Journal of Speech and Hearing Research, Volume 27, 456–465, September 1984

Research Note

A PROCEDURE FOR PHONETIC TRANSCRIPTION BY CONSENSUS

LAWRENCE D. SHRIBERG JOAN KWIATKOWSKI KIT HOFFMANN
University of Wisconsin–Madison

A consensus transcription procedure is presented, and representative data are reported. The procedure includes explicit response definitions for mapping narrow phonetic transcription onto dichotomous, nonerror/error scores and 17 rules for obtaining consensus between two transcribers. The data include descriptive statistics for consensus rule use based on over 18,000 sounds transcribed during a normative study of speech sound acquisition in children aged 3–6 years. Validity and reliability issues in phonetic transcription are discussed.

Valid and reliable phonetic transcription is central to the study and management of persons with communicative disorders.

Validity

Three validity concerns may be raised whenever perceptual judgments are made about speech and particularly when dealing with disordered speech. First, even under optimal live listening conditions, phonetic transcription data may not agree with physiologic or acoustic records. Weismer (1984), for example, presented evidence that phonetic transcription may be inconsistent with acoustic records, and Oller and Eilers (1975) demonstrated expectancy effects associated with lexical alternatives. Second, when transcription is done from recorded materials, with or without a video signal, perceptual judgments of some behaviors may disagree both with instrumental records and with descriptions made under live listening conditions (Daniloff, Wilcox, & Stephens, 1980; Hoffman & Schuckers, 1978; Stephens & Daniloff, 1977). Although technical improvements in audio and video recording devices have been extensive, and explicit procedures to maximize the efficiency of phonetic transcription have been proposed (Shriberg & Kent, 1982), the validity of both live transcription and transcription from recorded materials remains a challenging methodological concern. Current trends are to supplement auditory perceptual data with on-line or hard copy acoustic data (typically, spectrograms) to maximize the validity of phonetic transcription (see, e.g., Kent, 1981).

A third validity issue in phonetic transcription concerns the relationship between phonetic transcription and judgments of "correct" versus "incorrect" speech sounds. Whereas phonetic transcription presumably only describes speech events, judgments about the correctness or social acceptability of speech sounds involve subjective estimates of normalcy. Consider what occurs when a narrow phonetic transcript of sounds produced by a speech-delayed child is to be dichotomized into correct versus incorrect speech sounds. Phoneme deletions and phoneme substitutions (whether based on acoustic or perceptual data) unarguably are speech production "errors." However, not all phonetic distortions are considered speech errors. For example, a dentalized /s/, [s̪], is generally considered a socially relevant speech error, but a dentalized /d/, [d̪], typically is not. The concept of validity in this situation refers to what the lay person considers to be a speech error, that is, what one would perceive as distracting or "defective" speech. Beyond early sociometric studies of the effects of speech errors in the school milieu (e.g., Colby, 1944; Freeman & Sonnega, 1956; Kleffner, 1952; Perrin, 1954) and recent simulation studies of dentalized (Mowrer, 1978) and lateralized (Silverman, 1976) /s/ errors in adults, few textbooks have suggested guidelines on which classes of articulatory distortions should be judged incorrect (however, see Van Riper & Irwin, 1958, for an early treatment of this issue).

The first two validity concerns discussed above, associations between any type of perceptual data and instrumental records, require considerations beyond the scope of this note. The third validity issue, the question of response definitions, seems timely to address. Although phonetic transcription continues to be widely used in contemporary child phonology research, there is little agreement on transcription conventions or response definitions. As computer technology allows phonetic transcription to be recoded into correct/incorrect scoring decisions (Shriberg, 1982), it becomes necessary for researchers to develop and commit to a set of explicit response definitions.

Table 1 includes two lists of narrow phonetic transcriptions. List A includes diacritics that are not considered to be "errors," and List B includes those that are considered to be errors. The nonerror diacritics were included in List A because they meet one or more of the following three criteria: (a) They symbolize optional allophones in casual speech forms (e.g., unreleased stop [p˺]), (b) they are difficult to transcribe reliably (Hoffmann, 1982; Shriberg & Kent, 1982; Shriberg & Kwiatkowski, 1980), and (c) they presumably would not be perceived by the lay person as articulation errors (e.g., *bat* [bæ̃t]). Error diacritics included in List B, in contrast, (a) symbolize nonop-

0022-4685/84/2703-0456$01.00/0

TABLE 1. Response definitions for mapping narrow phonetic transcriptions onto nonerror/error phonemic distortion decisions.

