Determining the dimensions of speech quality from PCA and MDS analysis of the Diagnostic Acceptability Measure

D. Sen
AT&T Laboratories Research
Fax: +973 360 7111
Email: dsen@research.att.com

Keywords: Multidimensional Speech Quality

Summary

Subjective judgment of quality is based on an internal multidimensional perceptual representation. That the *quality* of any substance, whether it be synthesized speech or a beverage, is multidimensional should not surprise anyone. A beverage for instance might be described using the dimensions of *saltiness*, *sweetness*, *temperature*, *dryness*, etc. Different subjects, asked to judge the *overall* quality of the beverage, will judge it in different ways based on their unique tastes of those dimensions. To deterministically predict the abstract notion of *quality*, the logical approach would be to predict constituent dimensions that contribute to the overall quality. However, prevalent objective measures of speech quality [1] attempt to predict the *quality* unidimensionally, without consideration for various characteristics of the speech material. In this study, a database of Diagnostic Acceptability Measure (DAM) scores is analyzed to determine the dimensions of speech quality.

Introduction

The speech/audio stream from any source has many attributes. Different subjects will rate the *quality* of the stream depending on their individual, unique and often time-varying tastes of those attributes. The assumption, often taken for granted, that there should be some correlation in judgments from different subjects, is not at all clear. Most objective measures of speech quality are however based on that assumption. These measures are thus only valid if the synthetic speech is limited to a single type of distortion. A slight broadening of the distortion characteristics would result in the failure of these objective measures. This is evidenced when the objective measures are used on very low rate coders, speech corrupted by additive background noise or systems with channel errors [1][2].

A measure designed to predict specific attributes and characteristics of the speech signal would allow more accurate evaluation of speech synthesis systems and a much wider variety of system distortions. This would enable synthesis algorithms to be designed to minimize particular distortions and perhaps allow tailoring speech systems toward particular audiences and environments.

In order for objective measures to extract specific attributes of speech characteristics, we need to know the constituent dimensions of *quality*. In 1977, Voiers [3][4] described a subjective measure of quality that is based on the ability of a group of listeners to detect different types of distortions. The measure, termed Diagnostic Acceptability Measure (DAM), is based on the assumption that listener responses about the detectability of particular distortions are far more likely to be correlated than their opinions of the overall quality. The distortions that the listeners are asked to detect however are not orthogonal and many of them are indeed highly correlated. The rationale behind using correlated feature sets is to distribute the error allowing the speech sample to be located precisely in the quality space. In this work, I analyze a database of DAM scores in an attempt to determine the dimensions of the speech quality space. Knowledge of the dimensions will enable the future design of objective measures which will essentially aim to predict the location of speech sample in the *quality space.*

Principal Component Analysis

The DAM database I used is made up of 56 speech synthesis systems of various speech coding algorithms, whose rates range from 1.2 kbps to 16 kbps. The source speech material includes clean speech as well as speech corrupted with various background noise such as HMMWV military vehicles and office noise. Each system has six sets of DAM scores for the three male and three female speakers that were used to record the original set of speech samples. While the DAM has parametric ratings for both signal and background quality, we have concentrated on the signal ratings, for this particular study. The eight signal quality parameters are described in Table 1.

We used Principal Component Analysis (PCA) to find orthogonal dimensions in the quality space. In Figure 1, the data are plotted in the Component One (PC-1) versus PC-2 space. A plot showing the amount of variability explained by each of the principal components is shown in Figure 2. The first two components account for 70% of the variance while the third component accounts for another 10% of the variance in the data. This vindicates the multidimensionality of the speech *quality space.*

The first Principal Component is mostly weighted by the SF, SB and SI parameters and least by the SH parameter. The second principal component, on the other hand, is weighted mostly by the SH parameter and least by the SF, SB and SI parameters. Plotting the weights for each of the principal components allows us

to visualize the *quality space*. In Figure 3, the weights of PC-1 vs PC-2 have been plotted, while in Figure 4, a 3-dimensional plot of the weights of PC-1, PC-2 and PC-3 have been plotted.

	Description	Example
SD	Harsh	Peak Clipped Speech
SI	Interrupted	Packetized Speech with Glitches
SF	Fluttering	Interrupted Speech
SB	Babbling	Systems with Errors
SH	Thin	High Passed Speech
SL	Muffled	Low Passed Speech
ST	Thin	Band Passed Speech
SN	Nasal	2.4 kbps Systems

Table 1: Signal Parameters in the DAM test

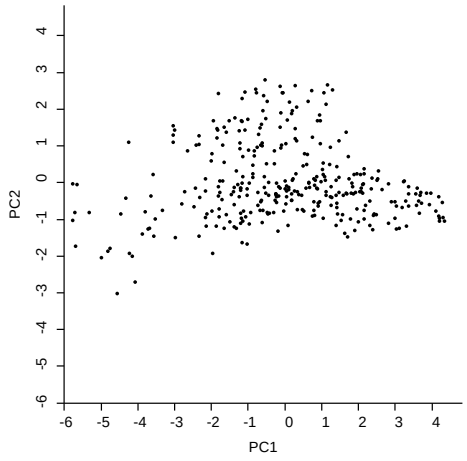

Figure 1. Principal Component One vs Principal Component Two.

