Comparison of speech sounds: distance vs. cost metrics

John J. Ohala

1. Introduction

Making comparisons between speech samples is a very fundamental procedure in the phonetic sciences. In automatic speech recognition (ASR), a speech sample whose linguistic identity is unknown is compared with a list of previously identified utterances. The unknown is assigned the value of the stored sample it most closely resembles. In phonology two sounds involved in sound changes, e.g., /kw/ and /p/ (Latin *aqua* 'water'; Rumanian *apa*) are compared in order to judge the likelihood of the change having occurred due to the acoustic similarity and thus confusability of the two sounds (Ohala 1993). Many more examples could be given from other phonetic domains including the clinical and the pedagogical. In general, it seems that one of the main reasons we are interested in making comparisons is that we need to categorize speech sounds and that we recognize that two or more entities presented to us can exhibit some variation but may still merit a similar categorization. But what does it mean to 'compare' things?

2. What is the nature of comparison?

Much could be said about the process of comparison: that the procedures and criteria applied may differ depending on (a) the taxonomic level at which the comparison is being made, e.g., (to step outside the domain of speech) is this object a fruit, is it a fruit of the *Rosaceae* family, is it an apple (species *Malus*), is it a Mackintosh apple?), (b) the purpose of the comparison (e.g., one can imagine that in a plagiarism lawsuit, the plaintiffs will give more weight to similarities between the texts whereas the defendants will emphasize differences that may be present). Most comparisons are made in a relative way, i.e., with reference to a well defined universe of investigation, e.g., speech sounds are compared with other speech sounds, not with people or vegetables. This implies that some kind of preliminary comparison precedes any detailed comparison.

In all cases, however, it seems that whether done intuitively or by algorithm, the items to be compared are analyzed into some sort of *primitive*

features and then the similarity of the compared items is estimated in terms of these features. So the question of how comparison is done boils down to how these feature differences are assessed.

In phonetics and many other fields, it is common to simply deal with the *magnitude* of the differences. The differences can be expressed in a discrete all-or-none way ([+ nasal] vs [- nasal]) or in a continuous, parametric, way (a certain amplitude of energy in a given frequency band). In order to get overall difference of the two sounds compared it is common to use one of two metrics. First, one can compute the sum of the differences in feature magnitudes (normalized if necessary); this is the so-called 'city-block' distance. They may also be characterized as the Pythagorean 'distances' in an n-dimensional Euclidean space under the assumption that all feature are orthogonal; see Hanson (1967:35). Both metrics make use of the notion and the mathematics of spatial measurement. Whichever spatial metric is used, their crucial element is that only the magnitudes of the featural differences are used.

The use of the spatial metaphor to express degree of similarity or difference of objects is also very common in ordinary speech; witness such expressions as 'You're not *far off* (in guessing the truth)', 'a *close* translation', 'it costs *next to* nothing.'

The practice of characterizing speech sounds as points in some kind of space has a long history; it was explicit in Panini's works of about two and a half millennia ago. Originally, of course, phoneticians considered the space they used as a real, physical, one, that inside the vocal tract. It wasn't until fairly recent times that the dimensions of some of the spaces used were no longer inherently spatial, e.g., Trubetzkoy (1939) included manners of articulation as dimensions in his phonetic spaces and Joos (1948) plotted vowels on a formant one vs. formant two space. More recently many researchers have plotted speech sounds in a perceptual 'space' constructed by statistical means, e.g., multi-dimensional scaling of listeners' reactions to speech sounds (Hanson 1967). The use of the spatial metaphor in phonetics undoubtedly is very useful for certain purposes but I wish to argue here that it is not useful in all cases.

3. The problem: asymmetries in confusions of speech sounds

A clear warning signal that spaces fail to reflect fully the mechanisms of speech perception is the fact that numerous perceptual studies have exposed asymmetrical confusions, i.e., where one sound, A, is confused as B more often than B is confused as A. For example, in one of the experimental conditions used by Winitz, Scheib, and Reeds (1972), they obtained the following asymmetries of confusion:

pi > ti	34%	but	ti	> pi	6%
ku > pu	27%	but	pu	> ku	16%.

Also, the confusion matrices obtained by Miller and Nicely (1955) show a stronger tendency for /T/ to be confused as /f/ rather than vice-versa. These asymmetries are also found in sound change—presumably because they originate in perceptual confusions of the same type that occur in the lab (see Ohala 1983a, 1985, 1993). If incidence of confusion may be taken as a measure of the similarity of the two sounds involved, such data signify a relation that is incompatible with a spatial representation: that, e.g., /pi/ is closer to /ti/ than /ti/ is to /pi/.