List A diacritics—Nonerror			*List B diacritics—Error*		
Term	*Symbol*		*Term*	*Symbol*	
Nasalized vowel	[̃]		Nasal emission	[̃̃]	on oral consonants
Denasalized consonant	[ⁿ]		Labialized	[ʷ]	glides and liquids only
Rounded vowel	[ᵓ]				
Nonlabialized consonant	[ᵚ]		Dentalized	[̪]	sibilants and bilabials only
Dentalized	[̪]	not including sibilants	Lateralized	[ˡ]	
Palatalized	[ʲ]		Retroflexed	[˞]	
Whistled	[ˎˏ]		Velarized	[̴]	prevocalic and intervocalic /l/ only
Glottalized	[ˀ]				
Breathy	[ʱ]		Derhotacized	[˞]	
Frictionalized	[ˣ]		Unaspirated	[ʰ]	only when aspiration is obligatory
Weakly articulated[a]	[√]				
Lengthened	[ː]		Consonant additions	[^{xy}]	unless considered a nonlinguistic behavior
Shortened	[ˑ]				
Partially voiced	[̬]				
Partially devoiced	[̥]				
Unreleased	[̚]				
Aspirated	[ʰ]				

Note. From *PEPPER (Programs to examine phonetic and phonologic evaluation records) users manual* by L. Shriberg, 1982, Madison, WI: Waisman Center on Mental Retardation and Human Development, Research Computing Facility. Copyright 1982 by the Waisman Center. Adapted by permission.

[a]All phonetic symbols follow the narrow phonetic transcription system described in Shriberg and Kent (1982), except [√] which is used here to symbolize *weakly articulated*, e.g., [t̲].

tional allophones, (b) are more reliably transcribed using only perceptual phonetics, or (c) presumably are included in the set of socially relevant articulation errors for which speech management services are provided. Note the previous example that contrasts these two response definition classes: the dentalized /d/ [d̪] (List A) is not considered a socially relevant speech error, whereas a dentalized /s/ [s̪] (List B) is considered a socially relevant speech distortion. Entries in these two lists of nonerror and error diacritics have emerged from our studies of normal speech sound acquisition (Hoffmann, 1982; Hoffmann & Shriberg, 1982), from studies of children with severely delayed speech development (Kwiatkowski & Shriberg, 1983; Shriberg & Kwiatkowski, 1982), and from associated literature in normal and delayed speech acquisition. These explicit response definitions were used in the study to be described in this paper.

Reliability

The distinction between a validity issue and a reliability issue is not clear when considering transcription data or correct/incorrect scoring data. In our view, the difference between a validity and a reliability concern is determined by the nature and degree of training of the transcribers/judges. If little or no shared discrimination training among transcribers or judges has taken place, their independent decisions about a speech segment would seem to represent a form of validity. That is, each transcriber or judge may be considered a different measurement instrument; the consensus decision on a segment is valid in the same way that qualified but different instruments yield the same measurement result. As before, this form of validity occurs when the lay person agrees with the speech-language pathologist or when two or more speech-language pathologists from different training backgrounds agree with one another. Data of this sort actually reflect consensual validity.

When transcribers or judges are intensively pretrained to make similar judgments about stimuli, however, their subsequent agreement on new stimuli reflects what is usually considered transcription reliability. The reliability of phonetic transcription is generally assessed in one of two ways. If only one transcriber's data are to be used, reliability is estimated by intrajudge and interjudge agreement checks. Interjudge reliability estimates typically reach .90 and above for transcription of segments

that are clearly correct or clearly incorrect (e.g., Irwin & Wong, 1983). For transcription or judgments of more difficult segments, typically those characterized as distortions, interjudge percentages of agreement that fall below .80 are common (Brungard, 1961; Diedrich & Bangert, 1980; Hoffman, 1983; Shriberg, 1972).

A second reliability procedure is to attempt to reduce measurement error by using consensus transcription. A consensus transcription from two or more transcribers presumably will reduce errors of measurement such as those that might occur when one transcriber is momentarily inattentive or has particular response biases. Consensus transcription has received little methodological study.

Three general problems are associated with consensus transcription. One practical problem is that it requires a heavy investment in personnel and in total transcription time per subject record. For many and perhaps most research and clinical situations, it is not feasible to enlist several well-trained persons to serve as independent transcribers. Hence, although the term consensus transcription suggests the outcome from a panel of judges, more typically it reflects the consensus opinion of only two transcribers—perhaps the author of a paper and a colleague.

A second problem concerns procedural conventions. Methodological descriptions of consensus transcriptions in the literature on normal and disordered speech and language typically report only that sample tapes were replayed until disagreements were resolved, with little or no information provided on the rules or processes used to reach consensus. Our experience in transcription of large data sets suggests that consensus transcription procedures invoke individual differences in transcribers' technical competence, professional status, and personality styles. Differences in competence, status, or personality may affect the intended reliability gains afforded by consensus transcription. That is, the consensus transcription may be biased to reflect the judgments of the more competent, higher ranked, or "forceful" transcriber.

A third problem with consensus transcription is that data can be biased by decisions made about how to treat segments on which consensus was difficult or impossible to obtain. Such data may be included with the other data, treated separately from other data, or excluded entirely from all data analyses. Depending on the research question and the proportional occurrence of segments on which consensus was difficult or impossible to obtain, such decisions may seriously affect the interpretation of results. Specifically, what might be discarded are subsections of data that reflect the very phenomena under study.

Table 2 is a set of consensus procedures generated specifically to address the problem of reducing measurement error when only two, equally competent transcribers are involved. This set of consensus rules was developed for a normative study of speech sound acquisition (Hoffmann, 1982) and has been used subsequently in a series of studies of children with delayed speech acquisition.

