DISTORTION MEASURE EVALUATION USING SYNTHETIC SOUNDS AND HUMAN PERCEPTION

D. Tuffelli, H. Ye

Institut de la Communication Parlée (LA. CNRS. No.368)
46, Av. Félix-Viallet 38031 Grenoble
France

In this paper, we try to compare several distortion measures with the human's perception using synthetic sounds. Correlation and another measure of coherence is used. The goal of this research is to study the coherence between mathematical distortion measures and the human's perception. The results show there are some differences between them. But Itakura distortion measure is the best in the case of our isolated vowels.

I. INTRODUCTION

Distortion measures of speech is an important problem for speech processing: speech recognition; speaker identification; speech coding...etc.

Generally, there are 2 kinds of distortion measures. The first one is defined by means of a mathematical criterion, such as Itakura-Saito; cepstral; likelihood ratio and weighted Itakura-Saito[1,2] ... etc. The second is perceptually based measures, such as weighted slope metric(WSM)[3]; euclidean distance of critical-band spectra[5] and weighted likelihood ratio[6] ... etc.

The first approach is purely mathematical without any perceptual constraint. The second approach try to make use of perceptual properties with some model made from human's perception.

An early study has been done with difference limens of formants[7]. A recent study has been done on perceived phonetic distance[3].

Another more global type of comparison[8] was carried out between human performance (presented by confusion matrix) and an automatic recognition algorithm.

The work presented here tries to examine and to compare the previous 2 kinds of distortion measures with the data of a test of psychoacoustics which was especially designed for this goal.

II. EVALUATION OF DISTORTION MEASURES

Different Tested Distortion Measures

*Itakura distortion[10] is gain optimized Itakura-Saito measure which was originally introduced as an error matching function in maximum likelihood estimation of autoregressive spectral models.

$$d_{ita}(x,x') = \log(\alpha/\alpha_m)$$

where α is any residual energy and α_m is minimal residual energy.

*Cepstral distortion measure is an approximation of the L_2 norm of the log spectral distortion by first N terms.

$$d_{cep}(x,x') = \sum_{i=-N}^{N} (c_i - c'_i)^2$$

*2 other kinds of distortion measure a priori bad are tested: euclidean distance of linear prediction coefficients and autocorrelation coefficients (from LPC preprocessing).

*Weighted slope metric proposed by Klatt is a perceptually based distortion measure[3].

$$d_{wsm}(x,x') = Ke|E-E'| + \sum_{i=1}^{Q} K(i)*[S(i)-S'(i)]^2$$

Ke and K(i) are coefficients. We take Ke=0,K(i)=1 (according to [9] error is minimal with these values). Here Q=18 (Some values differ from Klatt).

*Another perceptually based distortion measure was proposed by Plomp[5]. Late it was used by Carlson (1979) and Blomberg(1983).

$$d_{plm}(x,x') = \left(\sum_{i=1}^{Q} |L_i - L'_i|\right)^{1/p}$$

where L_i is critical band spectra in band i and p=1 or 2.

*Another simple slope distortion measure (called here D_{hsl}) is defined by a Hamming distance on a set of Fn parameters[4]. Where

F_n = 1 if X(n+1)>=X(n)and X(n+1)>=threshold
0 otherwise

and X(n) is smoothed spectrum either in linear or in Mel frequency scale.

A classic method of evaluation of different distortion measures is to test them in a recognition system. So one can judge their performance according to their error percentage of recognition. This is often expensive and time consuming.

Psychoacoustic Tests

A test of psychoacoustics has been designed to produce pertinent histograms which can be easily compared with the curves of distortion measures.

The test was carried out with steady state synthetic vowels. 12 pairs of french vowels have been chosen. Each pair vowel is close so that there is not a third vowel between the vowels of a pair. A series of 11 sounds has been synthesized for each vowelpair by linear interpolation of their formants. The data of formants are from Mrayati (1976).

During the test, an auditor had to listen to the previous series of sounds between 2 references (these 2 references are phonetic references, that is vowels labels and the sounds were not given in the test) and discriminate every presented sound to one of the 2 asked references with forced choice. 12 histograms have been built with 9 auditors from 132 sounds (12*11).

In fact this is a similarity measure of the tested sound to vowels. Auditor will discriminate a sound to one class if it seems more similar to its reference than another one.

Distortion Measure Curve

The same signals have been used for distortion measure calculation. For reason of comparison we calculate

$$D_s(x,V1,V2) = d(x,V2) - d(x,V1)$$

where V1, V2 are 2 references and x is any sound of the series of sounds synthesized by linear interpolation between V1, V2. d is a distortion measure.

The evaluation is made by correlation and percentage of errors which will be defined in next section.

III. EXPERIMENTAL RESULTS

Normalized Correlation Measure

It is often used to compare a distortion measure and human perception.

If x and y are regarded as Euclidean vector,
$r = \cos\theta = (x,y) / (||x||.||y||)$

Percentage of Error

*For human perception, there is a statistic frontier (an arrow below) between 2 vowels. Every auditor made some error with respect to this frontier. The mean of this error for all auditors is denoted by E_h. For example, a histogram is presented in Fig 1, E_h is the sum of shaded region.

*For distortion measure, a percentage of error E_{dis} is defined. This percentage is computed by the ratio of 2 lengths: the length of the interval between the distortion measure frontier (zero crossing) and human statistic frontier, and the length between 2 references in Fig 2.

