PREDICTING FOREGROUND SH, SL AND BNH DAM SCORES FOR MULTIDIMENSIONAL OBJECTIVE MEASURE OF SPEECH QUALITY.

D. Sen

University of New South Wales, Sydney, Australia.

ABSTRACT

Current objective measures of speech quality [1,2] attempt to evaluate degraded speech by calculating a single distance measure between the original signal and the synthesized signal being evaluated. The distance measure is usually carried out after both the original and synthesized signal have been transformed to represent the effect of the auditory periphery. However, the fact that the subjective judgement of quality is based on a multidimensional perceptual space representation suggests that a measure that is based on predicting a multitude of independent perceptual characteristics, would yield better results and be applicable to a wider range of distortions and speech synthesis systems. This paper presents such a multidimensional approach to objective evaluation of speech quality and is directly motivated by the work of Voiers [3] from which the subjective evaluation procedure known as Diagnostic Aceptibility Measure (DAM) was created. While the DAM is a subjective measure of the detectability of the distortions identified by Voiers, this work reports on the first baby steps taken for objective evaluation of a subset of those same parametric distortions determined to be the principal components of the quality space from a previous statistical analysis [4].

1. 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 distortion in the synthetic speech signal is limited to a single type. A slight broadening of the distortion characteristics would result in the failure of these objective measures. This is evidenced when the objective measures fail to evaluate 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*. To this end, Voiers identified a number of different distortions or parameters [3] which he used to develop the DAM subjective measure of speech quality. The subjective test involves asking a number of listeners to detect the different types of distortions and 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. While in more recent work [5], it has been shown that only a few dimensions are required for the subjective evaluation of speech quality, the rationale behind using a multitude of different but correlated feature sets may have been to distribute the error allowing the speech sample to be located precisely in the quality space.

In previous work [4], we have analyzed a database of DAM scores (DAM-IIc scores obtained from Dynastat Inc.) in an attempt to determine the dimensions of the speech quality space. In the next section, we present the essential results of that work required to explain the motivation and methodology of the current work. In the subsequent section, we describe algorithms which attempt to isolate and thus independently evaluate the detectability of a small subset of these dimensions of speech quality. Results are presented in Section 4.

2. STATISTICAL ANALYSIS OF DAM SCORES

The DAM scores used for the statistical analysis is made up of 56 speech synthesis systems of various speech coding algorithms, the rates of which 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-babble 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 bothforeground signal (SQ) and background quality (BQ), we have concentrated on the signal ratings for this initial study. Some of the foreground signal quality parameters and background parameters are described in Table 1. More detailed information on the DAM may be available from [3] and [6].

We used Principal Component Analysis (PCA) to find orthogonal dimensions in the quality space. A plot showing the amount of variability explained by each of the principal components is shown in Fig. 1. 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 Fig. 2, the weights of PC-1 vs PC-2 have been plotted, while in Fig. 3, 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
BNH	Background Hiss	Background White Noise
BNL	Background Muffled	Background Muffled Noise

Table 1: Signal Parameters in the DAM test

From the graphical representation of Fig 2 and 3, 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 Fig 4, seem to indicate that they are excitable both by temporal and frequency localized distortions.

A Multidimensional Scaling (MDS) Analysis reinforced the findings of the above PCA Analysis [4].

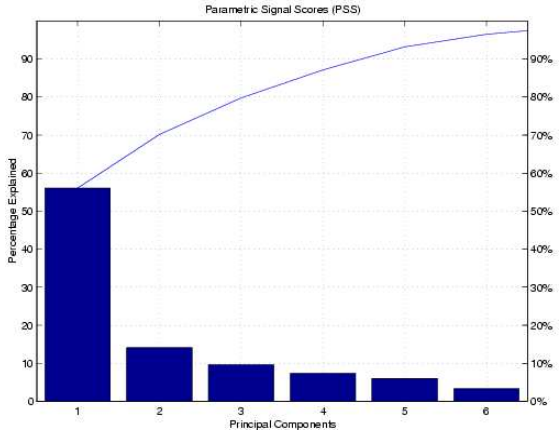

Figure 1: Amount of variability explained by each of the principal components.

