Assessing transcription agreement: methodological aspects

CATIA CUCCHIARINI

Department of Language and Speech, University of Nijmegen,
The Netherlands

(Received 5 August 1994; accepted 22 March 1995)

Abstract

In recent years it has become common practice among speech researchers to report transcription agreement coefficients, when describing findings based on phonetic transcription. Although such coefficients are intended to give an indication of the degree of transcription accuracy, in reality it is not clear to what extent high agreement coefficients do indeed guarantee great transcription accuracy. In this paper it is argued that the most commonly used agreement coefficient, percentage agreement, has three major disadvantages when applied to phonetic transcription: (a) it is based on the assumption that agreement between transcription symbols is all-or-none, (b) it is strongly influenced by chance agreement, and (c) it gives no account of the criteria used for transcription alignment. After a detailed discussion of these drawbacks, an alternative approach to calculating transcription (dis)agreement is proposed. In this approach experimentally derived feature matrices are used as input to a program that aligns transcription pairs automatically and at the same time calculates a (dis)agreement measure for each transcription pair, the average distance. It is argued that this metric is more adequate to express the degree of (dis)similarity between transcriptions than the more usual percentage agreement. The results of an evaluation experiment corroborate this view.

Keywords: phonetic transcription, transcription assessment, transcription agreement.

Introduction

In a great variety of linguistic disciplines written records of speech data are used as a basis for research or therapy. One of the most utilized techniques to obtain such written representations is segmental transcription. A segmental transcription is the auditory analysis of an utterance into a linear sequence of discrete units of speech represented by phonetic symbols. The latter are defined by parameters that are mainly articulatory, although some auditory categories are also used.

In early times phonetic transcription constituted the only means of recording speech material. Fieldworkers made their recordings instantaneously, while interviewing the informants. Nowadays symbolic notation is no longer the only

means of recording speech material. As a matter of fact, tape-recorders have been available for this purpose since the second half of this century. Moreover, fieldworkers can now resort to more advanced instruments such as video cameras and electropalatographs (Ladefoged, 1992).

In spite of this, transcriptions are still employed in various fields of linguistics, including phonetics, phonology, dialectology, sociolinguistics, second-language teaching, speech pathology and in studies of language development and language disorders. The way in which transcriptions are carried out varies considerably depending on the discipline and, more specifically, on the aim for which the transcription is made. These factors also determine the degree of detail that is required of a segmental transcription. For instance, in phonology it is common practice to use transcriptions that identify the distinctive elements of a language (phonemic transcription). This type of transcription, however, may turn out to be inadequate when it comes to examining subphonemic variation, as is sometimes necessary in sociolinguistics (Knowles, 1978: 81) or in speech pathology (Grunwell, 1984: 40).

Another major point that should be emphasized is that a transcription, whatever the type, is always the result of an analysis or classification of speech material. Far from being the reality itself, transcription is an abstraction from it. In practice this point is often overlooked, with the result that transcriptions are taken to be the actual phonetic 'data' (Abercrombie, 1965; Rischel, 1990; Kerswill and Wright, 1991). These data are then used as the basis for research or, in the case of speech pathology, for treatment (Grunwell, 1984: 43).

Over the years considerable evidence has been gathered for the fact that segmental transcriptions are subject to variation deriving from different sources (Henderson, 1938; Witting, 1962; Laver, 1965; Ting, 1970; Oller and Eilers, 1975; Amorosa, von Benda, Wagner and Keck, 1985; Pye, Wilcox and Siren, 1988; Shriberg and Lof, 1991). These studies reveal that transcriptions of the same utterance may show considerable differences, either when they are made by different transcribers (between-subjects variation) or when they are made by the same transcriber, but at different times or under different conditions (within-subjects variation). For linguistics and other disciplines that employ transcription to obtain data for further research this could be problematic, because the existence of discrepancies throws doubt on the reliability of segmental transcription as a research tool. As a matter of fact, some authors have addressed the issues of reliability and validity with respect to segmental transcription (Oller and Eilers, 1975; Pye *et al.*, 1988; Vieregge and Cucchiarini, 1988, 1989; Van Borsel, 1989; Kerswill and Wright, 1991).

Concern for the value of segmental transcription as a research tool was particularly evident among speech pathologists. In disciplines such as speech pathology and child phonology perceptual analysis still constitutes the main instrument for data collection, the data being either speech intelligibility scores, speech assessment scores, or transcriptions representing certain characteristics of the speech of a subject. These data are used as a basis for speech assessment, therapy, and the evaluation of treatment effectiveness.

Like many other data, perceptual judgements are subject to measurement errors. Moreover, they tend to contain an element of subjectivity. In view of this, many efforts in the study of communicative disorders have been directed to developing adequate methods for checking the objectivity and accuracy of observational data. As a result it has become customary among speech researchers using transcription data, to provide so-called transcription reliability or agreement indices when

reporting their research findings. Indices that exceed 80% are not uncommon in the literature. An extensive review of articles dealing with the reliability of transcriptions is provided by Shriberg and Lof (1991).

Although reliability indices should have the function of substantiating the findings based on the transcription data, there is a growing tendency among researchers to consider these indices with scepticism (Pye *et al.*, 1988). The main problem seems to be that it is not clear to what extent high reliability coefficients do indeed guarantee transcription accuracy.

The present article aims at discussing some of the main problems connected with the use of transcription reliability indices and at finding an alternative solution. The article begins with a note on terminology. Although the terms agreement and reliability are often used interchangeably, it will be shown that they indicate different properties of a set of data. Furthermore, it will turn out that the majority of studies on transcription reliability in reality concern transcription agreement. Subsequently, attention is paid to the agreement indices usually reported in the literature as a measure of transcription accuracy. Some of the properties of these measures will be discussed, together with some of the disadvantages they present when applied to transcription data. In the following section a different approach to transcription evaluation is considered. This new approach was developed within the framework of a study on the factors influencing intra-transcriber agreement (Cucchiaroni, 1993). Adapted versions of the feature-based matrices developed by Vieregge, Rietveld and Jansen (1984) were used as input to a transcription alignment program that makes it possible to compare transcriptions symbol by symbol and to calculate average distances between consonants and vowels separately. The advantages of this procedure over the traditionally used percentage agreement are discussed.

Reliability and agreement

Reliability is an important property of measuring instruments, and as such it occupies a prominent position in measurement theory. The reliability of the scores assigned by different judges to one and the same object is usually questioned when the scores are found to vary. This explains why this notion is often mentioned in connection with transcription variability. The use of the term 'reliability' seems justified when transcription is analysed as being an instrument of research. However, this term has a specific meaning in statistics and test theory, which is not directly applicable to observations about transcription, as will be shown below.

In general terms the reliability of a measuring instrument represents the degree of consistency observed between repeated measurements of the same object made with that instrument. It is an indication of the degree of accuracy of a measuring device. The notion of reliability is related to the idea that each measurement is subject to some degree of error. In science this phenomenon is known under the name of measurement error (see also Nunnally and Bernstein, 1994: 212–215). Owing to the existence of measurement errors the score resulting from any observation or measurement is seen as a combination of error and something different from it, i.e. the so-called true value. It is obvious that the aim of any measuring activity is to arrive at the true value. However, since any measurement contains a certain amount of error, obtaining the true value appears to be beyond one's possibilities. Indeed, the idea of limit is inherent in the concept of true value. Mathematically, the true value is defined as the limit of the average as the number of observations

approaches infinity. According to this definition the true value can be approached by increasing the number of observations and then taking the average value, on the assumption that error is randomly distributed.

Although this concept of true value may be appropriate for the physical sciences, it is clearly not applicable in disciplines studying human behaviour. There are obvious reasons for this. While the height of a wall can be measured hundreds of times without any observable consequences for the wall, administering a test to a subject hundreds of times is bound to have consequences for the subject. Moreover, administering the test will in itself have an effect on the subject, thus implying that (s)he will change from one measurement to the other.