Rationale for the consensus transcription rules is provided in the introductory materials in Table 2. The consensus rules attempt to retain "conservative" consensus transcriptions for subsequent analyses whenever possible. Most of the rules suggest a consensus transcription that makes the least severe claim about a child's error type; for example, distortion errors are considered less severe than phoneme substitutions. Other rules suggest a consensus transcription based on probabilistic hierarchies, as determined from our prevalence data of distortion types in normal and speech-delayed children. Overall, given a disagreement between transcribers, the consensus transcription reflects a compromise based on acoustic-perceptual considerations, severity of error considerations, and prevalence of error type considerations. When a compromise transcription that achieves a conservative solution is not available, the consensus transcription is circled as questionable data. Notice that rule 1 is not a consensus rule in the same sense as the other 16 rules because disagreements are resolved immediately on replay. This rule addresses the problem of momentary inattention by one transcriber. The remaining 16 rules in Table 2 provide explicit decision rules for resolving transcription differences.

The primary purpose of this research note is to stimulate research interest in validity and reliability issues in phonetic transcription and to present some data on a procedure for consensus transcription. The procedures described in Table 2 were derived to address the three problems in consensus transcription described earlier. That is, the procedure requires only two transcribers, includes explicit rules for resolving disagreements between transcribers, and provides a principled approach to retain the maximum amount of the consensus data for phonetic/phonologic analyses.

SOME PRELIMINARY DATA

Method

The data for this note were gathered in the context of a normative study of speech sound development in 3- to 6-year-old children (Hoffmann, 1982). Continuous speech samples collected from 72 children consisted of approximately 5 min of free conversation during which a child talked about a color form picture he or she was constructing. All samples were obtained by the same examiner using a Marantz C-207LP audio cassette tape recorder, matching external microphone, and Sony LNX cassette tape. Lip-to-microphone distance was maintained at 15–30 cm. These methods yielded excellent quality tapes in which consonant allophone features such as aspiration were clearly audible and vowels were not distorted.

Prior to consensus transcription of the 72 original continuous speech samples by two of the authors, extensive transcription training was conducted among the three authors. Each author had completed an audiotape training program in narrow phonetic transcription of normal and disordered speech (Shriberg & Kent, 1982) before spending approximately 15 hr of additional practice with the first author.

TABLE 2. Procedures for consensus transcription.

Independent Transcription Procedures

1. The clinical response definition for independent transcription is to transcribe each segment as "incorrect unless heard as correct." That is, when unsure between two transcriptions, e.g., [t̰] vs. [t], [ɾ̰] vs. [ɾ], use the transcription that describes the segment as less correct.

2. Each transcriber transcribes the entire utterance.

3. Transcribers compare transcripts and derive a consensus transcription for each utterance, using all segments on which there is exact agreement. Consensus transcription for the remaining segments is obtained by following the procedures described below.

Consensus Procedures for Disagreements

1. Replay a maximum of three times the segments on which transcribers disagreed. During each replay, both transcribers attempt to "hear" the other's transcription, in addition to attempting to confirm his/her own transcription.

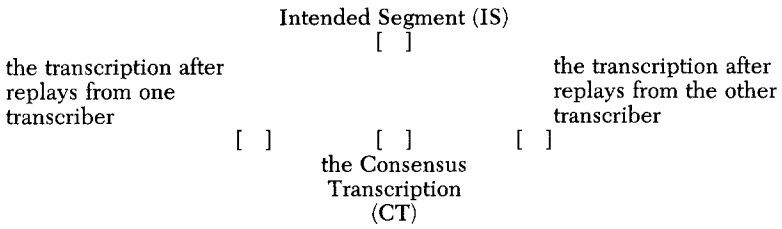
2. After the replays and second transcription, transcribers apply one of the 17 consensus rules listed below.

3. The goal of the 17 consensus rules is to give the speaker the "benefit of the doubt," given two transcriptions that disagree. Specifically, in each of the rules below the best approximation to normal is selected as the consensus transcription. The strict criterion, "incorrect unless correct," used in independent transcription is therefore balanced by the "benefit of the doubt" rule at the level of consensus transcription.

4. To separate questionable from unquestionable consensus transcription a circling convention is used. The 17 consensus rules are organized according to whether or not application of the rule will result in questionable data. There are three rule categories with respect to circling: *Never Circled* (Category I, rules 1–6), *Always Circled* (Category II, rules 7–13), and *Sometimes Circled* (Category III, rules 14–17). The arbitrary distinction among circled and noncircled consensus transcriptions is based on the cost of including possibly unreliable data versus the benefit of retaining possibly useful data for phonetic/phonologic analyses.