From the graphical representation of Figure 3 and 4, the quality space can be seen to be separated into temporally localized and frequency localized distortions. The temporally localized distortions seem to be confined within the SD (*harsh*) parameter on one end and the SB (*babble*), SI (*interrupted*) and SF (*fluttering*) parameters which form a tetrahedron. The difference between SD and the {SB, SI and SF} parameters can be interpreted to be in the distribution of the temporal distortions. A harsh effect is perceived when the temporal distribution of the distortions is dense as opposed to a sparse temporal distribution which excites the {SF, SI and SF} parameters.

The SH (*high passed*) and SL (*low passed*) parameter seem to form the vertices of the frequency localized distortion space. The position of the remaining two parameters, ST (*thin*) and SN (*nasal*) in Figure 4 seem to indicate that they are excitable both by temporal and frequency localized distortions.

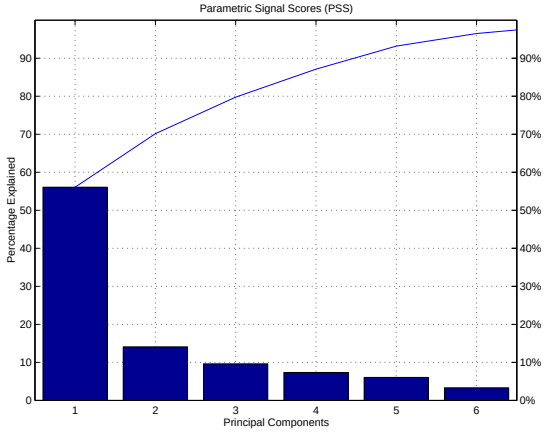

Figure 2: Amount of variability explained by each of the principal components.

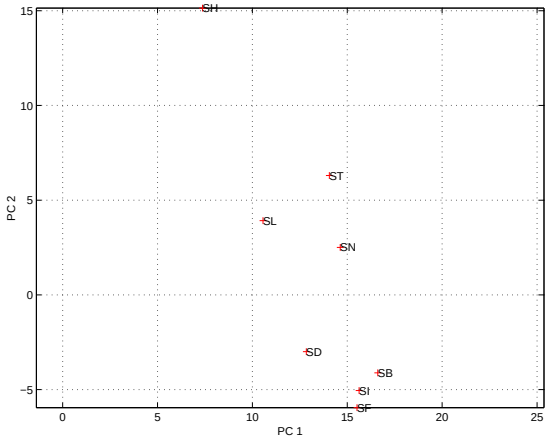

Figure 3: Weights of Principal Component One and Two.

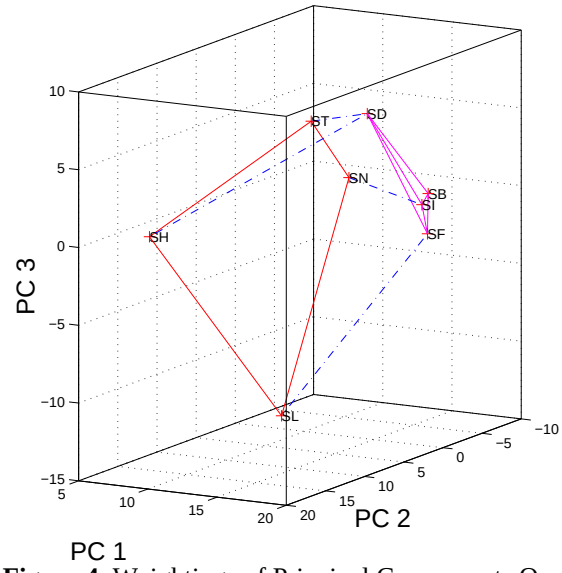

Figure 4: Weightings of Principal Components One, Two and Three.

Multidimensional Scaling

To further elucidate the quality space, a Multidimensional Scaling (MDS) analysis was carried on the DAM database. Figure 5 shows the result of the

3-dimensional Kruskal-nonmetric MDS analysis, with a stress of 0.0001 percent. In this figure, the distance between the features indicate their relative correlations such that highly correlated features will be located in close proximity to each other.