Researchers have recognized the problem, of course, and those using data from confusion matrices as the input to a multi-dimensional scaling construction of a perceptual space have dealt with such asymmetries in two ways. One approach is to impose symmetry by simply taking the probability of mutual confusion between A and B as the average of the probability of A being confused as B plus that of B being confused as A. In addition, the assumption of 'response bias' can level out some of the asymmetries (see Goldstein 1977). But averaging obviously side-steps the problem of asymmetrical confusions and the assumption of response bias is often made without independent justification (other than the existence of the asymmetries in the confusion matrix).

To put the problem in perspective it may be useful to examine similar asymmetries in other perceptual domains. In both visual and vibro-tactile displays subjects' perception of the 26 letters of the roman alphabet show virtually identical asymmetrical confusions (the particular confusions vary, of course, depending on whether upper or lower case letters are involved). For example, the following asymmetries crop up (where ">" means the symbol on the left is more often confused with the symbol on the right rather than vice-versa): R > P, B > P, P > F, Q > O, J > I, W > V, E > F (Gilmore—Hersh—Caramazza—Griffin 1979; Craig 1979). Response bias is of no help here since many of the letters on the left side of the arrow, e.g., R, E, occur much more frequently in English text than those on the right side that they are confused with—and it is well recognized that genuine response bias correlates with frequency of occurrence. As Garner (1978) discusses, such asymmetries may be accounted for by considering that the two stimulus arrays which are confused are structurally similar to each other except that one, the one on the left side of the arrow, has an 'extra' feature that the other

lacks. Directional confusion will occur if it is more probable that this extra feature is not detected than it is that that particular feature is erroneously "filled in" when it is absent in the stimulus array.

4. Comparison by estimating entropy differences between entities compared

A more general characterization of the circumstances under which such asymmetries occur is the presence of an *entropy gradient* on the scale which represents a given feature. The natural course of events, as is well known, is for entropy—a state of greater randomness or "noise", of less available energy—to increase. In short: order gives way to chaos. Other equivalent expressions associated with this phenomenon is that achieving a state of lower entropy requires more *energy* or greater *cost* (of some resource, e.g., money, time, attention). When two items have different values on such a scale, the item which has a lower entropy value tends very easily to change toward the item with the higher entropy but a change in the reverse direction is less likely. Crucial to a comparison metric that recognizes an entropy gradient along a given featural scale is not only to use the magnitude of the differences between two objects on such a scale, but also to estimate the energy expenditure needed to move along this scale from one position to the other.

Use of the space or distance metaphor or mathematics is appropriate in those special cases where the entropy gradient along a scale is zero, i.e., where it takes as much energy to go from point A to B as it does from B to A. This is the default case in the most commonly thought of space, the two-dimensional surface of earth (ignoring hills and valleys). But there are many domains where the costs of moving between two points on a scale are not symmetrical.

In the case of the perception of the capital letters, an 'E' would be associated with lower entropy (less randomness) than an 'F' in that more energy would be required for its correct perception. The natural tendency then would be for that energy not to be spent—for greater randomness to gain ascendancy—and for the extra feature, the "foot" of the 'E'—to go undetected and thereby taken by the viewer to be an 'F'. We must assume—plausibly, I think—that although the stimulus array for an 'F' would also be subject to the drift towards greater entropy resulting in some details being missed and some others which were not physically present being added, the chances of this noise creating an 'E' is less likely.

Some other non-linguistic examples may be helpful.

In mystery novels the presence of a badly mangled body at the foot of a cliff naturally leads the detective to suspect that the body may have originated at the top of the cliff whereas the same body at the top of the cliff would not be likely to lead him to think that the body was originally at the foot of the cliff. The body has less entropy at the top of the cliff than it does at the foot. On the scale of position with respect to the cliff, the state of a body at the top is "closer" to its state at the foot, than vice-versa.

A rich person presumably has less entropy than a poor person. Rich people are therefore more similar to the poor than vice-versa since it is very easy for the former to divest themselves of their wealth whereas considerable expenditure of effort would be required for the poor to acquire wealth.[1]

A living person has less entropy than a dead person. If we imagine that 'living' and 'dead' occupy opposite ends of a scale, people quite readily characterize a gravely ill but still living person as "close to death" but no one would say that a dead person, even if only recently deceased, is "close to living". (I exclude `here a possible temporal meaning for 'close.')