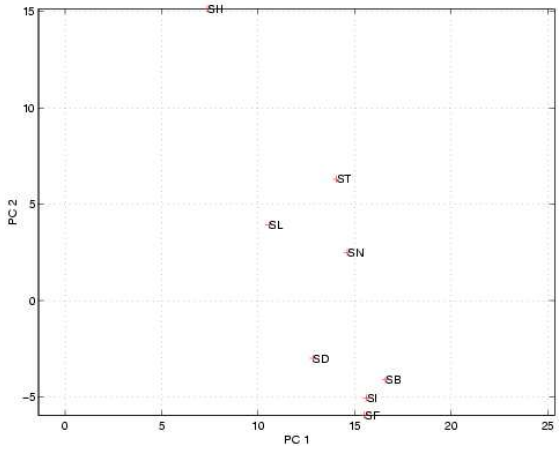

Figure 2: Weights of Principal Component One and Two.

3. OBJECTIVE DETECTION OF DAM FOREGROUND SCORES

Both the PCA and MDS results reveal that the SL-SH vertices form one of the three principal axis of the quality space. SH has the highest weighting for PC2 whereas SL has the highest weighting on PC3. PC2 and PC3 combine to explain 25% of the variability of the DAM scores (Fig.1). Since the perception of "low-pass" and "high-pass" distortion is localized in the frequency domain, it

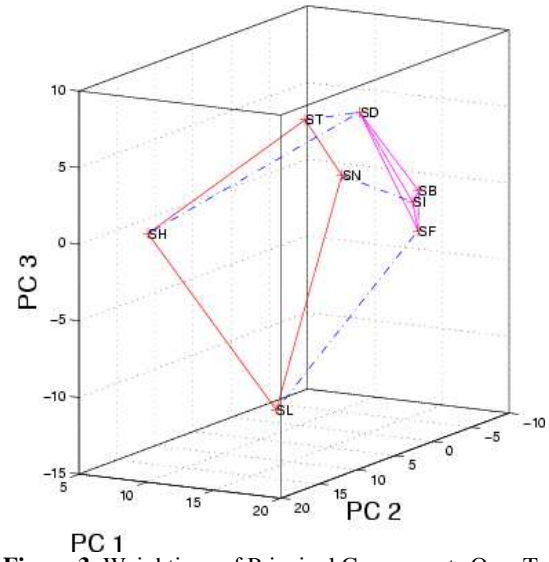

Figure 3: Weightings of Principal Components One, Two and Three.

ought to be possible to isolate these distortions from the output of a cochlear model where the signal is essentially resolved into frequency components along the length of the cochlea. This is thus the motivation for the procedures depicted in Fig. 4 (and described in the following paragraphs) to predict the detectability of the SL, SH and BNH distortions. While other parameters carry more weight in the quality space, the algorithms required to isolate the time localized distortion parameters will require further work outside the scope of this initial pilot study into the feasibility of a multidimensional measure.

The cochlear model used for this work is a non-linear two-dimensional model of the cochlear hydrodynamics incorporating a model of the Outer Hair Cell motility [7]. It is envisaged that the use of this physiologically accurate model of the auditory periphery will yield better results than the usual functional linear models of hearing used in [1] and [2].

In order to isolate individual distortions, we had to make a number of assumptions. The first hypothesis was to assume that the perception of background quality distortion was due to additive noise that is not correlated with the speech signal. A simple model to extract the uncorrelated noise from the correlated noise is to sample the "silence periods" between sentences. This background noise subtracted from the segments where there is voice-activity would "eliminate" the BQ distortions from the SQ distortions. The second hypothesis for the cognitive model is that the perception of foreground signal distortions (SQ parameters) is mainly carried out in the voiced regions of the signal rather than unvoiced regions. Since the statistical characteristics of the voiced segments are usually quite distinguishable from noise, this hypothesis is not totally unfounded.

