DISTANCE MEASURES: PHYSICAL AND PERCEPTUAL ASPECTS [1)]

Louis C.W. Pols

1. INTRODUCTION

Distance measures are used to specify the *(amount of) (dis)similarity between (two) events, on the basis of specific properties*. This definition of distance measures makes it possible to emphasis various characteristics which I consider to be important. In the speech domain one can replace the general term 'event' in this definition by '(multidimensional) speech signal'. Uni-dimensional physical characteristics, such as time, amplitude, or fundamental frequency, as well as their perceptual counter parts (duration, loudness, and pitch), are generally represented on a physical or perceptual scale. And although also with scale definitions there are already considerable difficulties, I will not discuss these any further. I will rather concentrate on spectro-temporal (multi-dimensional) aspects of speech signals, such as one individual vowel spectrum, or a sequence of spectra describing a word, or one measure such as fundamental frequency as a function of time ('pitch contour').
A measure of similarity or distortion on the basis of such signal characteristics is an ever returning problem in all the areas of speech research indicated below. In *speech analysis and coding* an error signal is frequently used to optimize a parametric representation, such as linear predictive coefficients, or pitch extraction. In *rule synthesis* one has again to optimize parametric fits, such as pitch and formant contours. The area of *speech recognition and understanding* is most strongly related to distance measures. They are needed, among other things, for segmentation, clustering, vector quantization, and for testing the input parameters against reference data. However, also in *speech perception* there is a strong emphasis on (dis)similarity between signals. Confusion matrices as a result of identification experiments, similarity matrices as a result of triadic comparison or other similarity judgments, or scores on semantic scales, are all indications of perceptual distances. Furthermore it is important to study the relation between perceptual similarity and physical distance measures, or in other words, the *perceptual relevance of distance measures* . Especially in speech perception, one also should be aware of the distinction between physical identity and perceptual and phonetic identity: Two physically different signals might be identified as the same vowel /a/, for instance in various phonetic contexts, or from different speakers, or in different noise backgrounds. Depending upon its purpose, a distance measure should include all relevant information and neglect information irrelevant for that condition.
In the following sections I will first give some general properties of distances and then describe perceptual and physical approaches for studying and optimizing distance measures.

[1)] Written version of a paper presented at the 'Colloquium Signaalanalyse en Spraak', 23-25 Sept. 1987, Amsterdam.

2. GENERAL PROPERTIES OF DISTANCES

The *Euclidean distance* is probably the best known measure to define the distance between any two points x_i and x_j in an n-dimensional space:

$$d_{i,j} = [\sum_{k=1}^{n} (x_{i,k}-x_{j,k})^2]^{1/2} \qquad [1]$$

This formula can be made more general by using the Minkowski parameter p:

$$d_{i,j} = [\sum_{k=1}^{n} |x_{i,k}-x_{j,k}|^p]^{1/p} \qquad [2]$$

For p=2 this formula represents again the Euclidean distance, p=1 represents the "city block distance", and for large values of p the distance is exclusively determined by the largest value of $(x_{i,k}-x_{j,k})$. If $\mathbf{m}_i$ is considered to be the mean value of a distribution of points, then one can use the distance from any other point to $\mathbf{m}_i$ as the probability that that point belongs to distribution i. The more one knows about the statistics of those distributions, the more specific the distance measure can be, see Fig. 1.