Computer scientists are interested in finding ways to correct computer users' typing and spelling mistakes and accordingly have worked on string-to-string similarity metrics (Morgan 1970; Wagner—Fischer 1975). They recognize first that there are certain 'primitive' typing errors that can distort a string from its intended form: addition, deletion, substitution, and transposition or metathesis (others might be added). Second, they compute the minimum 'cost' to transform a given unrecognized letter string into a recognizable one assuming that these primitive transformations have occurred, each of which has a certain cost or weight attached to it. For example, if additions "cost" more than deletions (which seems intuitively reasonable), then 'top' would be more similar to 'stop' (and thus a better candidate as a mistyped version of it) than the latter would be to the former.

Greenberg and Jenkins (1964), Vitz and Winkler (1973), and Derwing and Nearey (1986) have explored the use of similar (though simpler) string-to-string comparison algorithms for the sake of estimating the phonetic difference between two words or phoneme strings. Although none of them proposed differential weighting or costs for the different transformations, these algorithms could (and probably should) be developed in that way.

The example of the spelling-correction algorithms is interesting for another reason. From a purely logical point of view it would have been possible to reduce the number of primitive typing errors needed to two: addition and deletion, or, for that matter, to just one: substitution, if the null symbol could be used, too. This was not done for the reason that the choice of primitive transformations is not something to be decided on purely logical grounds, but rather on empirical grounds based on how typists actually operate. For a typist, an error of metathesis, e.g., "ot" for "to" does not consist of

the deletion of the "o" from second position and the introduction of an "o" in first position, it is a unitary error. Similarly, whatever comparison metrics we eventually use in phonetics should also be based on empirical considerations, not purely logical ones or those determined by computational efficiency.

At present we do not know enough about the process of speech perception to be able to work out deductively the feature scales and their entropy gradients which are appropriate for characterizing speech perception and the asymmetrical confusions found in it. In Ohala (1985) I offered some preliminary speculations as to factors involved in the asymmetrical confusion of /gi/ with /di/ (but seldom vice-versa). The presence of well-defined, highly structured details in spectra, e.g., narrow bandwidth peaks in the burst spectrum, may be associated with low entropy—just as the 'E' is in comparison to 'F'—and thus subject to misdetection, yielding a higher-entropy percept. Velar obstruents, by virtue of having a longer resonating cavity downstream of the point where their noise is created, have sharper, better-defined formant peaks than do apical or labial obstruents. This may account in part for the widely noted asymmetries of confusion involving velar obstruents, including those found in sound change (see opening paragraph).

5. Entropy gradients may differ in production and perception domains

We should also keep in mind that the scales we use may be different or may differ in the direction of their entropy gradient depending on whether we are trying to account for speech sound similarity due to acoustic-auditory factors or articulatory or other factors. Aerodynamic factors would increase the likelihood of a slightly affricated release on stops before high, close vowels or glides (Ohala 1983b). In the articulatory domain, then, affricated releases in certain cases would have higher entropy than non-affricated release. One might imagine, however, that in the acoustic-auditory domain the reverse would be true: affricated releases, like other transient noises, might easily be missed by listeners and thus more subject to change than non-affricated releases. In ASR this would lead to the necessity of forming an estimate of what types of distortions and variation the speech samples one gathers are subject to, whether just those due to the speech production apparatus or in addition, those added 'downstream' by the ambient acoustic environment or the reception process itself.

6. Prior work

I am aware of at least two prior cases where the possibility of asymmetries in the direction of change of speech sounds has been recognized.

There is, first, the technique of 'dynamic time warping' (DTW) in ASR for the comparison of speech samples represented as a temporally ordered series of spectra. In DTW a search finds the best spectral sample of the unknown input word to compare with a given spectral sample of the stored known word. It is possible to constrain the search (called the 'local path constraint' or the 'slope constraint') in such a way as to recognize that it is more likely that some spectral samples in the known would be missing in the input than to be repeated or to have spurious samples introduced (Moore 1985). This is a step in the right direction but it is incapable of dealing with the vast majority of sources of asymmetries in confusion (e.g., the /gi/ > /di/ confusion discussed above).

In phonology the notion of 'markedness' is an acknowledgement that certain speech sounds are more frequent, more likely to be the end product of change (e.g., neutralization of contrast), than others (Trubetzkoy 1939; Chomsky—Halle 1968: chapter 9). As implemented in Chomsky and Halle's *Sound Pattern of English*, though, it includes a number of implausible claims, e.g., that there is zero cost attached to changes to the 'unmarked' state (high entropy) and that all marked (low entropy) features have the same cost. In addition most phonologists seem content with inductively-arrived at statements of asymmetries of change ('marking conventions'); few have attempted to go beyond this in order to discover their physical and psychophysical causes.