In view of this it has become customary in psychology and other disciplines concerning human behaviour to use a small number of observations or scores per subject. Reliability is then estimated by analysing the scores of different subjects. The average of the scores of all subjects is an approximation of the true value. The degree of accuracy of the scores, or reliability, will depend on the deviations from the mean: the greater these deviations, the greater the amount of error in the scores. Statistically, reliability is defined as the variance of true values divided by the total variance (true variance plus error variance) (Rietveld and Van Hout, 1993). It follows that if there is no true variance, i.e. the objects to be rated are identical, the degree of reliability will be zero.

At this point it may be useful to consider the difference between reliability and agreement. In the literature these terms are used in different ways by different authors and sometimes they are treated as synonyms (Kelly, 1977; Hopkins and Hermann, 1977; Hartmann, 1977; McReynolds and Kearns, 1983; Kearns and Simmons, 1988). Attempts to solve the confusion in terminology do not appear to have been very successful (Suen and Ary, 1989: 101–129). The definitions of agreement and reliability presented here are in line with those adopted in Tinsley and Weiss (1975), Mitchell (1979), and Rietveld and Van Hout (1993). While the precise meaning of reliability may appear to contrast with our intuitions, that of agreement is rather straightforward and very close to its common use. An example will clarify the difference between these two concepts. By way of illustration, let us suppose that three judges are asked to rate the degree of intelligibility of six speech fragments on a five-point scale. We consider the three hypothetical situations represented in Table 1.

In situation 1 both agreement and reliability are high: the judges assign equal ratings to each of the fragments and also order the fragments in the same way with respect to their degree of intelligibility (in this specific case the fragments are in an order of increasing intelligibility from *a* to *f*). In situation 2 agreement is low because each fragment receives different ratings from the three judges. However, the judges still agree on the ordering of the fragments (also in this example intelligibility increases from fragment *a* to fragment *f*). In other words the judges detect the same differences in intelligibility among the six fragments. Therefore, reliability is said to be high. In situation 3 agreement is high because each speech fragment is assigned very similar ratings by the three judges. By contrast, reliability is low because all judges assign similar ratings to all objects. In other words they do not detect clear-cut differences in intelligibility between the speech fragments. As a result, variability is low in this set of data. Recall that reliability was defined as the proportion of the observed variance which is due to variance in the true scores. It follows that if there is small variability in the scores, while the speech fragments are known to differ with respect to intelligibility, reliability will be low.

Table 1. *Different levels of agreement and reliability for hypothetical intelligibility scores (after Tinsley and Weiss, 1975: 359)*

Speech fragment	Situation 1: high agreement high reliability			Situation 2: low agreement high reliability			Situation 3: high agreement low reliability		
	Judge			Judge			Judge		
	A	B	C	A	B	C	A	B	C
<i>a</i>	1	1	1	1	2	3	3	2	2
<i>b</i>	2	2	2	1	2	3	3	2	3
<i>c</i>	2	2	2	2	3	4	2	3	2
<i>d</i>	3	3	3	2	3	4	3	3	2
<i>e</i>	4	4	4	3	4	5	2	2	3
<i>f</i>	5	5	5	3	4	5	3	3	4
$\bar{x}$	2.8	2.8	2.8	2	3	4	2.7	2.5	2.7

To sum up, agreement has to do with the absolute values of a set of ratings. It indicates to what extent the ratings have identical values. If a panel of judges assign similar ratings to a number of objects, the degree of agreement will be high in any case. In contrast, under the same circumstances, the degree of reliability will be high to the extent that the amount of variation in the objects judged is high. Conversely, if the objects are not sufficiently varied, the degree of reliability will be low. Reliability is related to the idea of covariance. It indicates to what extent the judgements of a series of objects vary in the same way or, put otherwise, 'the degree to which the ratings of different judges are proportional when expressed as deviations from their means' (Tinsley and Weiss, 1975: 359).

This brings us to another important difference between agreement and reliability, i.e. the level of measurement of the variables involved. Since the definition of reliability is based on the notion of proportionality, determining reliability presupposes at least an interval level of measurement. Although exceptions can be made for certain types of ordinal variables (Tinsley and Weiss, 1975: 361), for nominal variables reliability cannot be established. At the nominal level of measurement objects are assigned to a number of mutually exclusive classes for which only statements of equality or difference can be made. If no quantitative differences exist between the categories, it is not possible to speak of deviations from the mean or proportionality. This means that reliability cannot be determined. Agreement, on the other hand, does not require any specific level of measurement and can therefore be established for every type of variable (for a thorough treatment of the notions of agreement and reliability and their implications for statistical analyses, the reader is referred to Rietveld and Van Hout, 1993 and Nunnally and Bernstein, 1994).

From the foregoing discussion it can be concluded that the term reliability does not make much sense with respect to transcription. The speech sounds the phonetic symbols stand for constitute a nominal variable. They can either be the same as, or different from, each other, while ordering is possible only in a limited number of cases (Ashby, 1990). Given that observations about transcriptions are not amenable to interpretation in terms of mean, deviation from the mean, and variance, reliability cannot be calculated. Although some authors do use the term reliability index when dealing with nominal variables such as transcription symbols (Shriberg, 1972;

Shriberg, Kwiatkowski and Hoffmann, 1984; Shriberg and Lof, 1991; Otomo and Stoel-Gammon, 1992; Morrison and Shriberg, 1992), it should be noted that the label ‘agreement index’ would be more correct in this case, because it is ‘more consistent with the terminology employed at the ordinal and interval levels of measurement’ (Tinsley and Weiss, 1975: 361).

It could be argued, however, that if speech sounds are analysed in terms of distinctive features (Singh, 1976), an ordering can be established according to some of these underlying properties. As a matter of fact, there are reasons to believe that two sounds (or symbols) may be more similar than two other sounds. If a multidimensional relationship can indeed be established among speech sounds, this information could then be used to quantify phonetic difference. In turn, this would make it possible to assign weights to the different disagreements between phonetic symbols. Since in this case statements of proportionality would be allowed (Tinsley and Weiss, 1975: 371), it seems that reliability can in theory be established for transcription data.

However, there is another important reason why transcription data are not suitable for reliability computations, which will be explained by means of the following example. Suppose we ask the same two subjects to judge the degree of intelligibility of seven speech fragments (Case A) on a five-point scale and to transcribe the word ‘defend’ (Case B). The data obtained are displayed in Figure 1.

In Case A we may want to determine whether the two subjects assigned similar scores to the same speech fragment. In our example this means that we compare the scores vertically. Alternatively, we may want to know whether the two subjects detect the same differences in intelligibility between the speech fragments. In this case it is important to compare the scores also in the horizontal dimension, to establish whether the difference discerned by Subject 1 between fragment 1 and fragment 2 is comparable to the difference detected by Subject 2 between the same two fragments. As explained above, if the differences are comparable, i.e. the scores covary, reliability is said to be high.

Case A: intelligibility ratings

Subject	Speech fragment					
	1	2	3	4	5	6
1	1	4	2	5	1	3
2	2	3	3	4	2	2

Case B: transcriptions

Subject	Phonetic symbol					
	1	2	3	4	5	6
1	[d	i	f	e	n	d]
2	[d	i	p	e	n	d]

Figure 1. *Hypothetical results of an experiment in which two subjects judge the degree of intelligibility of seven speech fragments on a 5-point scale (Case A) and transcribe the word ‘defend’ (Case B).*

Similarly, in Case B we may also want to compare the subjects' responses along the vertical dimension, to determine whether they used the same phonetic symbol for a given speech sound. However, in general we are not interested in comparing the responses along the horizontal dimension. Of course it would be possible to compare each phonetic symbol with the following one and calculate the difference between them. Dissimilarity values of this kind would make little sense, though. The point is that in Case A we know that if the first response is 2 and the difference between the first and the second response is 1, the second response must be 3. In Case B, on the other hand, this does not apply. If the first response is [d] and the difference between the first and the second response is 2, we still do not know what the second response is. It could be any of the sounds whose distance to [d] is 2. So, the essential difference between Case A and Case B seems to be that while the responses in Case A can be meaningfully ordered along one dimension, this is not possible for the responses in Case B. It follows that the notion of reliability should not be used in connection with transcription data.

