ASSESSMENT OF SUBJECTIVE AUDIO QUALITY FROM EEG BRAIN RESPONSES USING TIME-SPACE-FREQUENCY ANALYSIS

Charles D. Creusere, Jim Kroger, Srikant R. Siddenki, Philip Davis, Joe Hardin

New Mexico State University
Las Cruces, NM
Email: ccreuser@nmsu.edu, pdavis@nmsu.edu, srikants@nmsu.edu, jkk251@gmail.com

Colorado State University
Fort Collins, CO
Email: jhardin@engr.colostate.edu

ABSTRACT

In this paper, we consider the problem of quantifying changes in the perceived quality of audio by directly measuring the brainwave responses of human subjects using a high-resolution electro-encephelogram (EEG). Specifically, human subjects are presented with audio whose quality varies with time while being monitored by a 128-channel EEG; some of the time, they move a slider bar up and down to indicate their perception of the changing quality while at other times they listen passively. Our focus here is to identify low-level features in the brainwave responses that correlate well with temporal quality variations across multiple base audio sequences and different test subjects with our ultimate goal being to implement a classifier based on such features. The results presented here attempt to quantify the quality classification performance versus the perceptual uncertainty of the subjective data. We find that the proposed approach is much more effective in estimating the perceptual quality for one of the two distortion types considered, frequency truncation, than it is for the other type, scalar quantization.

Index Terms— Audio quality assessment, perceptual quality, EEG, brainwave analysis

1 INTRODUCTION

Interest in analyzing how humans perceive audio goes back many years. The groundbreaking work of Flechter in the early 1940s led to the first detailed understanding of the spectral sensitivities of the auditory system [1], although it was not until Johnston developed his perceptual noise masking approach in the 1980s that such concepts were incorporated into a compression algorithm [2]. While human subjects can only consciously rate variations in perceived audio quality to an accuracy of about 1 second, evidence from the neuroscience community suggests that the brain responds to changes in audio stimuli much more rapidly. The first low level processing of auditory information begins as soon as information from the ears reaches the primary auditory cortex, probably on the order of 10-20ms post-stimulus onset. In a variety of different experimental scenarios, it has been found that changes in an audio signal induce changes in the brainwave responses of human subjects in time intervals ranging from 25 to 100 ms [3, 4, 5, 6, 7, 8]. This is important because if we wish to create temporally-dynamic computer-based models of audio perception that have sufficient resolution for coding and transmission applications, we must be able to validate them subjectively. Thus, it should be possible to extrapolate subjective opinions provided by test subjects over longer time intervals to much shorter time frames by identifying common precursors embedded within the EEG waveforms.

Research supported by NSF Grant 1065603

2 EEG ANALYSIS APPROACH

The Active-2 EEG collection system that we used in these trials captures data from 128 electrode channels and samples it at a rate of 1024 Hz. Thus, for each trial we have a very large multidimensional dataset from which we are attempting to extract signals corresponding to changes in the human perception of audio quality. Innumerable approaches are possible from the simplistic—monitoring only changes in alpha wave rhythms [9]—to the highly complex—evaluating phase synchronization behavior amongst widely separated portions of the brain [10]. In this early research, we have chosen an approach based on time-space-frequency analysis of the EEG waveform set. This approach is motivated by the qualitative results for one test subject that are shown in Figure 1. Here, we have plotted space-frequency energy maps comparing the brain responses (with the frequency computed over a three second interval) of a single test subject to impaired and reference audio. Over the time interval selected, the impaired audio being presented to our subject was bandlimited to 5.5 kHz. An average reference was used here, an FIR lowpass filter with a cutoff frequency of 80 Hz was applied, and individual EEG channels from four different presentations of the

