A MULTI-DIMENSIONAL APPROACH TO PREDICTING SPEECH QUALITY USING A PHYSIOLOGICALLY MOTIVATED MODEL OF THE COCHLEA

D. Sen, Wenliang Lu

School of Electrical Engineering and Telecommunication
University of New South Wales, Australia

ABSTRACT

This paper introduces a framework to measure speech quality in a multi-dimensional space. A hydro-mechanical Cochlear model is used to convert speech into the perceptual domain, where salient features are extracted to detect temporal and frequency localized distortions. Overall quality evaluation is subsequently achieved using a weighted sum of the detectability of individual distortions.

Index Terms—Objective measurement of speech quality, Diagnostic Acceptability Measure

1. INTRODUCTION

Existing objective measures of speech quality provide a single dimensional output - a score that represents the overall perceived quality of the degraded speech signal (or system that produced the signal). The single output is ideally meant to replicate subjective Absolute Category Rating (ACR) scores. However, the design of objective measures has often ignored the fact that the subjective evaluation process is multidimensional - where the listener characterizes the signal based on the perceptibility of various distortions. Research into the dimensionality of speech quality [1, 2, 3] have established that at least three orthogonal dimensions are required for accurate representation of speech quality.

The principle hypothesis of the current approach is to replicate the listening process and analyse the signal based on the detectability of various distortions. While this multidimensional approach is novel for objective measurement of speech signals, the approach is central to the proprietary subjective Diagnostic Acceptability Measure (DAM) [1, 4]. In DAM, listeners are asked to focus on the detectability of individual distortions (or Elementary Perceptual Qualities - EPQs) - the premise being that this makes for an easier subjective task thereby improving reliability and repeatability of the evaluation. The DAM, in addition to the detectability of individual distortions, also provides “Composite Acceptability Estimate” (CAE) score that is computed from a combination of the individual EPQs. In our analysis the CAE has proven to be highly correlated to ACR scores.

The multi-dimensional method described within is unique in that it attempts to decompose the perceptible quality space into multiple orthogonal physiological representations of perceptible distortions. It is thus quite different from methodologies such as the E-Model [5] which attempts to predict quality based on multiple transmission and/or network factors.

The subjective databases used in this paper are DAM scores for 56 different speech degradation conditions. The degradations include environmental noise (helicopter, babble, hmmv noise), transmission noise due to bit-errors (upto 1%) as well as various speech coding algorithms ranging from 2.4 kbps to 32 kbps). Previous research and statistical analysis [3] has determined that the DAM foreground EPQs can be represented fairly accurately (accounting for 80% of the variance) using three dimensions: Temporally localized distortions (TLD); Frequency localized distortion (FLD); and distortions that do not directly belong to the above two groups. The three groups contribute 55%, 15%, and 10% of the total variance, respectively.

As it is impossible to come up with exact lexical descriptions of distortions which would be perfectly orthogonal to each other, we have instead attempted to cover the three dimensional space with four different distortions - which while not perfectly orthogonal to each other - does span the three dimensional space adequately. These four foreground distortions (two FLDs and two TLDs) along with a single background noise EPQ parameter which comprise our objective quality space are described below:

- **FLD Parameters:** *SH* and *SH* (Low and High Frequency suppression).
- **TLD Parameters:** *SB* and *SD* (Slow and rapidly varying discontinuity).
- **BN-Background noise.**

Analysis by other researchers [2, 6, 7, 8, 9, 10] have all come to a similar conclusion as to the number of dimensions (three) required to span the speech quality space. The different databases covering different types of distortions will of-course have affected those analysis. While our choice of using five distortions might seem to contradict both our own
statistical analysis and those of others, it is justified by the fact that while the four lexical descriptions are not perfectly orthogonal to each other, they are highly uncorrelated — and thus cover the three dimensional space more adequately than would a three lexical descriptions of distortions.