Having established that only agreement indices can be calculated for transcription data, let us take a closer look at the indices commonly reported in the speech pathology literature and see exactly what they express.

Agreement indices for transcription data

The kind of transcription agreement indices presented in much of the speech pathology and language acquisition literature (Amorosa *et al.*, 1985; Pye *et al.*, 1988; Ferrier, Johnston and Bashir, 1991; Shriberg and Lof, 1991; Otomo and Stoel-Gammon, 1992) are generally calculated by means of an item-by-item comparison procedure. The agreement statistic thus obtained is known under the name of percentage agreement or point-by-point agreement, and is computed by the following formula:

$$\text{percentage agreement} = \frac{\text{number of agreements}}{\text{number of disagreements} + \text{number of agreements}} \times 100\%$$

Given that percentage agreement is the most widely used index of agreement for phonetic transcription, it seems worthwhile to discuss it in some detail. Although a complete review of its advantages and disadvantages would be beyond the scope of this paper, it may be interesting to discuss some of the problems that can arise when this statistic is employed for phonetic transcription. More precisely, the following three points will receive special attention:

- (a) percentage agreement is based on the assumption that agreement between transcription symbols is all-or-none,
- (b) percentage agreement is strongly influenced by chance agreement, and
- (c) percentage agreement gives no account of the criteria used for string alignment.

These topics are dealt with in the following three sections.

Agreement between transcription symbols is not gradual

In calculating percentage agreement for transcription data, two transcriptions are compared symbol by symbol. For each symbol pair it is determined whether there is agreement or not (the symbols are identical or not). This procedure is usually

applied separately to vowels and consonants, and in certain cases even to diacritics. Sometimes further distinctions are made based on the position of the sound in the word. For example, agreement indices may be calculated separately for initial, medial, and final sounds (Shriberg and Lof, 1991).

One of the disadvantages of the use of percentage agreement for transcription data is that in this statistic agreement between transcription symbols is treated in an all-or-none way. In general, this lack of gradualness of agreement is not viewed as a disadvantage in the case of nominal variables. It certainly does not pose any problem for dichotomous variables, where indeed there are only two logical possibilities and disagreements are equally serious. For instance, when percentage agreement is used to indicate how often two researchers agree on the occurrence or non-occurrence of a given behaviour, the variable can only assume the values 0 and 1. In this case it is obvious that disagreements will all have the same gravity. However, lack of gradualness could be problematic when dealing with multivalued nominal variables in which disagreements might have different degrees of severity. Transcription symbols or, more precisely, the speech sounds they represent, are a case in point. For the purpose of illustration, suppose we have the following sequences:

T1 [pent]
T2 [bent]
T3 [xenk]
T4 [mɪŋk]

Percentage agreement values calculated for the transcription pairs T1–T2, T1–T3, and T1–T4 would be as follows:

$$P(\text{T1-T2}) = \frac{3}{1+3} \times 100\% = 75\%$$

$$P(\text{T1-T3}) = \frac{2}{2+2} \times 100\% = 50\%$$

$$P(\text{T1-T4}) = \frac{0}{4+0} \times 100\% = 0\%$$

The above values show that T1 is more similar to T2 than to T3 or T4. This will probably coincide with our intuitions about the differences between the above strings. However, there are cases in which this procedure gives less satisfactory results. For instance, let us consider the following transcriptions:

T1 [pent]
T2 [bent]
T3 [xent]

In this case all three sequences differ by one symbol, but everybody would probably agree that T1 and T2 are somehow more similar than say T1 and T3 or T2 and T3, which is the same as saying that [p] and [b] are more similar than [p] and [x] or [b] and [x]. Although such observations seem rather intuitive, a rationale can be found if we consider the phonetic symbols in more detail. As explained in the revised version of the Principles of the IPA (Roach, 1989), the aim of the phonetic alphabet is to describe how speech sounds are made, by means of a number of phonetic

categories. The symbols of the alphabet then are 'shorthand ways of indicating certain intersections of these categories'. If we now consider which intersections the symbols [p], [b], and [x] stand for, we notice that they share different categories:

- [p] voiceless bilabial plosive
- [b] voiced bilabial plosive
- [x] voiceless velar fricative

Indeed, we can see that [p] and [x] share the category voiceless and that [p] and [b] share both bilabial and plosive. Put otherwise, [p] and [b] differ in one respect, [p] and [x] in two and [b] and [x] in three. This might explain why the first two sounds may appear to be more similar than the other pairs.

In calculating point-to-point agreement such relationships between speech sounds cannot be taken into account. Since this statistic treats agreement in an all-or-none way, it does not allow for the possibility that between two transcriptions differing by one symbol there may be more agreement than between two other transcriptions that also differ by one symbol. This means that there is a discrepancy in specificity between the level at which speech sounds are transcribed and the level at which agreement is determined. This discrepancy obviously increases with the amount of detail recorded in transcription.

In the study of communicative disorders transcriptions are often rather detailed because this appears to be desirable for research purposes (Grunwell, 1984). Narrow transcriptions offer the possibility of indicating 'minute shades of sound' (Roach, 1989: 68) by using more specific symbols (Abercrombie, 1964) or by means of diacritical marks. Although the label 'minute shade of sound' may be somewhat arbitrary (what is minute in one language may be distinctive in another), it is used here because it illustrates this point nicely. So, certain differences between speech sounds, especially subphonemic differences, are considered to be minute. However, in determining agreement all differences between two transcriptions count the same, minute ones and less minute ones. This means that for both transcription pairs shown below, T1-T2 and T1-T3, an agreement coefficient of 33% is obtained:

T1 [otun̥kɛ]
T2 [olun̥kɛ]

T1 [otun̥kɛ]
T3 [obaste]

Some authors have tried to distinguish between different types of disagreements by computing percentage agreement separately for basic symbols and diacritics (Shriberg and Lof, 1991). This makes it possible, for instance, to indicate that two transcribers agree on the basic symbol and disagree only on the diacritic to be added to it. Although this could be a step in the right direction, it has considerable disadvantages.

First of all, minute differences, understood to be subphonemic differences, can be indicated not only by means of diacritics, but also by different basic symbols. For instance, the two /h/ allophones of Dutch are represented by [h] and [ɦ].

Second, analysing basic symbols and diacritics separately is not methodologically sound. Each diacritical mark is inextricably bound to the phonetic symbol to which it is added. As a matter of fact, a transcriber does not choose a diacritic in its own right, but a diacritic in conjunction with a specific consonant or vowel symbol. Were

(s)he to choose another basic symbol, (s)he would probably also select another diacritical mark. For example, supposing that a given vowel has been transcribed as [ɛ̃] by one subject and as [ẽ] by the other. In this case the diacritics have the effect of making the two transcriptions more similar. However, if agreement is assessed separately for vowels and diacritics, we find that the two subjects disagree in both respects. This interpretation of fine-grained transcription clearly differs from that of other authors according to whom transcribers 'hear' diacritics (Shriberg and Lof, 1991).

The role of chance agreement

Another drawback of percentage agreement is that it can be inflated by chance agreement. When addressing this issue, researchers usually mention one of the possible effects of chance agreement, i.e. the fact that percentage agreement is influenced by the percentage of occurrence of a given event. The reason why attention is paid to this aspect of chance agreement is probably that percentage agreement is generally employed in observational research for dichotomous responses. For instance, it is used to indicate how often two researchers agree on the occurrence or non-occurrence of a previously specified behaviour. In this kind of situation percentage agreement can be very high if the percentage of occurrence of the behaviour in question is either very high or very low. In the mid-range, on the other hand, inflation by chance agreement is limited (McReynolds and Kearns, 1983: 152; Suen and Ary, 1989: 107–108).