test and reference audio were spectrally averaged. Studying Figure 1, one notes that the spatial energy distributions at 6 and 15 Hz are considerably different for the two cases while the differences at 32 Hz are more subtle. Of particular interest is a significant shift in positive (red) power from the left posterior hemisphere, including primary auditory areas, to the posterior right hemisphere processing at 6 Hz, signaling less reliance on routine processing in auditory areas and recruitment of less defined evaluation areas. Beta (15 Hz) activity evidences an even more pronounced rightward shift, now accompanied by a recruitment of right analogs of left linguistic processing areas, as is often seen in non-linguistic or abstract auditory comprehension. Simultaneously, the left negativity (blue) around Brocaõs region for linguistic perception, seen in the reference condition, has disappeared in the impaired condition as presumably the subjects are in a less rote linguistic and auditory processing mode for the impaired audio. In short, this preliminary data is consistent with what might be expected as quality degradation impairs routine processing: subjects are using mechanical auditory processing in the reference condition and shift to a less defined, less constrained, and more global evaluation, typical of right hemisphere.

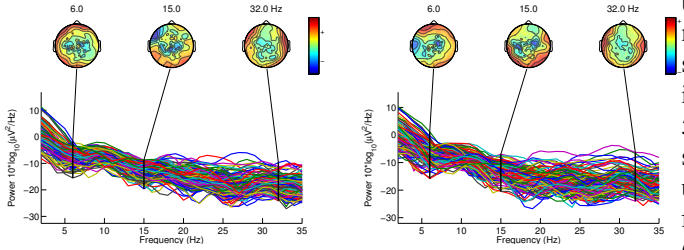

Fig. 1. Space-frequency energy map for reference (a) and impaired (b) audio

Based on these observations, we formulate feature vectors for an EEG-based audio quality classification algorithm as follows. First, for each 5 second time interval of constant audio quality we form 12 128x1 vectors, each containing the power in one of 12 spectral bands (2 Hz to 24 Hz in 2 Hz steps) at the 128 different electrode positions. The set of feature vectors is illustrated in Figure 2. A 1024-point non-overlapping short-time Fourier transform is applied and the spectral power is calculated over each 5 second interval of constant audio quality by averaging. Thus, when taken together our set of 12 feature vectors represents the time-space-frequency distribution of the EEG signal power and it should be able to capture the qualitative differences seen in Figure 1.

3 SUBJECTIVE TESTING METHODOLOGY

The facility used to conduct the human subjective trials was purpose built for EEG research and consists of an RF shielded testing chamber and a 128-channel Active-2 EEG

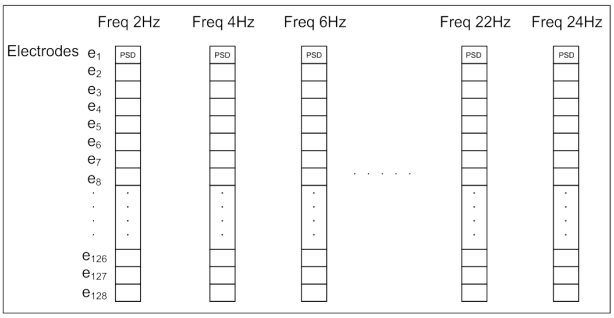

Fig. 2. Feature vector set corresponding to a given time interval.

collection system featuring preamplifiers constructed directly on the electrodes for increased environmental noise immunity. Monaural audio sequences are presented to participants under computer control with subjects being able to repeat any given trial as many times as desired until they are satisfied with their quality scoring (only the final score is retained). Perceived variations in audio quality are input by the subject to the computer via a sliderbar and real-time synchronization markers are embedded with the captured EEG signals by the system. Each trial takes about one hour per test subject (not including preparation and cleanup time) and generates about 5G bytes of data with each of the 128 spatial channels being sampled at 1024 Hz. Three 30 second base sequences were used to create all of the test sequences, each of which was formed by applying one of two different forms of distortion (frequency band truncation or quantization) with each having 3 possible impairment levels and one of two time-varying distortion patterns. Multiple presentations of both the impaired test sequences and the unimpaired reference sequences were made in a randomized fashion. Finally, subjects do not use the sliderbar to rate the distorted test sequences approximately two-thirds of the time and it is these unrated sequences that we use here for our analysis so that the motor cortex responses do not overwhelm the fainter signals of interest. The results presented in this paper are based on data collected from only five test subject and thus must be considered to be highly preliminary.