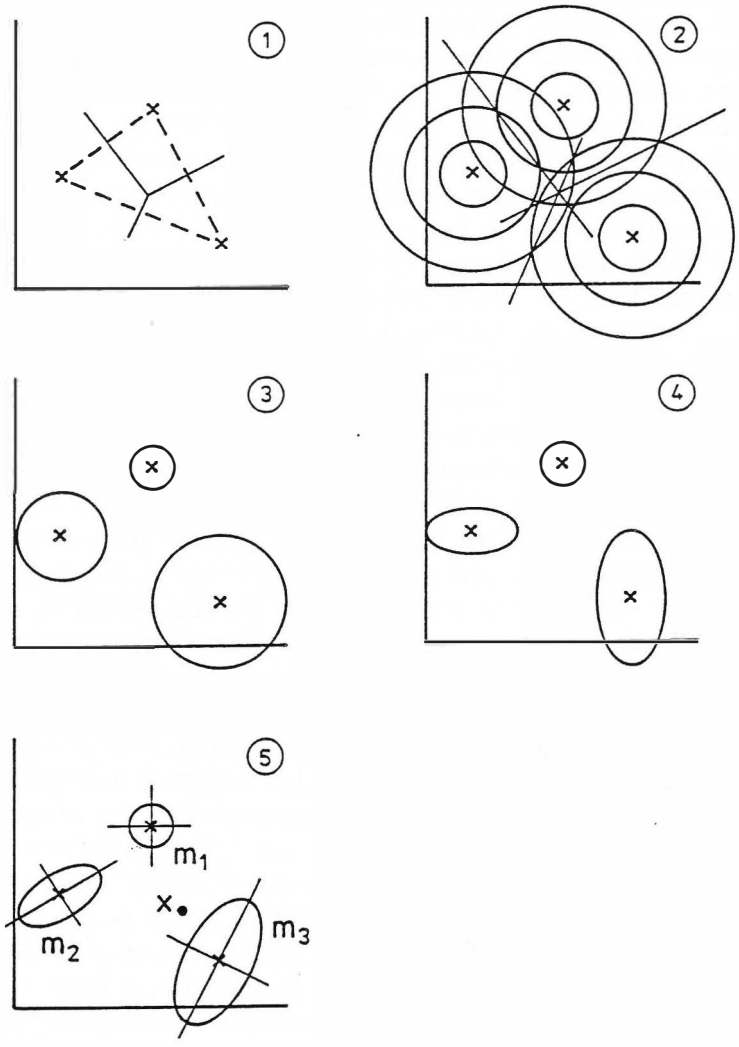


Fig. 1 Various representations of distributions around average positions. In 1) and 2) the distribution around each point is the same. In 3) the variance per dimension is the same, but differs per point, whereas in 4) also the variance per dimension differs. In 5) the axes of maximal variance no longer coincide with the original dimensions, indicating covariance.

The multidimensional probability density function defines the probability that vector **x** belongs to the i-th distribution:

$$f_i(\mathbf{x}) = (2\pi)^{-n/2} \cdot |S_i|^{-1/2} \cdot \exp[-1/2(\mathbf{x}-\mathbf{m}_i)^T \cdot S_i^{-1} \cdot (\mathbf{x}-\mathbf{m}_i)] \quad [3]$$

This function takes into account the (n x n) covariance matrix S_i of average position $\mathbf{m}_i$. The exponent of [3] represents the *Mahalanobis distance* between **x** and **m**:

$$(\mathbf{x}-\mathbf{m})^T \cdot S^{-1} \cdot (\mathbf{x}-\mathbf{m}) \quad [4]$$

For S equal to the identity matrix, formula [4] reduces again to the Euclidean distance (O'Shaughnessy, 1986). The n-dimensional vector **x** could well represent the n filter levels of a vowel analyzed with a bank of n filters, or n reflection coefficients as a result of an n-th order LPC-analysis, or even n articulatory parameters. So far we have supposed that the n values resulting from a certain measurement are used straightforwardly, however, more often than not these data are processed in one or more of the following ways:

- data transformations, such as from linear to dB level, or from linear to Bark frequency scale, or from prediction to cepstral coefficients, or from harmonic power spectrum to auditory model representation (Bladon and Lindblom, 1981; Seneff, 1986);
- data truncations, such as reduced cepstral coefficients;
- data reduction, such as principal-components or discriminant analysis;
- data weighting, such as weighted likelihood ratios, or band-pass liftered weighted cepstral coefficients.