However, there is another way in which percentage agreement can be affected by chance agreement, which is sometimes overlooked. As a matter of fact, in the case of multivalued variables, the percentage of agreement is dependent on the number of classification categories involved. In general, percentage agreement decreases as the number of categories increases. When two judges are asked to assign a number of objects to certain categories, the chance of the judges agreeing with each other will decrease as the number of categories increases (Rietveld and Van Hout, 1993: 217). An example may clarify this point. Two phoneticians are asked to judge 20 speech sounds twice. The first time they have to say whether the sound in question is an /r/ or an /l/, whereas the second time they can choose among the following four categories: [r], [R], [l], [ɿ]. Let us consider what the chances are that the two phoneticians completely agree in the two situations.

Possible outcomes:

Case 1: r-r r-l l-r l-l

Case 2: r-r r-R r-l r-ɿ R-r R-R R-l R-ɿ
l-r l-R l-l l-ɿ ɿ-r ɿ-R ɿ-l ɿ-ɿ

The probability of obtaining complete agreement by chance is $2/4 = 0.50$ in Case 1 and $4/16 = 0.25$ in Case 2. Let us now suppose that two phoneticians have judged the 20 speech samples in the two cases as shown in Figure 2. If we now calculate percentage agreement (P), we obtain the following values:

$$\text{Case 1: } P = \frac{15}{5 + 15} \times 100\% = 75\%$$

$$\text{Case 2: } P = \frac{15}{5 + 15} \times 100\% = 75\%$$

Case 1

		phonetician 1		
		/r/	/l/	
phonetician 2	/r/	8	2	10
	/l/	3	7	10
		11	9	20

Case 2

		phonetician 1				
		[r]	[R]	[l]	[ɫ]	
phonetician 2	[r]	4	1			5
	[R]		3	1		4
	[l]	2		2		4
	[ɫ]		1		6	7
		6	5	3	6	20

Figure 2. Hypothetical results obtained when two phoneticians identify 20 speech sounds as either /r/ or /l/ (Case 1) and when they can choose among [r], [R], [l], and [ɫ] (Case 2).

In both cases the two phoneticians agree 75% of the times. Quite clearly this index does not take account of the difference in chance agreement between the two cases. To overcome this difficulty, alternative agreement indices have been developed which discount chance agreement (Tinsley and Weiss, 1975). One such index is the kappa coefficient presented in Cohen (1960). The general formula for kappa is:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Where: P_o = observed percentage of agreement
 P_e = percentage of agreement expected by chance alone

P_e is defined as the sum over the agreement diagonal of the joint probabilities of the marginal proportions. In the two cases above P_e would be:

$$\text{Case 1: } P_e = \frac{11}{20} \times \frac{10}{20} + \frac{9}{20} \times \frac{10}{20} = 0.5$$

$$\text{Case 2: } P_e = \frac{5}{20} \times \frac{6}{20} + \frac{4}{20} \times \frac{5}{20} + \frac{4}{20} \times \frac{3}{20} + \frac{7}{20} \times \frac{6}{20} = 0.26$$

If we calculate kappa for the two cases presented above, we obtain:

$$\text{Case 1: } \kappa = \frac{0.75 - 0.5}{1 - 0.5} = \frac{0.25}{0.5} = 0.5$$

$$\text{Case 2: } \kappa = \frac{0.75 - 0.26}{1 - 0.26} = \frac{0.49}{0.74} = 0.66$$

As we can see, correction for chance agreement leads to different agreement indices for the two cases. The kappa index clearly reflects the fact that complete agreement for 15 cases out of 20 is more significant when there are four categories rather than two. Since kappa gives a more realistic indication of the degree of agreement, it seems natural to ask why transcription investigators still use percentage agreement instead of kappa. The reason lies in the problems involved in computing chance agreement (P_e) for phonetic transcription.

One of the conditions for the application of kappa is that the categories be independent. So, when two phoneticians are asked to identify as [r], [R], [l], or [I], a number of speech sounds that are presented at random, each of the four categories stands equal chances of being selected each time. In phonetic transcription, on the other hand, the transcriber hears one speech fragment which has to be segmented into different parts. It is obvious that there are interdependencies between the different segments and therefore between the different categories. This means that in this case P_e could not be computed as was done above, because the categories are not independent.

At this point the reader will probably wonder about the utility of this digression on the kappa coefficient given that it cannot be employed for phonetic transcription. The reason for introducing kappa was to draw attention to the issue of chance agreement. The importance of chance agreement for phonetic transcription becomes apparent when we compare transcriptions of differing degrees of detail, for instance broad and narrow transcription. It is obvious that chance agreement will be higher for broad transcription than for narrow transcription, since the latter involves a greater number of categories.

Although chance agreement cannot be easily computed for phonetic transcription, this by no means implies that its possible impact should be completely disregarded. The latter seems to be the case, though, in certain studies of communicative disorders. For example, Henderson (1938) calculated two types of percentage agreement for the same data. One index indicated whether the transcribers agreed on the correctness or incorrectness of the transcribed speech sounds, whereas the other one showed to what extent the transcribers had recorded the same sounds. The author concluded that the degree of agreement was considerably higher (above the commonly accepted standard of 75%) in the first case than in the second. This is precisely what one would expect on the basis of chance agreement, since there are only two categories (correct or incorrect) in the first case and many more in the second.

Similarly, in comparing agreement percentages for broad and narrow transcription Shriberg and Lof (1991: 254–256) fail to mention that in broad transcription a greater part of agreement is due to chance agreement than in narrow transcription. In their study 41 basic symbols were used in broad transcription, while in narrow transcription 35 diacritics were available over and above the 41 symbols. It follows that there was a considerable difference in specificity (number of categories) between

broad and narrow transcription, which in turn affected the agreement indices. As McReynolds and Kearns (1983: 141) point out, 'higher coefficients of agreement that result from use of simpler observation codes do not guarantee that observer recordings are accurate'.

The substantial changes in percentage agreement that result from decreasing or increasing the number of categories, are a direct consequence of the fact that in this statistic agreement is treated in an all-or-none way. The more details transcribers record, the more different their transcriptions are likely to be. Since all differences between transcriptions count the same (there is either complete agreement or no agreement), it follows that percentage agreement will be very sensitive to the number of categories, i.e. degree of detail recorded in transcription.

On the basis of the foregoing discussion it would seem advisable to depart from the notion of absolute agreement between phonetic symbols, as it is expressed in percentage agreement, and adopt a system of distance values. In this respect it may be important to note that the idea of gradual agreement between speech sounds is central to transcription comparison. Even when percentage agreement is used to indicate the difference between two transcriptions, some kind of gradualness of agreement is assumed because the transcriptions have to be aligned in the first place. Although this aspect of the comparison procedure is hardly mentioned in the literature, it is an important stage that deserves more attention. In the following section string alignment is discussed in more detail.

String alignment

The fact that in percentage or point-to-point agreement transcription symbols are compared pairwise presupposes that the two transcriptions contain the same number of symbols. A legitimate question at this point would be: what happens when the two transcriptions do not contain the same number of symbols? This is certainly not far-fetched, as it is very possible that one of the transcribers omits or inserts one or more symbols. As a result the two strings of symbols are not of equal length. In this case it has first to be determined which symbol(s) has/have been deleted in one string or inserted in the other, depending on the point of view. This sounds very elementary, but turns out to be important for the comparison procedure.

Point-to-point agreement is computed by comparing corresponding elements, that is symbols describing the same articulatory event. It follows that, before two strings of symbols can be compared, they have to be aligned, i.e. each symbol in string A has to be matched with the corresponding symbol in string B. This is easy when there are neither deletions nor insertions, since in this case there is a kind of natural correspondence between the elements in the two strings. However, any departure from this ideal situation is likely to complicate the alignment procedure considerably. For instance, in the presence of deletions or insertions, following the sequential order would result in very low agreement percentages, as will appear from the following example. Let us consider the following two transcriptions of the word 'artist':

T1 [a:rtɪst]

T2 [a:rtɛst]

As we can see, T1 contains six symbols and T2 only five. From the second position onwards the two transcripts differ in all successive elements. In this case the agree-

ment percentage would be very low ($1/(4+2)=0.17$), and one would wrongly conclude that the two transcriptions are considerably different.