4 CLASSIFIER DESIGN

Feature vectors are created using the time-frequency-space approach outlined in Section 2. A block diagram of the complete system is illustrated in Figure 3. In this figure, the preprocessing steps applied to the 128 signals are average referencing following by FIR lowpass filtering of each electrode waveform to a 50Hz bandwidth. Average referencing was selected based on analysis presented in [11] proving that this is as effective as any other referencing approach for high-resolution EEG data while the 50Hz cutoff frequency was chosen based our own subjective analysis of the time-space-frequency data. The next two blocks in Figure

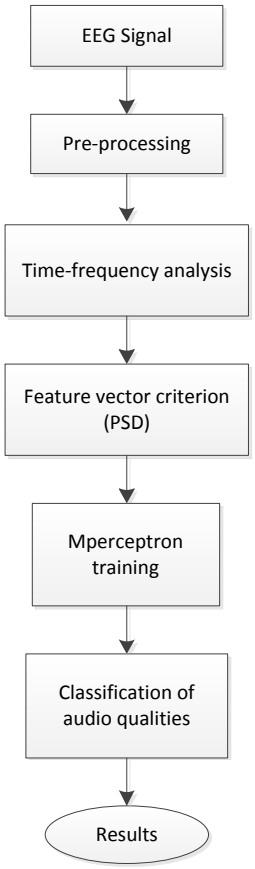

Fig. 3. Block diagram of EEG-based audio quality classifier.

3 implement the feature extraction process as previously described in Section 2 while the Mperceptron training process and the implementation of the resulting linear discriminant process are described below.

The Mperceptron tool described in [12] uses the perceptron-learning rule to train a linear machine (multi-class linear classifier) on the extracted feature vectors. The 'perceptron criterion' error function used by this tool minimizes the number of misclassifications, and the approach has the advantage that it will find a separable partitioning of the training set data if one exists. While we also evaluated the effectiveness of *k*-Nearest Neighbor and neural network-based classifiers over the course of this study, we found that the Mperceptron-based approach performed the best. An iterative stochastic gradient descent algorithm is applied to the error function, resulting in the optimal weight matrix **W**. The multi-class problem is then transformed into a single-class one using the Kessler's construction. The final result of Multi-class perceptron training algorithm is a model **W** and a bias vector *b*.

In order to fairly evaluate classification performance, we partition our data set into separate training and testing sets. We do this as follows: every degraded and reference audio sequence is presented to each subject four times in the passive listening state. We use the EEG data generated by three of these presentations as training data while the 4th is used as testing data to generate the results presented in Section 5.

Actual classification is performed using the Linclass tool which is also part of the Statistical Pattern Recognition Toolbox described in [12]. This tool accepts test data vector *i* and outputs class decision *Y(i)* according to

$$Y(i) = \arg\max_{y}(\mathbf{W}'(:,y)\mathbf{X}(:,i) + b(y))) \qquad (1)$$

based on the model that resulted from the application of the Mperceptron algorithm to each specific training set where $\mathbf{X}(:,i)$ is the 128x1 single-frequency spatial test vector for time *i* (e.g., a single column from Fig. 2). Because we know here that only four different levels of audio quality were used in these tests, we constrain our classifier to have only four possible output classes, each corresponding to a quality level: low (L), medium (M), high (H), and full (F) (i.e., undegraded reference audio). While the real-world problems we are trying to address lack such *a priori* constraints, we felt that it was important to limit the problem as much as possible in this early work in order to maximize our chances of finding potential quality indicators buried within the EEG data sets.

Finally, note that a separate linear classifier (e.g., Equation (1)) is designed for and applied to each 128x1 frequency vector in Figure 2 for each time interval. Thus, we get 12 classification decisions for each interval. To make a final decision, the *mode* (or median) of these 12 decisions is selected.