7. Conclusion

Comparison of speech sounds using distance metrics are inappropriate to represent the full range of factors which determine the similarities/differences between speech sounds. These metrics take into account only the magnitude of the differences the sounds have in their values of component features. A more general and more useful comparison metric for speech would be one that takes into account the "energy expenditure" required to transform one value of a feature to another, i.e., one which recognizes a entropy gradient on a given feature's scale. These will allow us to account for asymmetries in the direction of speech sound variation.

Notes

1. The following famous exchange was supposed to have taken place between F. Scott Fitzgerald and Ernest Hemingway: F: "You know, the rich are different from us"; H: "Yes, they have more money." Given the discussion in the text, the following alternative reply to Fitzgerald's comment would be equally appropriate and no less "Hemingwayesque": "Yes, but not so different as we are from them."

References

Chomsky, Noam—Morris Halle
1968 *The sound pattern of English.* New York: Harper & Row.

Craig, James C.
1979 "A confusion matrix for tactually presented letters", *Percep. Psychophysics* 26: 409-411.

De Mori, Renato —Ching-Y. Suen (eds.)
1985 *New systems and architectures for automatic speech recognition and synthesis.* [NATO ASI Series, Series F: Computer and System Sciences, Vol. 16] Berlin: Springer-Verlag.

Derwing, Bruce L.—Terry Nearey
1986 "Experimental phonology at the University of Alberta", in: J. J. Ohala—J. J. Jaeger (eds.), 187-209.

Garner, Wendell R.
1978 "Aspects of a stimulus: features, dimensions, and configurations," In E. Rosch —B. B. Lloyd (eds.), 99-133.

Gilmore, G. C.— H. Hersh—A. Caramazza—J. Griffin
1979 "Multidimensional letter similarity derived from recognition errors,"*Percep. Psychophysics* 25: 425-431.

Goldstein, Louis
1977 "Three studies in speech perception: Features, relative salience and bias," *UCLA Working Papers in Phonetics*, No. 39.

Greenberg, Joseph H.—James J. Jenkins
1964 "Studies in the psychological correlates of the sound system of American English, I and II," *Word* 20: 157-177.

Hanson, Göte
1967 *Dimensions in speech sound perception. An experimental study of vowel perception.* Stockholm.

Hattori, Shirô—K. Inoue (eds.)
1983 *Proceedings of the XIIIth International Congress of Linguists, Tokyo, 29 Aug.-4 Sept. 1982. Tokyo.* [Distributed by Sanseido Shoten.]

MacNeilage, Peter F. (ed.)
1983 *The production of speech.* New York: Springer-Verlag.

Charles Jones (ed.)
1993 *Historical Linguistics: Problems and Perspectives.* London: Longman.

Joos, Martin
1948 *Acoustic phonetics.* [Language Monograph No. 23] Baltimore: Linguistic Society of America.

Miller, George A.—Patricia E. Nicely
1955 "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* 27: 338-352.

Moore, Roger K.
1985 "Systems for isolated and connected word recognition," in: R. De Mori—C.-Y. Suen (eds.), 73-143.

Morgan, H. L.
1970 "Spelling correction in systems programs," *Comm. Assoc. Comput. Machin.* 13.2:90-94.

Ohala, John J.
1983a "The phonological end justifies any means," in: S. Hattori—K. Inoue (eds.), 232-243.
1983b "The origin of sound patterns in vocal tract constraints," in: P. F. MacNeilage (ed.), 189-216.
1985 "Linguistics and automatic speech processing," in: R. De Mori—C.-Y. Suen (eds.), 447-475.
1993 "The phonetics of sound change," in: C. Jones (ed), 237-278.

Ohala, John J.—Jeri J. Jaeger (eds.)
1986 *Experimental phonology.* Orlando, FL: Academic Press.

Passy, Paul
1890 *Étude sur les changements phonétiques.* Paris: Firmin-Didot.

Rosch, Eleanor—Barbara B. Lloyd (eds.),
1978 *Cognition and categorization.* Hillsdale: Lawrence Erlbaum Associates.

Trubetzkoy, Nikolai Sergeevich
1939 *Grundzüge der Phonologie*. Prag. [Bd. 7 der Travaux du Cercle Linguistique de Prague.]
Vitz, Paul C.— Brenda S. Winkler
1973 "Predicting the judged 'similarity of sound' of English words," *J. Verbal Learn. Verbal Behav.* 12: 373-388.
Wagner, Robert A.— Michael J. Fischer
1975 "The string-to-string correction problem," *J. Assoc. Comput. Machin.* 21.1: 168-173.
Winitz, H.— M. E. Scheib.—J. A. Reeds
1972. "Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech," *J. Acoust. Soc. Am.* 51: 1309-1317.