To avoid such mistakes one should first find the right correspondence between the symbols in T1 and those in T2, as shown below:

T1 [a: r t ɪ s t]
T2 [a: 0 t ə s t]

As appears from the above alignment, a null symbol has been inserted in the position corresponding to the deleted symbol. Apart from the two mismatches $r-0$ and $ɪ-ə$, the two transcriptions are rather similar and the agreement score obtained with this alignment is higher than the previous one: $4/(2+4)=0.67$.

Complex cases of string comparison, typically between sequences of symbols of different length, are dealt with in sequence comparison (Kruskal, 1983). In recent years sequence comparison has been studied by many researchers working in different areas such as geology, biology, chemistry, speech technology, etc. The methods developed to cope with the difficulties inherent in sequence comparison usually make use of dynamic programming. The algorithms applied make it possible to obtain the best correspondence between the symbols in the two strings, without having to try all possible solutions. In order to establish the best correspondence, the alignment program needs to know which elements are likely to appear in the same position or, put differently, which substitutions are more likely to occur. For example, to obtain the alignment presented above and not, say, the following one:

T1 [a: r t ɪ s t]
T2 [a: t ə s t 0]

the program needs to know that $[a:]-[a:]$, $[t]-[t]$, and $[s]-[s]$, have zero difference while $[t]-[r]$, $[t]-[ə]$, $[ɪ]-[s]$ and $[s]-[t]$ do not. This information can be fed into the computer in the form of a matrix containing costs or distance values for all possible substitutions. Quite clearly, the performance of such a program depends strongly on the values contained in the distance matrix. In fact, finding the appropriate distance measures is one of the fundamental problems in string alignment.

With regard to transcription comparison, the importance of distance relationships between speech sounds is usually overlooked. It becomes evident, however, when it comes to writing a computer program to perform alignment automatically. When phoneticians align transcriptions by hand, they use their knowledge of speech production and perception to arrive at what they think is the best alignment. They proceed almost automatically, guided by their internalised knowledge, without actually thinking why $[p]$ is matched with $[b]$ rather than, say, with $[r]$. Alternatively, when an automatic system is used, this knowledge has to be made explicit and available to the computer in the form of rules, constraints or matrices, with distances between the speech sounds represented by the phonetic symbols.

It should be clear now that the idea that there are varying degrees of similarity between speech sounds is somehow inherent in transcription comparison. Even when gradualness of agreement is not made explicit through the adoption of a distance matrix, and researchers employ percentage agreement to express transcription similarity, they nevertheless do assume that certain symbols are more similar than others when they align transcriptions. Sometimes the underlying idea of distance between speech sounds is revealed by the terms used in reporting the results obtained. For example, although Amorosa *et al.* (1985) calculate point-to-point agreement for

their transcription data, thus assuming all-or-none agreement between symbols, they state that some of the observed changes were 'further than the original transcription from the phoneme an adult would normally produce'.

The fact that deletions are often reported in studies of transcription variation, (Amorosa *et al.*, 1985; Shriberg and Lof, 1991) suggests that, in practice, transcription alignment is less straightforward than one might think. In spite of this, no mention is usually made of the criteria followed in aligning transcriptions before calculating percentage agreement. This is surprising, especially in speech pathology where the importance of transcription alignment cannot be overestimated.

A different approach to transcription evaluation

Within the framework of a study on the factors affecting intra-transcriber agreement (Cucchiaroni, 1993), we were looking for an adequate measure of transcription (dis)similarity to be used as dependent variable. In the preceding sections it was explained that the measure usually adopted in many transcription or transcription-based studies, percentage agreement, is unsatisfactory. In view of this we wondered whether it would be possible to develop an evaluation metric for transcription which does not exhibit the same disadvantages as point-to-point agreement, or at least one in which these disadvantages are less evident.

A possible solution would be to assign different weights to the various types of disagreements according to the degree of similarity or difference between speech sounds. The only problem with this approach is that it requires a quantification of speech sound similarity. If it were indeed possible to express the degree of similarity or difference between speech sounds in numerical form, then the values obtained could be employed to weight disagreements, thus obtaining a more refined overall agreement index for pairs of transcriptions.

The practice of assigning weights to different sorts of disagreements between categories of nominal variables is not completely unknown in observational research. As a matter of fact, a variant of the kappa index, weighted kappa (Cohen, 1968), was specially developed for situations in which not all disagreements are considered to be equally serious. This may be the case when multidimensional relationships are known to obtain among the different categories. Speech sounds are a case in point. As we saw above, speech sounds can be analysed in terms of a limited number of distinctive features, and the magnitude of difference between the various sounds can be expressed in terms of the number of features by which they differ (Van den Broecke, 1976). Along these lines, Vieregge *et al.* (1984) tried to develop a system of distance values for scoring discrepancies between transcriptions. This system formed the basis of the approach to transcription evaluation that is presented in this paper.

Feature-based distance matrices

In an attempt to develop a system for evaluating segmental transcriptions, Vieregge *et al.* (1984) constructed two matrices, one for consonants and one for vowels, in which distances between all possible pairs of sounds are represented by numbers. The two matrices were developed on the basis of experiments in which subjects were asked to judge the degree of articulatory difference between speech sounds.

The rationale behind Vieregge *et al.*'s experiments was that proprioceptively

derived distances are the most appropriate to determine overall distances between segmental transcriptions. When making a transcription a phonetician tries to identify the articulatory movements made by the speaker by experiencing them in his own vocal tract (Catford, 1974: 2502). The fact that transcribers analyse speech sounds in a proprioceptive way led the above-mentioned authors to choose judgements of articulatory (dis)similarity based on proprioception as a basis for transcription evaluation.

The speech material used in these experiments consisted of pairs of words presented visually in a random order. Each word contained one of the sounds under investigation in medial position. The words of each pair were chosen so as to obtain maximal similarity (equal number of syllables and equal stress pattern). The subjects' task was to try to reproduce the sounds contained in the stimulus words and indicate the degree of articulatory dissimilarity on a 10-point scale.

The dissimilarity scores obtained were then analysed by means of multidimensional scaling so as to retrieve the dimensions underlying the subjects' judgements. Through a combination of experimental results and phonetic knowledge, two feature matrices were obtained, one for vowels and one for consonants. Each speech sound was then defined by a unique combination of feature values. Subsequently, city-block distances were calculated for each pair of sounds. These distances are obtained by comparing each pair of sounds feature by feature. For each feature the absolute difference between the feature values is computed. These partial differences are then summed, thus obtaining an overall index of the difference between the two sounds in question. Following this procedure two distance matrices were constructed, which could subsequently be utilized to express the degree of difference between transcriptions (Vieregge, 1987).

The example below gives an idea of how the feature matrices are used. We consider the values assigned to [t], [s], and [l] for the various features:

Cons.	place	voice	nas	stop	glide	lat	fric	trill	high	distr
t	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
s	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
l	2.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

Suppose we want to calculate the difference between [t] and [s] and that between [t] and [l]. We first compute the absolute difference between the values assigned to [t] and [s] for each feature.

t	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
s	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0

By summing these differences we obtain the distance between [t] and [s], which is 2. Using the same procedure we can calculate the difference between [l] and [t] as follows:

t	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
l	2.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0

In this case the total difference is 3.

A transcription alignment program

In our study on the determinants of intra-transcriber agreement (Cucchiari, 1993) a total of 6720 transcriptions had to be analysed. It was clear that in this case alignment by hand was out of the question. Therefore a program was developed to align transcriptions automatically and determine their degree of distance (Vriens, 1990). The algorithm employed in this alignment program very much resembles the one developed by Picone, Goudie-Marshall, Doddington and Fisher (1986). This is an adapted version of the standard dynamic programming algorithm, which aligns two sequences of elements minimizing the cumulative distance between them (Kruskal, 1983). In our case the elements are phonetic symbols drawn from the Alphabet of the International Phonetic Association, which represent vowels and consonants. These basic symbols can be augmented with diacritics. So a string element can consist either of a phonetic symbol plus one or more diacritics or of a phonetic symbol only. In total there are 39 consonant symbols, 20 vowel symbols, and 22 diacritics. String alignment is performed on the basis of distance measures between the various symbols.