Consensus Rules

Several abbreviatory conventions are used for convenience in the 17 consensus rules below. The abbreviation *IS* is used for the *I*ntended *S*egment, that is, the correct segment or a correct segment with any optional allophones from List A. The abbreviation *CT* is used for the final *C*onsensus *T*ranscription, that is, the transcription that is ultimately entered in the transcript. The examples for each rule are arranged in the following format:



Rule	*Descriptive Label*	*Definition/Content*	*Example*
		Rules That Never Involve Circling	
1.	"Sure, I can hear it"	If one transcriber immediately "hears" and accepts the other transcriber's transcription as more accurate, select that transcription as the CT.	
2.	"Let's keep it simple"	If either transcription contains a List A diacritic (i.e., a nonerror diacritic), select as the CT the transcription that does not contain the nonerror diacritic.	[s] [s̻] → [s] ← [s] [k] [k̟̄] → [k̄] ← [k̄]
3.	"We both hear the diacritic"	If List A diacritics were used in both transcriptions the CT should contain only those included in both transcriptions.	[z] [z̪] → [z̪] ← [z̪]
4.	"Let's be conservative—nasals"	If optional nasal diacritics are used in both transcriptions, select as the CT the more conservative diacritic. The conservative hierarchy from most to least conservative is: [ⁿ] (e.g., [nⁿ]); [ᵑ] ([ᵑ] is used for *weak release*); [ᵐ].	[m] [m̃] → [m̥] ← [m̥]

Rule	*Descriptive Label*	*Definition/Content*	*Example*
		servative is: [˃] (e.g., [n˃]); [ˌ] ([ˌ] is used for *weak release*); [ˀ].	
5.	"Let's be conservative—stops"	If a final stop is perceived as unreleased in one transcription (i.e., is barely evident) and perceived as released in the other (with or without a List A diacritic), select as the CT the unreleased stop.	[d] [d̚] → [d̚] ← [d] [d] [d̚] → [d̚] ← [d̥]
6.	"Let's use the more probable diacritic"	If both transcriptions include List B diacritics, select as the CT the distortion highest on the distortion hierarchy. The distortion hierarchy is based on the frequency of error type, with the most frequently occurring error type highest on the hierarchy: /s/: [s̪], [s̻], [s̺]; /r/: [r̝], [r̠]; /l/ (prevocalic): [l̴], [l̪].	[s] [s̪] → [s̻] ← [s̺] [r] [r̝] → [r̠] ← [r̞] [l] [l̪] → [l̴] ← [l̴]
		Rules That Always Involve Circling	
7.	"Corrects outrank distortions"	If one transcription is a correct segment and the other a distortion (i.e., contains a List B diacritic), select as the CT the correct segment and circle it.	[z] [z] → (z) ← [z̪]
8.	"Corrects outrank deletions"	If one transcription is a correct segment and the other is a deletion, select as the CT the correct segment and circle it.	[l] [l] → (l) ← [ø]
9.	"Distortions outrank substitutions"	If one transcription is a phonemic distortion (i.e., contains a List B diacritic) and the other is a phonemic substitution, select as the CT the distortion and circle it.	[ɚ˞] [ɚ˞] → (ɚ˞) ← [ə]
10.	"Distortions outrank deletions"	If one transcription is a distortion and the other a deletion, select as the CT the distortion and circle it.	[l] [l̴] → (l̴) ← [ø]
11.	"Substitutions outrank deletions"	If one transcription is a substitution and the other is a deletion, select as the CT the substitution and circle it.	[l] [w] → (w) ← [ø]
12.	"Let's use the closest substitution"	If both transcriptions are substitutions, select as the CT the closest substitution to the IS (based on place/manner features) and circle it.	[θ] [f] → (f) ← [s]
13.	"Once circled, always circled"	If either (or both) of the transcribers has circled his/her transcription before or after replay, the CT also must be circled, even if transcribers agree on the segment. If the segments are different, apply the appropriate consensus rule and circle the CT. (Consensus Rule 9 also applies)	[s] (s) → (s) ← [s] [s] [s] → (s) ← (θ)
		Rules That Sometimes Involve Circling	
14.	"Sure, we both can hear it"	If (a) both transcribers can hear both transcriptions, (b) the transcriptions are functionally equivalent, and (c) neither transcription is a deletion, select as the CT the transcription that is closest (place/manner features) to the IS.	[ð] [d̪] → [ð] ← [ð] (or [ð]) (or [d̪]) [ð] [d̪] → [d̪] ← [d] (or [d]) (or [d̪]) [s] [s̻] → [s] ← [s̪] (or [s̻]) (or [s̻])