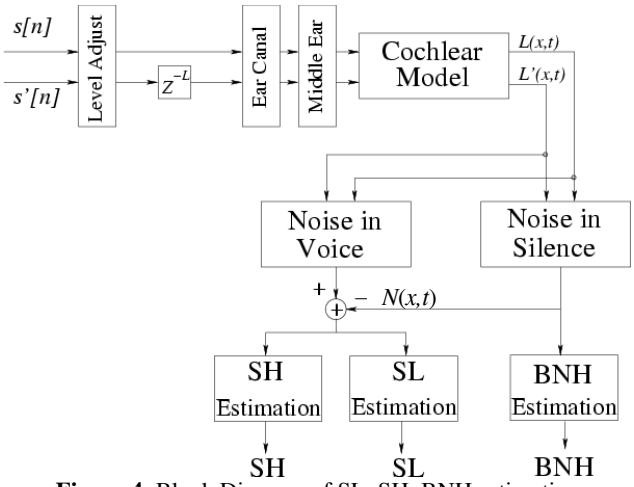

Figure 4: Block Diagram of SL, SH, BNH estimation algorithm.

The original and sytnthesized signals ($s[n]$ and $s'[n]$) were pre-processed to carefully match the pressure levels presented to the subjects (this is required as the non-linear cochlear model is sensitive to pressure levels), time-synchronized (with sub-sample resolution denoted by Z^{-L} in Fig. 4) [8] to account for delays in the synthesis/coding algorithms and put through an external ear model. Following the pre-processing, the auditory model is used to convert the pressure signal to Inner Hair Cell (IHC) responses. The IHC response was squared to produce a quantity which resembles instantaneous partial loudness, $L(x,t)$ ($L'(x,t)$ for the synthesized signal). Here x represents the length along the cochlea which has a one-to-one mapping with frequency (given by the cochlear map) and t is time. High frequencies are represented by small values of x due to the nature of the cochlea where high frequencies are resolved at the basal end of the cochlea. A low-pass operation thus involves considering $L(x,t)$ at the higher values of x while a high-pass operation looks at the lower values of x.

To ensure that the additive uncorrelated noise hypothesis for the BQ scores was correct, we averaged the high frequency $L(x,t)$ response difference (in the silence periods) and performed a linear regression with one of the high frequency BQ (BNH) parameters. The cutoff frequency/position X_{THI}, was found by maximizing a correlation cost function. As revealed in Fig. 6, the correlation between the predicted and subjective BNH parameter was very high (R=0.995). This highly precise measure (useful in itself), is capable of resolving some very fine differences between speech coders. The predicted BNH is thus given by,

$$P_{BNH} = \sum_{t\,=\,silence} \sum_{x\,=\,0}^{X_{THI}} |L(x,t) - L'(x,t)| \qquad (1)$$