The major differences between our program and that of Picone *et al.* concern the aim and the input matrices. While Picone *et al.*'s program only aims at aligning strings, our program also produces distance indices. Further, Picone *et al.* use phoneme distance matrices based on listening experiments (Peterson and Barney, 1952; Miller and Nicely, 1955) whereas our program employs matrices containing feature values (see Appendix). These matrices were obtained on the basis of the experiments carried out by Vieregge *et al.* (1984). There are 10 features to define consonants and three to define vowels. Distances between speech sounds are computed as the program needs them. Although this makes the system slower, it has the important advantage of making it possible to include diacritical marks.

Diacritics have the property of altering the feature values of basic symbols. Owing to the addition of a relevant diacritical mark, two basic symbols may become more similar or more different. For instance, the two vowel symbols [e] and [ɛ] differ with respect to the feature 'height'. By adding appropriate diacritics the two transcriptions can be made more similar: [e̞] [ɛ̞].

Taking into account the effect of diacritics during automatic comparison is difficult when phoneme distance matrices are used as input. Since in this case the feature values characterizing the speech sounds are not available, they cannot be changed. The only solution would be to expand the matrix so as to include entries consisting of symbols plus diacritics. To obviate these problems we opted for feature matrices whose values can be changed, if diacritics are added to the symbols. Only after the effect of diacritics has been discounted can the distance between the transcription symbols be computed. By way of illustration, in the case of [e̞] and [ɛ̞], the distance is computed as follows. First, the program calculates the effect of the diacritic [̞] on the value [e] has for the feature 'height'. Second, the effect of the diacritic [̞] on the 'height' value of [ɛ] is computed. After the 'height' values of [e] and [ɛ] have been changed into those of [e̞] and [ɛ̞], respectively, the total distance between the two vowels is calculated according to the procedure described above.

Another difference between the matrices used by Picone *et al.* and ours is that the former contain perceptually based distances, whereas the latter refer to articulatory features. Apart from a few exceptions, both programs disallow vowel-to-consonant matches. In Picone *et al.* this is achieved by adding an extra matrix in

which the costs for vowel-to-consonant matchings are greater than those for deletions and insertions. A restricted number of matches between vowels and consonants are allowed, by defining their costs to be lower than that associated with deletions and insertions. In our program vowel-to-consonant matches are prevented by rule. However, since matches between certain vowels and certain approximants turned out to be desirable (see below), a list of exceptions with their associated costs was drawn up.

If two transcriptions do not contain the same number of phonetic symbols, null symbols are inserted. On the basis of the distance values the alignment program determines which symbols are missing in one of the two transcriptions (or have been inserted in the other, depending on the point of view). Costs for insertions/deletions are also contained in a separate list. Figure 3 gives an example of the output of the alignment program. At best, an alignment program will perform as well as a human expert. Of course human performance does not mean 100% correctness, as there can be string pairs that are simply difficult to align, even for an experienced phonetician. This may be the case when two transcriptions are very different, both quantitatively (number of symbols contained) and qualitatively (nature of the phonetic symbols). It is evident that even the best combination of rules and distance values cannot guarantee the performance level of a human expert, as the latter has access to much more information, can use his intuitions and can be more flexible. In other words we have to settle for something which can only approach human performance. This means that in any alignment program human corrections will eventually be required.

However, we should not forget that the use of alignment programs also has some important advantages over manual alignment. Apart from the fact that with a program the number of transcriptions to be aligned is no longer an obstacle, an alignment program also has the advantage of being consistent. Given two transcriptions a computer will always align them in the same way. When alignment is performed by human beings this will not necessarily be the case. It is conceivable that even the same person will apply different criteria for aligning transcriptions on different occasions. For instance, he or she may judge the difference between the same two phonetic symbols in different ways at different times. This is all the more probable when the criteria for alignment are not explicitly defined, as is often the case when percentage agreement is used. Alternatively, if the various degrees of (dis)similarity between phonetic symbols are established beforehand, for example by using distance matrices, alignment is likely to be less subjective.

Figure 3. Sample output of the alignment program.

Before our program was employed to align transcriptions in the present investigation, its performance was checked on 1680 transcription pairs. As is usual in this kind of research (Picone *et al.*, 1986: 782), human experts' judgements were used to benchmark the degree of successfulness of the algorithm. In our case the investigator and a research assistant independently checked whether the transcriptions had been aligned correctly. The results of this evaluation are reported in Cucchiarini and Van Bezooijen (1991). Out of 1680 pairs of transcription 87 (5.17%) string pairs turned out to be incorrectly aligned. Two major causes of incorrect alignment were identified:

1. The first one, accounting for 52 cases (3.09%), was the impossibility of matches between vowels and consonants, which in certain cases led to the insertion of too many null symbols. When one and the same position is occupied by a vowel in one string and by a consonant in the other, the program inserts two null symbols as follows:

p 0
0 a

The reason for this is that the two sounds are so different that they cannot be transcriptions of one and the same articulatory event. Sometimes, however, the difference between vowels and consonants may be less clear-cut. For example, certain approximants are phonetically very similar to certain vowels. In this case it did not seem correct to avoid the matching between a consonant and a vowel and insert two null symbols. In effect, a vowel like [i] and a consonant like [j] could very well be transcriptions of one and the same articulatory event. Therefore, a number of vowels were allowed to be matched with corresponding approximants.

2. The second source of alignment errors was the incorrect matching of diphthongs with long vowels, which was responsible for 19 cases (1.13%). This type of error was due to the fact that the program aligns the long vowel with the first part of the diphthong and then matches the second part with a null symbol. To cope with this problem the whole of the diphthong was allowed to be matched with the long vowel.

It should be noted that the development of this alignment program was not our primary goal, but only a by-product of our research on transcription variation. Given our practical aim, which was developing a system that would make our job easier, we could not devote much time and energy to optimizing the program. In view of these limitations we decided not to incorporate the suggested changes in the program, since this could have had undesirable results for other string pairs. Consequently, the transcription pairs with incorrect alignment were retrieved by means of a program that scans all transcription alignments looking for a pre-specified pattern or subsequence. The changes were then made manually, and the distance values were changed accordingly. By eliminating the two main error sources, this two-step procedure guarantees a higher percentage of correct alignment (99% instead of 94.83%).

Average distances

The alignment program described above was used to align the 6720 transcriptions obtained in our investigation on intra-transcriber agreement. Once two transcriptions

had been aligned, the distance values for all consonant and vowel symbol pairs were summed and divided by the number of consonant and vowel pairs, respectively. The average distance is calculated according to the following formula:

$$D = \frac{1}{N} \sum_{i=1}^N d_i$$

where: D = average distance; N = number of vowel (or consonant) pairs; d = distance between two corresponding vowels (or consonants); i = index.

For each transcription pair two average distance measures were thus obtained, one for consonants and one for vowels. These distance measures were then submitted to analysis of variance to determine the effect of the factors involved in the investigation on intra-transcriber agreement. Preliminary results of these analyses are described in Cucchiarini and Van Bezooijen (1992).

The results obtained with the new metric, the average distance, were subsequently compared with those obtained with percentage disagreement (the complement of percentage agreement). As was to be expected, in certain cases the two measures yield very different levels of intra-transcriber agreement for the same transcription pairs (see Cucchiarini, 1993: 117–162). Given that the average distance was proposed as a possible alternative to the index that is often reported in the literature, percentage agreement, it is now opportune to check whether the new metric constitutes an improvement on percentage agreement on the three points mentioned above.