Rule	*Descriptive Label*	*Definition/Content*	*Example*
		Deletion Exception: If (a) both transcribers can "hear" both transcriptions, and (b) only one of the transcriptions is a deletion, select as the CT the nondeletion transcription and circle it.	[l] [ø] ← (w) ← [w] (or [w]) (or [ø]) [l] [ø] → (l) ← [l] (or [l]) (or [ø])
15.	"I hear yours, but you don't hear mine."	If (a) only one of the transcribers can "hear" both his/her own transcription and the other transcriber's, and (b) the transcription heard by both is not a deletion, select as the CT the transcription heard by both transcribers.	[h] [h] → [h] ← [ø] (or [h]) [z] [z̬] → [z̬] ← [z̬] (or [z̬])
		Deletion Exception: If the transcription both transcribers can hear is a deletion, select it as the CT and circle it.	[h] [h] → (ø) ← [ø] (or [ø])
16.	"We both hear something new"	If (a) on replay both transcribers hear the same transcription, one that is different from either of their original, and (b) neither original transcription was a deletion, select it as the CT.	[d] [d˞] [f] ↓ ↓
		Deletion Exception: If the new transcription is a deletion, circle it.	[ø] → (ø) ← [ø]
17.	"Let's find a good compromise"	If one transcription is the correct IS and the other is a substitution, select as a CT a transcription that is somewhere midway between the two transcriptions.	[ð] [d] → [d̪] ← [ð] [ɑ] [ɔ] → [ɑ̟] ← [ɑ] [l] [w] → [l̴] ← [l] [l] (final position) [o] → [l̴] ← [l]
		If a suitable compromise is not possible, select as the CT the correct transcription and circle it.	[m] [m] → (m) ← [w]

Figure 1 is a list of the 45 diacritic symbols and positional conventions that comprised the phonetic transcription system used in all transcription. These symbols were used to modify vowel, diphthong, and consonant symbols similar to those recommended in the International Phonetic Alphabet. The training procedure used for the 15 hr of additional practice required comparisons of independent transcriptions of normal and disordered speech samples to determine precisely where and why disagreements occurred and to generate tentative strategies to improve future discrimination of difficult response classes.

The two transcribers transcribed the 72 speech samples during daily sessions over a period of 6 weeks. Each daily session lasted 1–5 hr. A Dictaphone 2550 transcriber was positioned midway between the transcribers, with its speaker oriented perpendicular to and approximately 1 ft (30.5 cm) from each transcriber. The transcribers took turns operating a foot pedal control for replay of items as needed for the independent and consensus transcriptions. The general procedure was to transcribe an utterance independently and then to compare the two transcriptions. For each segment in an utterance on which transcribers disagreed in any way, up to three replays were used to reach consensus, as prescribed by the rules listed in Table 2.

Six weeks after the last tape was transcribed, the procedures used for the original consensus transcription were repeated on 25 utterances for each of eight randomly selected speech samples. Four weeks later, another eight speech samples were randomly selected for retest reliability assessment. Comparison of the original consensus transcriptions and the retest transcriptions yielded an overall exact agreement retest reliability of 68%. This