A next level of complexity arises when one has a temporal sequence of measurements, such as a formant or a pitch contour, or the trace of a word in n-space (Pols, 1971). There are no good *automatic* procedures ('t Hart, 1979) to stylize pitch contours, apart from median smoothing (Rabiner et al., 1975) and endpoint connection (O'Shaughnessy and Allen, 1983), which makes it also difficult to define the distance between two contours. In several of the projects of the Dutch SPIN program 'Analysis and synthesis of speech', much attention will be paid to formant tracking, stylization, and parameterization. Green et al. (1987) developed a form of stylization they called 'Speech Sketch', which enables acoustic evidence and phonetic knowledge to be represented in similar ways, so that like can be compared with like. Computationally very different approaches, such as Hidden Markov Modelling (HMM), also require different distance metrics. For a recent review, see Juang and Rabiner (1985).

3. PERCEPTUALLY-ORIENTED DISTANCE MEASURES

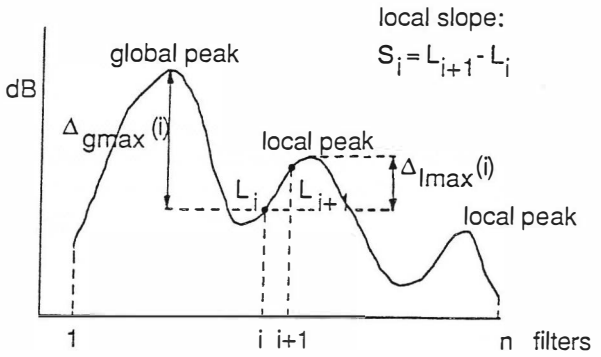
Klatt (1982) did a very interesting experiment to study systematically the effect of various signal modifications upon the perceived distance towards a reference signal: a 300-ms, stationary, harmonic, five-formant, signal sounding like /a/. Altogether 66 different signals, acoustically similar to /a/, were generated by adding together sinusoidal harmonics of the appropriate frequencies, amplitudes, and phases. Small changes to either formant frequencies, formant bandwidths, spectral tilt, and filtering passband or stopband were introduced. In a paired comparison task, with the reference stimulus always as one of the pair, subjects had to judge the distance on a 10-point scale. The data were normalized such that the total responses from each subject had equal mean and variance. The subjects were asked to judge the *phonetic distance*

Table I. Average psychophysical and phonetic distances to a reference signal /a/, for the largest changes in each of the 16 conditions. From Klatt (1982).

Condition	psychophysical	phonetic
1. high-pass filtered at 400 Hz	8.3	0.3
2. spectral tilt: +4, -4 dB/oct	8.1	1.3
3. random phase	7.9	1.1
4. low-pass filtered at 2500 Hz	7.6	3.0
5. F1 & F2: +10, -10%	6.6	8.3
6. F2: +10, -10%	5.6	7.2
7. spectral tilt: +4, -4 dB/kHz	5.2	0.7
8. B1 & B2 & B3 & B4 & B5: +100, -50%	5.0	3.0
9. F1: +10, -10%	4.3	5.4
10. notch filtering: 800-1180 Hz	4.3	3.2
11. F3: +10, -10%	3.3	0.6
12. overall amplitude: +3, -3 dB	3.0	0.8
13. B1: +100, -50%	2.6	2.8
14. notch filtering: 1370-2520 Hz	2.5	0.8
15. B3: +100, -50%	1.8	0.3
16. B2: +100, -50%	1.4	0.7