The first objection against percentage agreement was that, in this statistic, agreement between phonetic symbols is not gradual. In this respect the average distance is a considerable improvement. In this metric disagreements between phonetic symbols are weighted by means of feature matrices. This makes it possible to indicate various gradations of agreement between basic symbols and between transcription symbols differing in diacritical marks.

As was explained in the section on the influence of chance agreement, percentage agreement indices can vary considerably depending on the number of categories on which two judges have to agree. In general the level of agreement will be higher for simple forms of judgement such as correct/incorrect than for more complex evaluations involving a greater number of categories such as, for instance, the precise description of the sounds produced by a patient. This means that the meaning of an agreement coefficient such as percentage agreement depends not only on the coefficient, but on the combination of coefficient and number of categories involved in the judgement. Unfortunately, for transcription agreement coefficients there are no tables indicating the significance of a given coefficient in relation to the degrees of freedom, as is the case for correlation coefficients. Unsatisfactory percentages of agreement can be inflated by reducing the number of categories; that is, by making the judgement less specific, until so-called acceptable percentages are obtained. It is obvious, however, that this procedure does not enhance the accuracy of the observations (McReynolds and Kearns, 1983: 141).

The dissimilarity measure proposed here, the average distance, is less sensitive to the number of categories. This means that increasing or decreasing the amount of detail recorded in transcription will cause less dramatic changes in the degree of dissimilarity obtained. In turn this implies that if the dissimilarity indices are very high (i.e. agreement is low), they cannot artificially be inflated by adopting broader judgement categories.

This brings us to another important point that has not been mentioned so far: how can we establish what acceptable degrees of transcription dissimilarity are? Unfortunately there are no standard rules for this. The only reasonable criterion would be to relate the average distance to the maximum distance that could be obtained with the feature matrices used. For example, in Cucchiarini (1993) the average distance for consonants varied between 0.9 and 1.7 for different experimental conditions, while for vowels it varied between 1.0 and 1.9. Given that the maximum distance could be 9 for consonants and 8 for vowels, and could become even larger with the addition of diacritics, it seems that the values obtained indicate a reasonable degree of intra-transcriber agreement.

In this respect it is important to note that, even for percentage agreement, there are no standardized criteria for acceptance. Sometimes it is assumed that 75% is an acceptable degree of agreement (Henderson, 1938), while in other cases 85% is considered to be the minimum level (Pye *et al.*, 1988; Shriberg and Lof, 1991). However, we have seen that such percentages can have different meanings depending on the number of categories on which the judgement is based. In other words, simply reporting the percentage of agreement, without mentioning how it was arrived at, does not provide much information about the degree of transcription accuracy (see also Pye *et al.*, 1988).

A third remark about the measure of transcription agreement proposed here concerns string alignment. In the present approach the relations between the different symbols are defined by means of feature matrices, which in turn are used as input to a transcription alignment program. This means that the criteria for alignment are clearly established and are applied consistently to every transcription pair.

Further evidence in favour of the average distance is provided by the results of a small-scale evaluation experiment described in Cucchiarini (1993: 154–158)[†]. In this experiment 19 experienced phoneticians were asked to evaluate the degree of transcription dissimilarity for a balanced set of 50 transcription pairs. The subjects indicated their scores on a scale ranging from 1 (complete difference) to 10 (complete identity). The scores exhibited a reliability coefficient of 0.97 (reliability formula for composite ratings with raters as a random factor). For the same 50 transcription pairs the degree of dissimilarity was also determined by calculating the average distance and the degree of percentage disagreement (the complement of percentage agreement). Comparison of three measures for the same transcription pairs revealed that the phoneticians' judgements were more correlated with the average distance ($r = -0.86$, d.f. = 48, $p < 0.01$) than with percentage disagreement ($r = -0.68$, d.f. = 48, $p < 0.01$) (the coefficients are negative because the phoneticians' judgements are inversely proportional to the other two measures). Computation of the associated t statistic showed that the two correlation coefficients are significantly different ($t_{47} = -2.94$, $p < 0.01$). In other words, it seems that when phoneticians evaluate the degree of transcription dissimilarity they do not assign any transcription deviation the same value, as is done in percentage agreement. On the contrary, in scoring transcription deviations they take the degree of similarity between speech sounds into account, as is done in computing the average distance.

[†]For the realization of this experiment we are indebted to Monique Biemans, who carried out most of the practical work.

Conclusions

In this paper a new approach to transcription evaluation is presented, which makes use of feature matrices to calculate a dissimilarity or disagreement measure between pairs of transcriptions, the average distance. This metric is proposed as a possible alternative to the more commonly used percentage agreement. The reason for seeking an alternative metric for transcription evaluation is that percentage agreement has the following three disadvantages: (a) it is based on the assumption that agreement between transcription symbols is all-or-none, (b) it is strongly influenced by chance agreement, and (c) it gives no account of the criteria used for string alignment.

It is claimed that these three negative features are less evident, or completely absent, in the average distance. As a matter of fact, in this metric agreement between transcription symbols is gradual. This also has the advantage of making the metric less sensitive to the number of categories involved in the judgement (the degree of detail). Finally, in this metric the criteria for alignment are clearly defined in the form of feature matrices which are used as input to a transcription alignment program. On these grounds it seems reasonable to conclude that the average distance provides a more adequate measure of transcription (dis)similarity than the index used so far, percentage agreement.

Acknowledgements

The present research was supported by the Linguistic Research Foundation, which is funded by the Netherlands Organization for Scientific Research, NWO. I would like to thank Prof.-dr W. H. Vieregge and Dr A. C. M. Rietveld, who supervised the research reported on here, and commented on an earlier version of this paper.

References

- ABERCROMBIE, D. (1964) *English Phonetic Texts* (London: Faber & Faber).
- ABERCROMBIE, D. (1965) *Studies in Phonetics and Linguistics* (London: Oxford University Press).
- AMOROSA, H., VON BENDA, U., WAGNER, E. and KECK, A. (1985) Transcribing phonetic detail in the speech of unintelligible children: a comparison of procedures. *British Journal of Disorders of Communication*, 20, 281–287.
- ASHBY, M. (1990) Prototype categories in phonetics. *Speech Hearing and Language*, 4, 21–27.
- CATFORD, J. C. (1974) Phonetic fieldwork. In T. A. Sebeok (Ed.), *Current Trends in Linguistics* (The Hague: Mouton).
- COHEN, J. A. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- COHEN, J. A. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.
- CUCCHIARINI, C. (1993) *Phonetic transcription: a methodological and empirical study* (PhD thesis, University of Nijmegen).
- CUCCHIARINI, C. and VAN BEZOOIJEN, R. (1991) The effect of linguistic expectancy on phonetic transcription: developing an adequate alignment algorithm. *Proceedings, XIIth International Congress of Phonetic Sciences*, Aix-en-Provence.
- CUCCHIARINI, C. and VAN BEZOOIJEN, R. (1992) Familiarity with the language transcribed and context as determinants of intratranscriber agreement. *Proceedings, International Conference on Spoken Language Processing 1992*, Banff.
- FERRIER, L. J., JOHNSTON, J. J. and BASHIR, A. S. (1991) A longitudinal study of the babbling and phonological development of a child with hypoglossia. *Clinical Linguistics and Phonetics*, 3, 187–206.