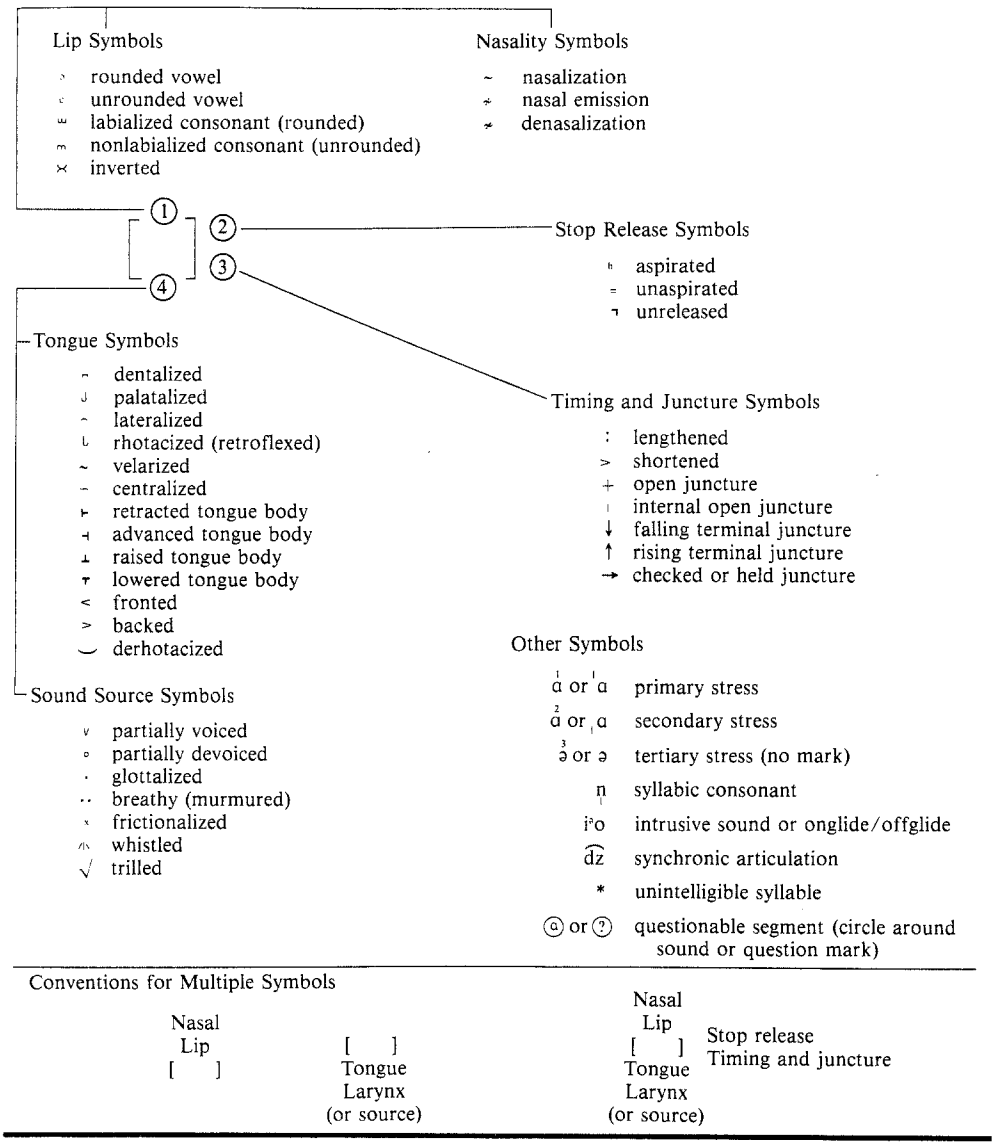


FIGURE 1. Diacritic marks for phonetic transcription. The numerals 1, 2, 3, and 4 show the placement of marks within a given category. For example, marks having to do with tongue position or adjustment are located under the phonemic symbol to be modified. The brackets surrounded by the numeral represent the position of the phoneme symbol (from Shriberg & Kent, 1982).

figure reflects an exact segment-by-segment comparison; additional interjudge and intrajudge analyses of diacritic-level agreement are provided in Hoffmann (1982). When computed with the disagreements on Table 1, List A diacritics removed, the overall exact percentage of agreement was 76%. As previously described, many of the diacritics listed in Table 1, List A, are among the most difficult to transcribe reliably.

For the purposes of the present study, one of the transcribers coded data on each consensus rule application while accomplishing the transcription of the 72 normative samples. Thirty of these 72 continuous speech samples were randomly selected for consensus transcription analyses. The focus in the present study is on how often each of the 17 consensus rules was used when the two transcribers disagreed and on the implications of these data for issues in consensus transcription.

Results and Discussion

Rule use. Table 3 includes summary statistics for the 30 continuous speech samples that were randomly selected for consensus transcription analyses. A total of 18,178 speech sounds were transcribed. Vowels and diphthongs constituted 41% of the phonetic data, and consonants constituted 59%. This ratio is consistent with values reported in several published statistical accounts of adult and child speech (Shriberg & Kent, 1982).

The findings for the consensus rule application are summarized in Tables 4 and 5. Table 4 includes the

TABLE 3. Descriptive statistics for the 30 continuous speech samples.

Sound class	*n*	*% of total*	*Glossed [a] segments per sample* $\bar{x}$	*SD*	*Range*
Vowels/ Diphthongs	7,393	41	246	30	191–299
Consonants	10,785	59	360	46	247–437

[a]Glossed segments are those for which transcribers could infer the child's intended segment, i.e., they were intelligible in the continuous speech context.

frequency of consensus rule use by the two transcribers. Percentages derived are based on the total number of segments in the 30 speech samples and on the percentage application of each rule relative to the total rule use. As shown in the totals at the bottom of Table 4, consensus rules were used more often on consonants than on vowels or diphthongs. For these normally developing children and within clinical disorders groups, errors on consonants are more frequent. The percentage totals indicate that consensus rules were needed for approximately 5% of all vowels/diphthongs transcribed (i.e., 363/7,393) and for approximately 16% of all consonants transcribed (i.e., 1,790/10,785). As shown in the last two columns by rule category and rule number, Rule 1 use accounted for approximately 90% of all consensus rule applications for vowels/diphthongs and approximately 81% for consonants. The remaining rules, therefore, accounted for approximately 10% of the vowel/diphthong rule applications and 19% of the consonant rule applications, respectively. Finally, the data in Table 4 indicate that relatively few rule uses involved circling or data that would not be retained for analysis. Overall, when the data generated by application of the 17 rules are added to the consensus transcriptions to which the rules did not have to be applied, 98% of the vowel data and 92% of the consonant data were retained for phonetic/phonologic analysis.

The data in Table 5 indicate the frequency of consensus rule use per sample. Rule 1 was used in each of the 30 samples, ranging 4–87 times per sample. With the exception of Rule 6, which was never used in these 30 samples, each of the other rules was used in 1–28 of the continuous speech samples 1–10 times per sample.

Additional Analyses of Rule 1 Use

The mnemonic phrase for Rule 1, *Sure, I can hear it,* reflects the altered perceptual experience that can occur when transcribers attempt to resolve disagreements by listening again to the segment in question. Additional analyses sought to understand the conditions under

TABLE 4. Frequency of consensus rule use by two listeners while transcribing 30 continuous speech samples. V = vowels/diphthongs, C = consonants.