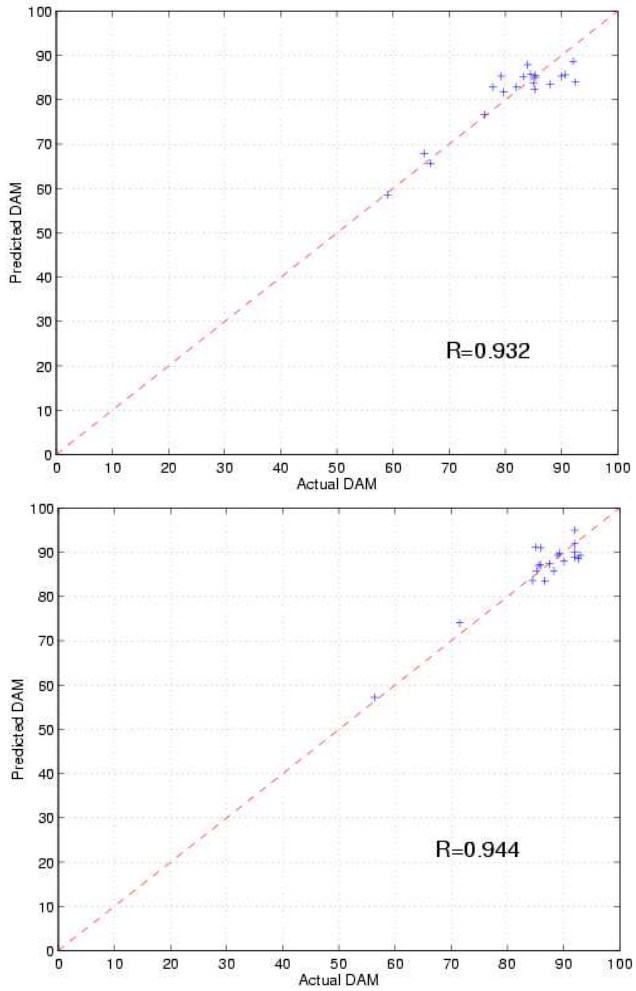

Figure 5. Scatter plot predicted vs actual SH and SL scores.

To isolate the SL-SH distortions from the background distortions, we first subtract the background noise energy $N(x,t)$, estimated in the silence regions of the signal (see Fig. 4). Next, we estimated this isolated SQ noise in voiced sections of the speech signal. If low-pass distortions are being perceived then there has to be large high frequency errors. Thus the SL score can be estimated using,

$$P_{SL} = \sum_{t=voiced} \sum_{x=0}^{X_{TH2}} |L(x,t) - L'(x,t) - N(x,t)| \quad (2)$$

Similarly, the SH score can be estimated using,

$$P_{SH} = \sum_{t=voiced} \sum_{x=X_{TH3}}^{X_N} |L(x,t) - L'(x,t) - N(x,t)| \quad (3)$$

The cutoff frequencies/positions X_{TH1} and X_{TH2} were again determined by maximizing the correlation cost function. X_N is the largest value of x-index from the cochlear model.

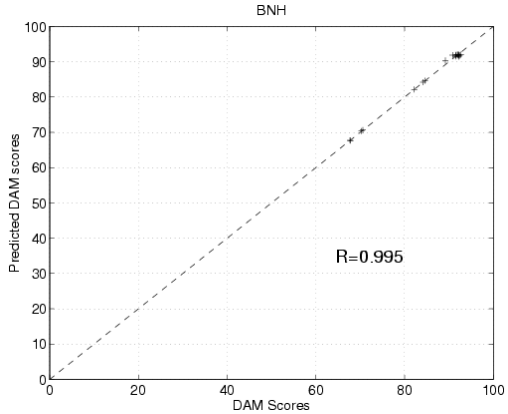

Figure 6. Scatter plot predicted vs actual BNH scores.

4. RESULTS AND CONCLUSION

A scatter plot showing the predicted SH, SH and BNH scores versus actual scores is shown in Figures 5 and 6 respectively. The results show the values for data which were outside of the training set during calculation of the regression coefficients. The Pearson correlation coefficient between the predicted and actual DAM scores were 0.932 and 0.944 for the SH and SL parameters respectively and 0.995 for the BNH score. Future work will focus on predicting temporally localized distortions to predict scores such as SD and SF which account for a large portion of the variability. The current results seem to suggest that a multidimensional objective meausre may well be viable.

5. 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] D. Sen, "Determining the dimensions of speech quality from PCA and MDS analysis of the Diagnostic Acceptability Measure," *MESAQIN*, 2001.

[5] J. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *JASA*, **110**, pp 2167, 2001.

[6] Dynastat Inc., "The DAM-IIc Diagnostic Acceptability Measure", *Dynastat*, 2000.

[7] D. Sen and J.B. Allen, "Benchmarking a two dimensional auditory model against experimental auditory data," *Midwinter Meeting, Association for Research in Otolaryngology (ARO)*, St Petersburg, Feb 2001.

[8] S. Voran, "An Algorithm for estimating the delay of telephony speech," *ITU-T SG-12*, DOC SQ-75, 1996.