between stimuli, ignoring as best as they could any changes that were associated with a change in 'speaker or recording conditions'. In an earlier experiment Carlson, Granström, and Klatt (1979) had used similar stimuli to judge the *psychophysical distance*, and in 1976 Carlson and Granström had already studied the discriminability of spectral slope changes in several synthetic vowels. In Table I the major results for the Klatt (1982) experiment are summarized, by giving the average responses to the largest (positive and negative) changes in each condition. As can be seen there are large differences between the two types of judgment. Only those conditions including formant frequency changes (cond. 5, 6, and 9) induce large changes in phonetic distance. Filtering and spectral tilt conditions (1, 2, 4, and 7) produce substantial spectral changes (and large psychophysical distance scores), but are apparently ignored when making phonetic judgments. The large distance for random phase (cond. 3) is somewhat surprising in the light of results found by Plomp and Steeneken (1969). Klatt found no indication for the Chistovich et al. (1979) claim that listeners perform spectral smoothing over about three critical bands prior to matching the stimuli. Klatt repeated the experiments for stimuli resembling the vowel /ae/ and found the same strong effect of formant frequency changes upon phonetic quality. However, he also admits that humans most probably do not detect and label formant peaks when making phonetic judgments, since they generally make only small misperceptions, whereas machines relying on formant extraction make gross errors whenever a formant is missed or a spurious formant added. So, Klatt tries to define a phonetic distance that does not include a formant peak labeling step.What is called by Klatt the "Plomp metric" [area between two critical band spectra; see Plomp (1970), but also Klein et al. (1970) and Pols et al. (1969)] seems to be a good first candidate for that. However, this might be more a psychophysical than a phonetic distance. According to Klatt the ideal phonetic distance "should attend to peak locations, but be insensitive to relative peak heights and to spectral details in the valleys between peaks". He defines a metric (see Fig. 2) based on spectral slope differences near the peaks in the critical-band spectra which correlates very well (0.93) with his perceptual data of Table I. For this optimal fit the constants had the following values: $k_E = 0$ (a neglect of overall level

differences), $k_{LMAX} = 1$ and $k_{GMAX} = 20$ (for units in dB x 10) indicating a large weight to the largest (global) peak. Nocerino et al. (1985) apllied the same measure in his comparative tests (see also Table II) but achieved best results with $k_E = 0$ and $k_{LMAX} = k_{GMAX} = \infty$, which implies an unweighted slope difference. More recently, Klatt (1986) was again studying metrics based on frequency locations of spectral peaks, including alternative spectral representations for disappearing peaks. Bladon and Lindblom (1981) applied auditory modeling in an attempt to improve the correspondence between spectral distance and vowel quality differences by using data of subjects who judged the similarity between pairs of stimuli on a 5-points scale.



WEIGHTED SPECTRAL DISTANCE MEASURE

$$d_{1,2} = k_E \cdot |E_1 - E_2| + \sum_{i=1}^{n-1} k_S(i) \cdot [\, S_1(i) - S_2(i) \,]^2$$

k_E weights the absolute **energy difference** $|E_1 - E_2|$

$k_S(i)$ weights the actual **slope difference** according to local and global peaks

$$k_S(i) = [\, k_{S1}(i) + k_{S2}(i) \,] / 2$$

$$k_{S1}(i) = \left[\frac{k_{LMAX}}{k_{LMAX} + \Delta_{LMAX}(i)} \right] \cdot \left[\frac{k_{GMAX}}{k_{GMAX} + \Delta_{GMAX}(i)} \right]$$

Fig. 2 Representation of the weighted spectral distance measure, as defined by Klatt (1982).

4. PHYSICALLY-ORIENTED DISTANCE MEASURES

Especially in speech recognition research, a great many different distance or distortion measures have been studied. Since linear predictive coding nowadays is so popular, most measures are based on this approach. The differences have, among other things, to do with using autocorrelation coefficients or cepstral coefficients, using linear or Bark-type frequency scales, apply weighting functions in various domains (e.g. bandpass liftering, Juang et al., 1987), stressing peaks, or making the measure symmetrical. Table II gives an overview of a number of them. It is rather difficult to compare performance of these measures over different studies since the conditions are never the same and there might be substantial interaction with:

- speakers (one or many, male or female)
- vocabulary (vowels, digits, alphadigits)

- input bandwidth (wide band vs. telephone bandwidth)
- acoustic environment
- front end (LPC and its order, filterbank and frequency spacing)
- level and gain (full or local normalization)
- training
- dynamic time warping
- reference set (one representative, clusters, vector quantization)
- recognition procedure (probability, k-nearest neighbour)
- accept/reject threshold

It is for this reason that in comparative studies, such as those of Mermelstein (1976), Gray (1976), Rabiner and Soong (1985), Nocerino et al. (1985), and Hanson and Wakita (1987) the above-mentioned conditions are kept as constant as possible.