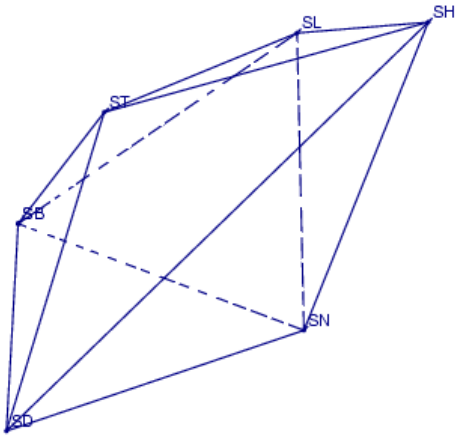

Figure 5: Three-Dimensional Multidimensional Analysis of the DAM database.

As in the PCA Analysis, SH and SL seem to form the vertices of the frequency localized distortion. Unlike the PCA analysis however, the subset {SB, SI, SF} collapse into one point (SB on the figure) due to their high relative correlations. These, along with SD form the vertices of the temporally localized distortion space. Finally, the position of the ST and SN nodes, again suggest that they are excitable by both temporal and frequency localized cues.

Discussion

The PCA and MDS analysis serve to portray the multiple dimensions of the speech *quality space*. In particular, it shows that an objective measure which is able to evaluate speech corrupted by a variety of distortions and conditions, would need to predict various dimensions of distortions rather than the unidimensional approach taken in current approaches. An even larger number of dimensions would be required if background-noise characteristics (neglected in the present study) are to be predicted. In order to predict the individual dimensions, cognitive models would be required to isolate the cues which excite the particular dimensions. Further discussion of how these cognitive models might be designed is beyond the scope of the present paper.

The parameters chosen by the DAM subjective measure can be broadly classified into the time-localized and frequency-localized distortions. The choice of the highly correlated features, SB, SI and SF, which seem to be cues for sparsely distributed temporally localizable distortions, can be explained by the high number of speech samples in the database which lie in that space. By using multiple, correlated descriptors, the designers of the DAM have effectively zoomed into that part of the quality space. Figure 6 shows how the use of non-orthogonal axes effectively distributes the errors and provides a more accurate way of locating a point in space. To illustrate the effect, we assume that there is very little variance in the *x* dimension (or that it is easy to predict). In comparison, the *y* dimension accounts for a large variance, requiring a very precise measurement of that dimension. This effect can be approximated by assuming $x \ll y$. This makes *l*, the distance of the point from the origin, approximately equal to *y*. An *n*% error in *y* is thus reflected as an *n*% error in *l*. If a non-orthogonal axis *P*, is used instead of *Y*, $p = y\cos(\theta - 90)$. Similarly, an *n*% error in *p* is scaled to produce a reduced $n\cos(\theta - 90)$% error in *l*. Multiple non-orthogonal axes in multiple dimensions may thus be used for a more accurate estimation of the location of the point.. In the absence of an exact descriptor for an orthogonal axis, the use of multiple correlated descriptors to distribute the error can well be appreciated.

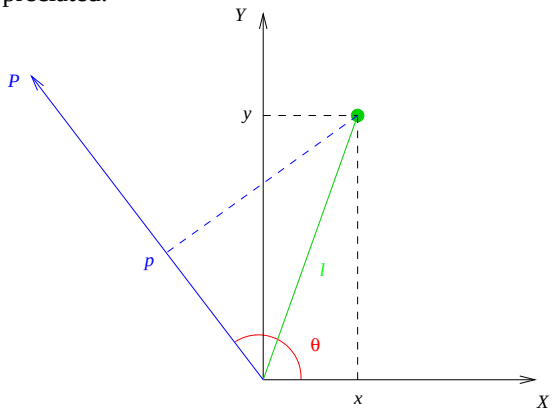

Figure 6: Figure depicting the effect of using non-orthogonal axes.

Acknowledgment

The author would like to gratefully acknowledge John Collura (of NSA) Alan Sharpley and Ira Panzer (both of Dynastat). The author would also like to thank Schuyler Quackenbush, Andreas Buja and Richard Cox (of AT&T labs research) for their words of encouragement.

References

[1] A.W. Rix, J.G. Beerends, M.P. Hollier and A.P. Hekstra, "PESQ - The New ITU Standard for End-to-End Speech Quality Assessment", 109th AES Convention, Pre-print No. 5260, September 2000.

[2] J.G. Beerends and J.A. Stemerdink, “A perceptual speech-quality measure based on a psychoacoustic sound representation”, Journal of the AES, 42 (3), pp 115-123, March 1994.

[3] W.D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems". Proc. ICASSP, pp. 204-207, 1977.

[4] S.R. Quackenbush, T.P. Barnwell III, and M.A Clements, “Objective measures of speech quality”, Prentice Hall, NJ, USA, 1988