- GRUNWELL, P. (1984) *Clinical Phonology* (London: Croom Helm).
- HARTMANN, D. P. (1977) Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, **10**, 103–116.
- HENDERSON, F. M. (1938) Accuracy in testing the articulation of speech sounds. *Journal of Educational Research*, **31**, 348–356.
- HOPKINS, B. L. and HERMANN, J. A. (1977) Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis*, **10**, 121–126.
- KEARNS, K. P. and SIMMONS, N. N. (1988) Interobserver reliability and perceptual ratings: more than meets the ear. *Journal of Speech and Hearing Research*, **31**, 131–136.
- KELLY, M. B. (1977) A review of the observational data-collection and reliability procedures reported in the *Journal of Applied Behavior Analysis*. *Journal of Applied Behavior Analysis*, **10**, 97–101.
- KERSWILL, P. and WRIGHT, S. (1991) The validity of phonetic transcription: limitations of a sociolinguistic research tool. *Language Variation and Change*, **2**, 255–275.
- KNOWLES, G. O. (1978) The nature of phonological variables in Scouse. In P. Trudgill (Ed.), *Sociolinguistic Patterns in British English* (London: Edward Arnold).
- KRUSKAL, J. B. (1983) An overview of sequence comparison. In J. B. Kruskal and D. Sankoff (Eds), *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison* (Reading, MA: Addison-Wesley).
- LADEFOGED, P. (1992) Knowing enough to analyze spoken languages. *Proceedings, International Conference on Spoken Language Processing 1992*, Banff.
- LAVER, J. D. M. (1965) Variability in vowel perception. *Language and Speech*, **8**, 95–121.
- MCREYNOLDS, L. and KEARNS, K. P. (1983) *Single-subject Experimental Designs in Communicative Disorders* (Austin, TX: Pro-Ed).
- MILLER, G. A. and NICELY, P. E. (1955) An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338–352.
- MITCHELL, S. K. (1979) Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, **86**, 376–390.
- MORRISON, J. A. and SHRIBERG, L. D. (1992) Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research*, **35**, 259–273.
- NUNNALLY, J. C. and BERNSTEIN, I. H. (1994) *Psychometric Theory* (New York: McGraw-Hill).
- OLLER, D. K. and EILERS, R. E. (1975) Phonetic expectation and transcription validity. *Phonetica*, **31**, 288–304.
- OTOMO, K. and STOEL-GAMMON, C. (1992) The acquisition of unrounded vowels in English. *Journal of Speech and Hearing Research*, **35**, 604–616.
- PETERSON, G. E. and BARNEY, L. (1952) Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175–184.
- PICONE, J., GOUDIE-MARSHALL, K. M., DODDINGTON, G. R. and FISHER, W. (1986) Automatic text alignment for speech system evaluation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **34**, 780–784.
- PYE, C., WILCOX, K. A. and SIREN, K. A. (1988) Refining transcriptions: the significance of transcriber 'errors'. *Journal of Child Language*, **15**, 17–37.
- RIETVELD, A. C. M. and VAN HOUT, R. (1993) *Statistical Techniques for the Study of Language and Language Behaviour* (Berlin: Mouton de Gruyter).
- RISCHEL, J. (1990) What is phonetic representation? *Journal of Phonetics*, **18**, 395–410.
- ROACH, P. (1989) Report on the 1989 Kiel Convention. *Journal of the International Phonetic Association*, **19**, 67–80.
- SHRIBERG, L. D. (1972) Articulation judgments: some perceptual considerations. *Journal of Speech and Hearing Research*, **15**, 876–882.
- SHRIBERG, L. D. and LOF, L. (1991) Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, **5**, 225–279.
- SHRIBERG, L. D., KWIATKOWSKI, J. and HOFFMAN, K. (1984) A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research*, **27**, 456–465.
- SINGH, S. (1976) *Distinctive Features. Theory and Validation* (Baltimore, MD: University Park Press).
- SUEN, H. K. and ARY, D. (1989) *Analyzing Quantitative Behavioral Observation Data* (Hillsdale, NJ: Lawrence Erlbaum).
- TING, A. (1970) Phonetic transcription: a study of transcriber variation. *Report from the*

- Project on Language Concepts and Cognitive Skills Related to the Acquisition of Literacy* (Madison: Wisconsin University).
- TINSLEY, H. E. A. and WEISS, D. J. (1975) Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, **22**, 358–376.
- VAN BORSEL, J. (1989) The reliability of phonetic transcriptions: a practical note. *Child Language Teaching and Therapy*, **5**, 327–333.
- VAN DEN BROECKE, M. P. R. (1976) *Hierarchies and Rank Orders in Distinctive Features* (Assen: Van Gorcum).
- VIEREGGE, W. H. (1987) Basic aspects of phonetic segmental transcription. In A. Almeida and A. Braun (Eds), *Probleme der phonetischen Transkription* (Wiesbaden: Franz Steiner Verlag).
- VIEREGGE, W. H. and CUCCHIARINI, C. (1988) Evaluating the transcription process. In W. A. Ainsworth and J. N. Holmes (Eds), *Proceedings, Speech '88*, Edinburgh.
- VIEREGGE, W. H. and CUCCHIARINI, C. (1989) Agreement procedures in phonetic segmental transcriptions. In M. E. H. Schouten and P.Th. van Reenen (Eds), *New Methods in Dialectology* (Dordrecht: Foris).
- VIEREGGE, W. H., RIETVELD, A. C. M. and JANSEN, C. I. E. (1984) A distinctive feature based system for the evaluation of segmental transcription in Dutch. *Proceedings, Xth International Congress of Phonetic Sciences*, Utrecht.
- VRIENS, D. (1990) *Het ophijsingsprogramma: handleiding en documentatie* (unpublished manuscript, Department of Language and Speech, University of Nijmegen).
- WITTING, C. (1962) On the auditory phonetics of connected speech: errors and attitudes in listening. *Word*, **18**, 221–248.

Appendix

The feature matrices used as input to the transcription alignment program are shown below. On the basis of these matrices the program calculates the distances between consonants and vowels, respectively.

Consonant feature matrix

Cons.	place	voice	nas	stop	glide	lat	fric	trill	high	distr
p	1·0	0·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	1·0
b	1·0	1·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	1·0
t	2·0	0·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0
d	2·0	1·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0
c	3·0	0·0	0·0	1·0	0·0	0·0	0·0	0·0	1·0	1·0
k	4·0	0·0	0·0	1·0	0·0	0·0	0·0	0·0	1·0	0·0
g	4·0	1·0	0·0	1·0	0·0	0·0	0·0	0·0	1·0	0·0
f	1·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0	0·0	0·0
v	1·0	1·0	0·0	0·0	0·0	0·0	1·0	0·0	0·0	0·0
s	2·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0	0·0	0·0
z	2·0	1·0	0·0	0·0	0·0	0·0	1·0	0·0	0·0	0·0
ʃ	2·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0	1·0	1·0
ʒ	2·0	1·0	0·0	0·0	0·0	0·0	1·0	0·0	1·0	1·0
x	4·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0	1·0	0·0
y	4·0	1·0	0·0	0·0	0·0	0·0	1·0	0·0	1·0	0·0
χ	4·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0	0·0	0·0
κ	4·0	1·0	0·0	0·0	0·0	0·0	1·0	0·0	0·0	0·0
m	1·0	1·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0
ŋ	1·0	1·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0

Cons.	place	voice	nas	stop	glide	lat	fric	trill	high	distr
n	2.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ŋ	4.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
ɲ	3.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
l	2.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
ɭ	4.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
r	2.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
ɾ	2.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R	4.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
w	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
ʋ	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
j	3.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
h	5.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
ʔ	5.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ç	3.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
ɹ	2.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
ɻ	3.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
ʃ	2.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
ʂ	5.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
ʈ	1.5	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
θ	1.5	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0

Vowel feature matrix

Vowel	front/back	tongue height	lip rounding
a	1.5	1.0	0.0
æ	1.0	1.5	0.0
ɛ	1.0	2.0	0.0
œ	1.0	2.0	1.0
e	1.0	3.0	0.0
ø	1.0	3.0	1.0
i	1.0	4.0	0.0
y	1.5	4.0	1.0
ɪ	1.5	3.0	0.0
ʏ	1.5	3.0	1.0
ə	2.0	2.5	0.0
ɜ	2.0	2.0	0.0
ɐ	2.0	1.5	0.0
ʊ	2.5	3.0	1.0
u	3.0	4.0	1.0
o	3.0	3.0	1.0
ɔ	3.0	2.5	1.0
ʌ	3.0	2.0	0.0
ɒ	3.0	1.0	1.0
ɑ	3.0	1.0	0.0