Rule category	*Rule number*	*Total number of rule uses* V	C	V+C	*% rule use based on total number of segments* V	C	*% rule use based on total rule use* V	C
Never circled	1	326	1409	1735	4.4	13.1	89.8	79
	2	0	20	20	0	<1	0	1.1
	3	0	3	3	0	<1	0	<1
	4	0	1	1	0	<1	0	<1
	5	0	10	10	0	<1	0	5.6
	6	0	0	0	0	0	0	0
							89.8%	80.6%
Always circled	7	0	45	45	0	<1	0	2.5
	8	2	24	26	<1	<1	1	1.3
	9	2	2	4	<1	<1	1	<1
	10	0	3	3	0	<1	0	<1
	11	0	4	4	0	<1	0	<1
	12	0	9	9	0	<1	0	<1
	13	0	27	27	0	<1	0	1.5
							1.1%	6.4%
Sometimes circled	14	4 (0)	20 (2)[a]	24	<1	<1	1.1	1.1
	15	9 (0)	104 (2)	113	<1	<1	2.5	5.8
	16	11 (0)	83 (2)	94	<1	<1	3.0	4.6
	17	9 (4)	26 (20)	35	<1	<1	2.5	1.5
		363	1790	2153	4.9%	16.4%	9.1%	13.0%
					12%			

[a]Number in parenthesis is the number of occurrences that were circled.

TABLE 5. Per sample frequency of consensus rule use by two listeners while transcribing 30 continuous speech samples. V = vowels/diphthongs, C = consonants.

Rule category	*Rule number*	*Number of different samples in which rule used at least once*			*Uses per sample*					
					$\bar{x}$		*SD*		*Range*	
		V	*C*	*V + C*	*V*	*C*	*V*	*C*	*V*	*C*
Never circled	1	30	30	30	10.9	47.0	5.5	15.4	4–21	19–87
	2	0	13	13	0	1.5	–	1.4	–	1–6
	3	0	2	2	0	1.5	–	0.7	–	1–2
	4	0	1	1	0	1.0	–	–	–	–
	5	0	6	6	0	1.7	–	0.5	–	1–2
	6	0	0	0	–	–	–	–	–	–
Always circled	7	0	21	21	0	2.4	–	1.2	–	1–5
	8	2	14	15	1	1.7	–	1.1	–	1–4
	9	2	2	4	1	1.0	–	–	–	–
	10	0	2	2	0	1.5	–	0.7	–	1–2
	11	0	4	4	0	1.0	–	0.4	–	1–2
	12	0	8	8	0	1.1	–	–	–	–
	13	0	14	14	0	1.9	–	0.9	–	1–4
Sometimes circled	14	2	15	16	2	1.3	–	0.6	–	1–3
	15	8	28	28	1.1	3.7	0.4	2.5	1–2	1–10
	16	6	25	25	1.8	3.2	0.8	2.0	1–3	1–9
	17	8	14	17	1.1	1.9	0.4	1.3	1–2	1–5

which this immediate resolution of transcription disagreements typically occurred. For convenience, the results of an item-by-item inspection of 1,569 available Rule 1 use comparisons can be summarized as follows.

1. Approximately 63% of both the vowel/diphthong and the consonant disagreements resolved by Rule 1 involved nonerror diacritics (see Table 1, List A). Differences of one element (e.g., unreleased symbol [˺] versus the absence of a released symbol) occurred on approximately 57% of Rule 1 applications, and differences involving two phonetic elements (e.g., fricated [ˌ] vs. dentalized [ˌ]) occurred on an additional 6% of rule applications. Most of the differences that resulted from transcriptions of error productions (use of List B diacritics) involved differences in segment transcription for /r/ and /l/ and in "distortions" of /r/, /l/, /s/, /z/, /ʃ/, and /ʒ/.

2. Both transcribers used the full range of diacritics (Figure 1) in Rule 1 applications, and differences did not follow a sequential pattern within subjects or over time. That is, observed differences were not obviously patterned such that strings of disagreements involved the same transcriptional difference between the two transcribers.

3. Summed over all Rule 1 applications, approximately 70% of the consensus resolutions reflected the original transcription of the more experienced transcriber. The more experienced transcriber generally used more diacritics to modify segments; however, transcribers were equally likely to transcribe a segment as deleted.

4. Anecdotal recall of the transcription period indicated that neither transcriber was aware that one or another of the original transcriptions was being retained more often. That is, neither transcriber was aware of any "dominance" bias during the several weeks of transcription. Transcribers reported that the most frequent reason for Rule 1 resolution was that one or another transcriber had been inattentive to a particular phonetic behavior, which was immediately obvious upon replay.

SUMMARY AND CONCLUSIONS

A central issue in the use of narrow phonetic transcription for clinical speech samples is whether such perceptual data are valid and reliable. Validity questions involving relationships between instrumental records and transcription from live and audio/videotaped samples have not been addressed in this note; research in these areas is encouraged. Validity concerns involving the concept of speech "errors" have been addressed here by presenting a table of nonerror/error response definitions as a stimulus for future research. Reliability concerns, as viewed here, have been approached methodologically in two ways in the literature: by either estimating the measurement error of one transcriber or judge (intrajudge and interjudge reliability coefficients) or by attempting to reduce measurement errors by using consensus of two or more transcribers. Elsewhere (Hoffmann, 1982) we submit that using perceptual phonetics alone, point-to-point interjudge percentage of agreement in the mid 70s may represent the ceiling for narrow phonetic transcription of very young children or disordered speech. Finally, to reduce measurement error, the results from a preliminary study of consensus transcription suggest that well-trained transcribers can successfully use explicit consensus rules to resolve disagreements and retain most of the speech data for phonetic-phonologic analyses. Additional studies using perceptual phonetics in association with acoustic

displays are needed to explore the upper boundaries of the validity and reliability of narrow phonetic transcription.