The second novel approach of the current work involves the use of a spatial non-linear 2D hydro-mechanical model, which computes various electrical and mechanical response in the Cochlea. The Inner Hair Cell (IHC) response from the model is used in this work. The model is implemented in the time domain and provides excellent time-frequency resolution. Details of the model have been published in [11, 12] and will not be discussed further here. Contemporary objective measures of speech quality invariably use psychoacoustic auditory models (PAM) to convert speech signals to the perceptual domain. However, even the most evolved PAMs [13] are unable to provide both high time resolution and high frequency resolution simultaneously. Moreover, the PAMs used in these systems are only approximations to actual physiological cochlear behaviour and almost always ignore the highly non-linear behaviour of the physiological cochlea. The secondary hypothesis of this work is thus that the replacement of a PAM by a physiologically motivated cochlear model will provide a better representation of what is perceived — and hence provide a more accurate front-end to any attempt at predicting speech quality. This allows for the accurate extraction of salient features, especially TLD.

A third hypothesis of this work is that speech quality is mostly discerned when listening to voiced sections of speech. This hypothesis has also been alluded to various speech perception body of research [14]. Voiced sections typically have high-energy, and are of long durations, compared to other sections. They maintains a state of “steady-state”, which is especially important for TLD measurement.

The paper is structured as follows: The following section discusses salient feature extraction from the cochlear model and mapping to subjective FLD scores. These include high and low frequency suppressions, respectively. The next section discusses the feature extraction and mapping for temporally localized EPQs. These include slowly and rapidly varying discontinuity, respectively. Finally, the computation of an overall quality score using a weighted combination of the above predicted EPQs is discussed.

2. FREQUENCY LOCALIZED DISTORTIONS

Frequency localized distortions (FLD) includes the SH and SL EPQs. Both the definitions of these EPQs as provided to the listeners and by observation of the distortion that ‘excite’ these EPQs amongst the various degraded signals, indicate that the types of distortions involved are localized along contiguous regions along the frequency axis. For SH, the discrepancies are in the low frequencies or in apical regions of the cochlea whereas for SL the discrepancies are in the high frequencies (or in basal regions of the cochlea). The methodologies for predicting SH and SL are described in detail in [15] and depicted in Fig. 1.

Briefly, the spatio-temporal output from the CM is averaged over time over temporal windows of 5ms. This has the effect of smoothing out any temporal distortions and produces a spatial (frequency) dependent output. The output is analyzed in specific spatial lengths along the cochlea. For SL distortion, the length of the cochlea closest to the oval-window is analyzed for high-frequency distortions. Similarly, for SH, spatial locations far from the oval-window are analyzed for low-frequency distortions. The analysis is tempered by the type and amount of background noise (measured in silence regions). Results are reported in Section 5.

3. TEMPORALLY LOCALIZED DISTORTIONS

The aim of this section is to isolate and predict the detectability of temporally localized distortions such as single “clicks” and “pops” but also more temporally dense distortions which produces the perception of “harshness”. Generally speaking and in line with the PCA analysis [3], the distortions can be classified into a “slow” and “rapid” mode. The methodology
for their extraction is shown in Fig. 2. Detailed description of the methodology can be found in [11]. Briefly, a peak tracking algorithm on the 3-Dimensional cochlear response (as shown in Fig. 3), followed by “centre of mass” computations over pitch periods and the spatial axis produces the so called “Salient Feature Points” or SFPs. Distance between the original and degraded SFPs are calculated, to estimate the ‘slow’ (SB) and ‘fast’ (SD) components. The results from all voiced sections are averaged and used as the predicted outputs.

![Fig. 3. Cochlear model response for vowel /o/ showing the peak tracks.](image)

### 4. OVERALL QUALITY

With the successful prediction of FLD and TLD, it is possible to combine the perceptibility of individual distortions to correlate with subjective estimates of “overall quality”. We have chosen in this work to only predict the overall foreground quality. Thus, the “Background Noise” parameter (BN) is not used and only the predictions of FLD and TLD are used to estimate overall quality (ACR). An analysis of our DAM subjective database showed that the correlation between ACR scores and multiple-regression of foreground DAM parameters was $R = 0.887$. We thus do not expect achieve any higher with our objective prediction.