Some Figures

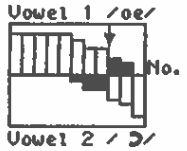


Fig 1.

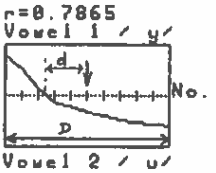
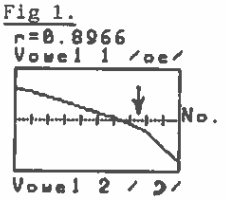
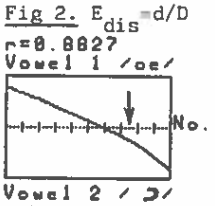


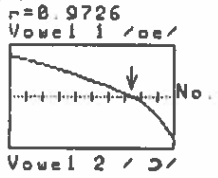
Fig 2. $E_{dis}=d/D$



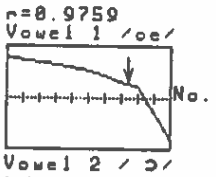
(a) Itakura



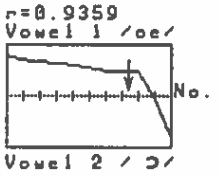
(b) Cepstral



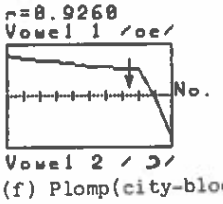
(c) Ri



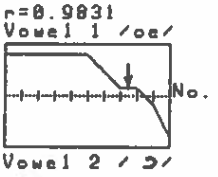
(d) Ak



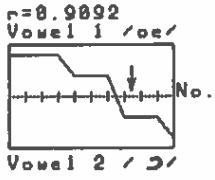
(e) Plomb(euclidean)



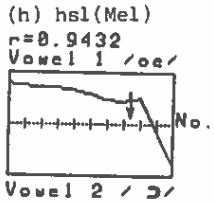
(f) Plomb(city-block)



(g) hsl(linear)



(h) hsl(Mel)



(i) WSM

Fig 3.
↓ is statistic frontier of test

Some Results

We present here a part of results about the correlations and the percentages of errors. All results are means of 12 tests. This correlation is between all points of 2 curves: D_s and histogram of human perception.

	Correlation	Percentage
Auditors		9.6%
Itakura	0.857	10.1%
Cepstral	0.832	13.7%
Plomp(city-block)	0.824	15.3%
Ri(euclidean)	0.77	17.0%
Plomp(euclidean)	0.80	17.8%
Hamm. slop(linear)	0.74	18.3%
Hamm. slop(Mel)	0.77	19.6%
WSM	0.70	21.0%
Ak(euclidean)	0.64	22.6%

Another type of correlation can be computed from the different frontiers. For example correlation between frontiers of Itakura and these of cepstral over 12 tests is 0.993, it corrresponds to an angle of 6.7°; and correlation between Itakura and Ri is 0.9, it corresponds to an angle of 25.8°.

IV. CONCLUSIONS

The main mathematical distortions are better than perceptually based distortions but the test we have done is favourable to mathematical distortions (the sounds vary by formants shifts only). As it was expected the Ak coefficients are not good ones. Sometimes very bad frontiers are obtained which are difficult to explain. A very high correlation between Itakura and Cepstral measures is observed.

The most difficult choice, in this work, is the set of formants of references. The chosen set is considered as representative of french vowels. Surprisingly it is very well adapted to Itakura distortion.

REFERENCES

1. R.M.GRAY, A.BUZO, A.H.GRAY, Y.MATSUYAMA, "Distortion Measure for Speech Processing" IEEE Trans. ASSP-28, No.4, pp367-376, 1980

2. P.L.CHU, D.G.MESSERSCHMITT, "A Frequency Weighted Itakura-Saito Spectral Distance" IEEE Trans. ASPP-30, No.4, pp545-560, 1982

3. D.H.KLATT, "Prediction of Perceived Phonetic Distance from Critical Band Spectra: a first step" ICASSP pp1278-1281, 1982

4.T.K.VINTSJUK,A.G.SHINKAJ,apco8,lvov,3,pp19-24,1974 (in russian)

5. R.PLOMP, "Timbre as a Multidimentional Attribute of Complex Tones" FREQUENCY ANALYSIS AND PERIODICITY DETECTION IN HEARING Ed. PLOMP, 1970

6.K.SHIKANO, M.SUGIYAMA, "Evaluation of LPC Spectral Matching Measure for Spoken Word Recognition" Trans. IECE, Vol. J65-D, No.5, pp535-541, 1982

7. R.VISWANATHAN, J.MAKOUL, W.RUSSELL, "Towards Perceptually Consistent Measure of Spectral Distance" ICASSP pp485-489, 1976

8. M.ESKENAZI, J.S.LIENARD, "Recognition of Static State French Sounds Pronounced by Several Speakers; Comparison of Human Performance and an Automatic Recognition Algorithm" Speech Communication 2(1983) pp173-177

9. N.NOCERINO, F.K.SOONG, L.R.RABINER, D.H.KLATT, "Comparison Study of Several Distortion Measures for Speech Recognition" Speech Communication 4(1985) pp317-331

10.F.ITAKURA, "Minimum Prediction Residual Principle Applied to Speech Recognition" IEEE ASSP-23, No.1, 67-72, 1975

11. M.MRAYATI, Contribution aux Etudes sur la Production de la Parole, Thése d'Etat, INPG 1976