REFERENCES

BRUNGARD, M. (1961). *Effect of consistency of misarticulation of /r/ and /s/ on gains made with and without therapy.* Unpublished doctoral dissertation, Pennsylvania State University.

COLBY, M. (1944). The early development of social attitudes toward exceptional children. *Journal of Genetic Psychology, 44,* 105–110.

DANILOFF, R., WILCOX, K., & STEPHENS, M. (1980). An acoustic-articulatory description of children's defective /s/ productions. *Journal of Communication Disorders, 13,* 347–363.

DIEDRICH, W., & BANGERT, J. (1980). *Articulation learning.* Houston: College Hill Press.

FREEMAN, G., & SONNEGA, J. (1956). Peer evaluation of children in speech correction class. *Journal of Speech and Hearing Disorders, 21,* 179–182.

HOFFMAN, P. (1983). Interallophonic generalization of /r/ training. *Journal of Speech and Hearing Disorders, 48,* 215–221.

HOFFMAN, P., & SCHUCKERS, G. (1978). Audio-recording effects upon judgment reliability of children's /r/ misarticulations. *Perceptual and Motor Skills, 47,* 451–456.

HOFFMANN, K. (1982). *Speech sound acquisition and natural process occurrence in the continuous speech of three-to-six year old children.* Unpublished master's thesis, University of Wisconsin–Madison.

HOFFMANN, K., & SHRIBERG, L. (1982, November). *Percentage of consonants correct (PCC): A normative study.* Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Toronto.

IRWIN, J., & WONG, S. (1983). *Phonological development in children 18 to 72 months.* Carbondale: Southern Illinois University Press.

KENT, R. (1981, April). *Normal speech development: A motor control perspective.* Paper presented at the Spring Meeting of the Wisconsin Speech and Hearing Association, Madison.

KLEFFNER, F. (1952). *A comparison of the reactions of a group of fourth grade children to recorded examples of defective and non-defective articulation.* Unpublished doctoral dissertation, University of Wisconsin, 1952.

KWIATKOWSKI, J., & SHRIBERG, L. (1983, November). *Classification studies of developmental phonologic disorders: Evidence for subgroups.* Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Cincinnati.

MOWRER, D. (1978). Effect of lisping on audience evaluation of male speakers. *Journal of Speech and Hearing Disorders, 43,* 140–148.

OLLER, D., & EILERS, R. (1975). Phonetic expectation and transcription validity. *Phonetica, 31,* 288–304.

PERRIN, E. (1954). The social position of the speech defective child. *Journal of Speech and Hearing Disorders, 19,* 250–252.

SHRIBERG, L. (1972). Articulation judgments: Some perceptual considerations. *Journal of Speech and Hearing Research, 15,* 876–882.

SHRIBERG, L. (1982, April). *PEPPER (Programs to examine phonetic and phonologic evaluation records) users manual.* Madison, WI: Waisman Center on Mental Retardation and Human Development, Research Computing Facility.

SHRIBERG, L., & KENT, R. (1982). *Clinical phonetics.* New York: Wiley & Sons.

SHRIBERG, L., & KWIATKOWSKI, J. (1980). *Natural process analysis (NPA): A procedure for phonological analysis of continuous speech samples.* New York: Wiley & Sons.

SHRIBERG, L., & KWIATKOWSKI, J. (1982). Phonological disorders I: A diagnostic classification system. *Journal of Speech and Hearing Disorders, 47,* 226–241.

SILVERMAN, E. (1976). Listeners' impression of speakers with lateral lisps. *Journal of Speech and Hearing Disorders, 41,* 547–552.

STEPHENS, M., & DANILOFF, R. (1977). A methodological study of factors affecting the judgment of misarticulated /s/. *Journal of Communication Disorders, 10,* 207–220.

VAN RIPER, C., & IRWIN, J. (1958). *Voice and articulation.* Englewood Cliffs, NJ: Prentice-Hall.

WEISMER, G. (1984). Acoustic analyses strategies for the refinement of phonological analysis. In M. Elbert, D. Dinnsen, & G. Weismer (Eds.), *Phonological theory and the misarticulating child* (ASHA Monographs. No. 22, pp. 30–52). Rockville, MD: American Speech-Language-Hearing Association.

Received January 1, 1983
Accepted September 14, 1983

Requests for reprints should be sent to Lawrence D. Shriberg, Ph.D., Department of Communicative Disorders, University of Wisconsin–Madison, 1975 Willow Drive, Madison, WI 53706.

A Procedure for Phonetic Transcription by Consensus

Lawrence D. Shriberg, Joan Kwiatkowski, and Kit Hoffmann
J Speech Hear Res 1984;27;456-465