Table II. List of a number of distance measures frequently used in speech analysis and recognition. The orders indicated are those found in two comparative studies.

	G&M[1]	Noc.[2]	R&S[3]
Euclidean spectral distance (Plomp, 1970)			
rms log spectral distance (Gray and Markel, 1976)	x		
(truncated) cepstral distance (Gray and Markel, 1976)	x	6	4
likelihood ratio (Itakura-Saito distance) (Itakura, 1975)	x	4	3
cosh measure (Gray and Markel, 1976)	x		
maximum likelihood (Itakura, 1975)		8	
log likelihood ratio (Itakura distance) (Itakura, 1975)		2/3	
(truncated) weighted likelihood ratio (Shikano and Sugiyama, 1982)		5	5
weighted slope metric (Klatt, 1982)		2/3	
likelihood ratio + norm. temporal energy (Nocerino et al., 1985)		1	
(Bark-scale frequency-) warped cepstrum (Nocerino et al., 1985)		7	
warped weighted likelihood ratio (Nocerino et al., 1985)		9	
weighted cepstral distance (Tohkura, 1986)			2
BP-liftered weighted cepstral distance (Rabiner and Soong, 1985)			1

[1] Gray and Markel (1976): theoretical and experimental discussion.
[2] Nocerino et al. (1985): LPC-based, alpha-digit vocabulary, 2 male, 2 female, 5-7 x training, 10 x test, telephone bandwidth.
[3] Rabiner and Soong (1985): 10 vowels, 4 male, 3 female, 5 x training, 5 x test, single frame, vector quantization (code-book size 1-8, in this table: 1).

Just as Klatt (1982) gave his opinion about the ideal phonetic distance, he also specified needs for metrics to work best in word recognition systems (Klatt, 1986):
(1) provide a uniform scoring algorithm, such as probability, in order to compare disparate events;
(2) respond in a monotonically increasing way to phonetically increasing distances; and
(3) ignore phonetically irrelevant acoustic details, to the extent that this is possible.
This leads to discussions about consistent feature extraction, invariant features, relative amplitude or slope metrics, and ways to accumulate distance over time. He concludes by saying that "no matter what metric is used, distance scores for phonetically similar spectra are usually too big, while those for very different spectra are not big enough".

5. CONCLUSIONS

It will be clear from the above discussion that *the best distance measure* cannot be defined. Its definition strongly depends on the application field (e.g. coding, recognition, perceptual similarity), on the descriptive parameters (e.g. critical-band spectra, cepstral coefficients, formant contour), on the signals under consideration (e.g. stationary vowels, alphadigits), on the type of similarity (e.g. psychophysic, phonetic), and probably many more. However, present literature has emphasized a number of important aspects, which should allow for a definition most appropriate for its needs. This is certainly much easier for stationary signals than for time sequences, whereas also the perceptual relevance of most measures is still far from optimal.

6. REFERENCES

Applebaum, T.H., Hanson, B.A., and Wakita, H. (1987), "Weighted cepstral distance measures in vector quantization based speech recognizers", Proc. IEEE-ICASSP87, 1155-1158.

Bladon, R.A.W. and Lindblom, B. (1981), "Modeling the judgment of vowel quality differences", J. Acoust. Soc. Amer. 69, 1414-1422.

Carlson, R. and Granström, B. (1976), "Detectability of changes in level and spectral slope of vowels", STL-QPSR 2-3, 1-4.

Chistovich, L.A., Sheikin, R.L., and Lublinskaja, V.V. (1979), "Centres of gravity and spectral peaks as determinants of vowel quality", In: B. Lindblom and S. Öhman (Eds.), Frontiers of speech communication research, Academic Press, New York, 143-158.

Carlson, R., Granström, B., and Klatt, D.H. (1979), "Vowel perception: The relative perceptual salience of selected acoustic manipulations", STL-QPSR 3-4, 73-83.