It is entirely possible that different subjects and demographics will weight the distortions differently in their evaluation of total quality. Nevertheless, if our original premise of being able to extract individual distortions more accurately than overall quality is correct, then it should be possible to predict overall quality – at least more accurately than objective measures which attempt to predict overall quality without any regard to individual distortions.

This paper attempts to predict ACR scores by a simple weighted average of the individual distortions. The weights are calculated by a training procedure. We have used half the available data to train for the weights and then used those weights to predict the rest of the half. The results plotted in Section 5, represents the test data points only (not the training data). The weighting was carried out on the raw scores and not the regressed fit.

The algorithm is open to more complex statistical models to perform this task. The weighting may also be a function of the type and amount of background noise present in the signal. This would however make the weights a function of noise characteristics. Results are reported in Section 5.

### 5. RESULTS

The results below represent the scatter plots between the detectability predictions of individual distortions (as described in Sections 2 and 3) and their closest DAM distortion mappings.

Section 5.1 shows the performance of predicting individual distortions while Section 5.2 shows the results of combining the individual distortions to predict ACR.

Figure 4 shows the predictions of the individual distortion types and overall quality. There are a total of 162 points for each of the six speakers over 27 systems. The prediction is attempted at the individual speaker level rather than at a system level (i.e. the mean over all the speakers was not used to measure the performance). Each plot also quotes the regression coefficients for first, second and third order regression.

#### 5.1. Predicting individual distortions [SL,SH,SB,SD]

The plot in Fig. 4 (a) shows the predicted SH distortion described in Section 2 versus the subjective DAM score of SH. The correlation coefficient magnitudes are 0.86, 0.86 and 0.89 for linear, quadratic and cubic regression respectively.

The plot in Fig. 4 (b) shows the predicted SL distortion described in Section 2 versus the subjective DAM score of SL. The correlation coefficient magnitudes are 0.68, 0.77 and 0.83 for linear, quadratic and cubic regression respectively.

The plot in Fig. 4 (c) shows the predicted SB distortion described in Section 3 versus an average of subjective DAM scores for babble, interrupted and fluttery noise (all three of which were highly correlated with each other and is interpreted as slow TLD). The correlation coefficient magnitudes are 0.83, 0.87 and 0.89 for linear, quadratic and cubic regression respectively.

The plot in Fig. 4 (d) shows the predicted SD described in Section 3 versus the DAM subjective score for ‘harsh’ noise. The correlation coefficient magnitudes are 0.69, 0.69 and 0.70 for linear, quadratic and cubic regression respectively.

#### 5.2. Combining individual distortions to predict ACR

Figure 5 (a) is a scatter plot of subjective Mean Opinion Scores (MOS) versus a weighted average of the raw (not the regressed fit) predicted individual distortions (SL,SH,SB and SD). The weights were calculated using multiple linear regression on a training database which was separate from
the test database. The resulting correlation coefficient of 0.78 is to be compared with the maximum possible of 0.887 (as the background distortions are not used) as mentioned above. Figure 5 (b) plots the result of evaluating the same test database using PESQ. The maximum correlation coefficient of 0.39 (using cubic interpolation) is significantly lower than the above multidimensional approach.

Fig. 4. Individual distortions predictions.

(a) SH prediction
(b) SL prediction
(c) SB prediction
(d) SD prediction

7. REFERENCES


6. CONCLUSION

This paper proposes a new framework for objective evaluation of speech signals by predicting the perceptibility of individual distortions. The work follows from previous research [3] which decomposed the quality space into the dimensions of temporally and frequency localised distortions and also previous research [11, 15] which proved the viability of isolating and extracting these distortions. The results presented in this paper vindicates the multi-dimensional methodology by estimating overall speech quality from a weighted sum of the individual distortion of SH, SL, SB and SD. The accuracy of the method supercedes that of PESQ.

When comparing with PESQ it has to be remembered that it requires clean speech at its input (which in the case of speech afflicted by background noise is not readily available) and also the fact that PESQ is not designed for speech codecs below 4 kbps. However, this does not detract from the performance of the proposed method. Future work will involve the addition of background EPQs and improvements to the statistical backend - which it is envisaged will further improve speech quality prediction.