

Hoffman, P. (1976). A comparison of judgment reliability and validity: live versus tape-recorded /r/ misarticulation in sentence con-

texts. Paper presented to the American Speech and Hearing Association National Convention, Houston, Texas.

Hoffman, P.R., & Schuckers, G.H. (1978). Audio-recording effects upon judgment reliability of children's /r/ misarticulation. *Perceptual Motor Skills, 47*, 451-456.

Irwin, R.B. (1977). Variability of phonemic perceptual abilities of clinicians. Paper No. 73, p. 29, XVI Int'l Cong. Logopedics Phoniatrics, Interlaken, Switzerland, Aut. 25, 1974. cited in M.I. Stephens & R. Daniloff, A methodological study of factors affecting the judgments of misarticulated /s/. *Journal of Communication Disorders, 10*, 218.

Kearns, K. P., & Simmons, N. N. (1988). Interobserver reliability and perceptual ratings: more than meets the ear, *Journal of Speech and Hearing Research, 31*, 131-136.

Kent, R.D. (1982). Contextual facilitation of correct sound production. *Language, Speech, and Hearing Services in Schools, 13*, 66-76.

McCauley, R. J. (1989). Measurement as a dangerous activity. *Journal of Speech-Language Pathology and Audiology. 13*, (1), 29-32.

McDonald, E. (1964). *Articulation Testing and Treatment: A Sensory-Motor Approach*. Pittsburgh: Stanwix House, Inc.

McDonald, E. (1976). *A Screening Deep Test of Articulation*. Toronto: J.M. Dent & Sons, Ltd.

Paynter, E.T., & Edwards, W.S. (1977). Effects of examiner race. *Journal of Communication Disorders, 10*, 221-229.

SAS Institute Inc. (1989). *SAS User's Guide*. Box 8000, Cary, North Carolina.

Scharf, D. (1968). Distinctiveness of "defective" fricative sounds. *Language and Speech, 11*, 38-45.

Shriberg, L.D. (1972). Articulation judgments: some perceptual considerations. *Journal of Speech and Hearing Research, 15*, 876-882.

Appendix A

Word pairs used in children's repetitions and listeners' judgements. Each of the phonemes /s/, /z/, /Θ/, /l/, and /r/ occurred 5 times in word initial and 5 times in word final positions.

1. south ear
2. rose leaf
3. zoo lace
4. ear bus
5. teeth rail
6. size ball
7. thumb miss
8. roll zone
9. think more
10. lice zoo
11. lose thief
12. lies mouth
13. zap sour
14. rice thumb
15. think rays
16. sore tooth
17. zip nail
18. mail teeth

Shriberg, L.D., Kwiatkowski, J., & Hoffman, K. (1984) A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research, 27*, 456-465.

Shriberg, L.D., & Kent, R. (1982). *Clinical Phonetics*. New York: John Wiley & Sons, Inc.

Stephens, I., & Daniloff, R. (1977). Trouble with /s/. A methodological study of factors affecting the judgement of misarticulated /s/. *Journal of Communication Disorders, 10*, 207-220.

Wright, H. (1954). Reliability of evaluation during basic articulation and stimulation testing. *Journal of Speech and Hearing Disorders, 19*, 5-19.