Green, P.D., Cooke, M.P., Lafferty, H.H., and Simons, A.J.H. (1987), "A speech recognition strategy based on making acoustic evidence and phonetic knowledge explicit", Proc. European Conf. on Speech Techn., Vol. 1, 373-376.

Hanson, B.A. and Wakita, H. (1987), "Spectral slope distance measures with linear prediction analysis for word recognition in noise", IEEE Trans. ASSP-35, 968-973.

Itakura, F. (1975), "Minimum prediction residual principle applied to speech recognition", IEEE Trans. ASSP-23, 67-72.

Juang, B.-H. and Rabiner, L.R. (1985), "A probabilistic distance measure for Hidden Markov Models", AT&T Techn. J. 64, 391-408.

Juang, B.-H., Rabiner, L.R., and Wilpon, J.G. (1987), "On the use of bandpass liftering in speech recognition", IEEE Trans. ASSP-35, 947-954.

Gray, A.H. and Markel, J.D. (1976), "Distance measures for speech processing", IEEE Trans ASSP-24, 380-391.

Gray, R., Buzo, A., Gray, A.H., and Matusyama, Y. (1980), "Distortion measures for speech processing", IEEE Trans. ASSP-28, 367-376.

Klatt, D. H. (1982), "Prediction of perceived phonetic distance from critical-band spectra: A first step", Proc. IEEE-ICASSP82, 1278-1281.

Klatt, D.H. (1986), "The problem of variability in speech recognition and in models of speech perception", In: J.S. Perkell and D.H. Klatt (Eds.), Invariance and variability of speech processes, Lawrence Erlbaum Ass., Hillsdale, N.J., 300-319.

Klein, W., Plomp, R., and Pols, L.C.W. (1970), "Vowel spectra, vowel spaces, and vowel identification", J. Acoust. Soc. Amer. 48, 999-1009.

Nocerino, N., Soong, F.K., Rabiner, L.R., and Klatt, D.H. (1985), "Comparative study of several distortion measures for speech recognition", Speech Comm. 4, 317-331.

Mermelstein, P. (1976), "Distance measures for speech recognition -- Psychological and instrumental", Haskins SR-47, 91-103.

Olano, C. (1983), "An investigation of spectral match statistics using a phonemically marked data base", Proc. IEEE-ICASSP83, 773-776.

O'Shaughnessy, D. (1986), "Speaker recognition", IEEE ASSP Magazine 34, 4-17.

O'Shaughnessy, D. and Allen, J. (1983), "Linguistic modality effects on fundamental frequency in speech", J. Acoust. Soc. Amer. 74, 1155-1171.

Plomp. R. (1970), "Timbre as a multidimensional attribute of complex tones", In: R. Plomp and G.F. Smoorenburg (Eds.), Frequency analysis and periodicity detection in hearing, Sijthoff, Leiden, 397-414.

Plomp, R. and Steeneken, H.J.M. (1969), "Effect of phase on the timbre of complex tones", J. Acoust. Soc. Amer. 46, 409-421.

Pols, L.C.W. (1971), "Real-time recognition of spoken words", IEEE Trans. Comp. 20, 972-978.

Pols, L.C.W., Kamp, L. J. Th. van der, and Plomp, R. (1969), "Perceptual and physical space of vowel sounds", J. Acoust. Soc. Amer. 46, 458-467.

Rabiner, L.R., Sambur, M., and Schmidt, C. (1975), "Application of a non-linear smoothing algorithm to speech processing", IEEE Trans. ASSP-23, 552-557.

Rabiner, L.R. and Soong, F.K. (1985), "Single-frame vowel recognition using vector quantization with several distance measures", AT&T Techn. J. 64, 2319-2330.

't Hart, J. (1979), "Explorations in automatic stylization of F_0 curves", IPO-APR 14, 61-65.

Seneff, S. (1986), "A computational model for the peripheral auditory system: Application to speech recognition research", Proc. IEEE-ICASSP86, 1983-1986.

Tohkura, Y. (1986), "A weighted cepstral distance measure for speech recognition", Proc. IEEE-ICASSP86, 761-764.