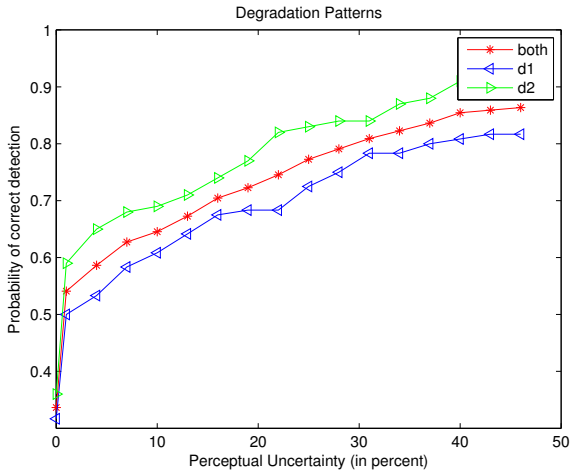

Fig. 4. Probability of correct detection versus the perceptual uncertainty in the quality rating for the two different distortion patterns.

5 RESULTS

The results achieved using the Mperceptron-based quality classification approach for a set of test cases are summarized

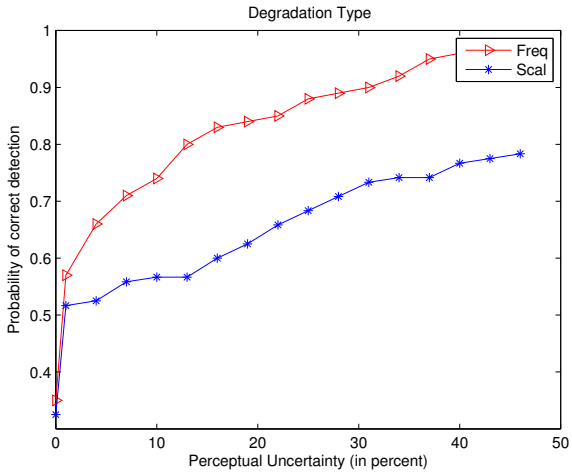

Fig. 5. Probability of correct detection versus the perceptual uncertainty in the quality rating for distortion types Scalar Quantization and Frequency Truncation.

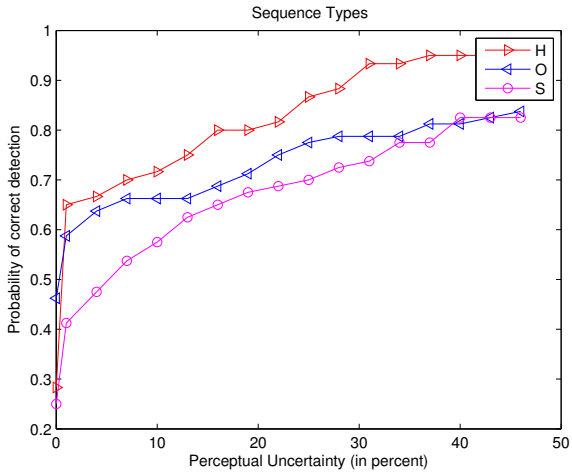

Fig. 6. Probability of correct detection versus the perceptual uncertainty in the quality rating for the three different base audio sequences.

in Figures 4, 5, and 6. In these figures, we plot the probability that the classifier selects the correct quality class relative the 'perceptual uncertainty' in the quality ratings. This 'perceptual uncertainty' allows us to accept an incorrect decision made by the classifier as a correct decision if the perceptual ratings giving by the individual participants to the same segments of the same base audio sequences are sufficiently close together. The baseline subjective quality ratings were collected twice from each subject for every degraded and reference audio sequence using the sliderbar as discussed in Section 3. This percentage difference is calculated in the following manner. For each 3 second interval over which the audio quality was held constant (technically, the middle 3 seconds of a 5 second interval of constant quality) and each specific degraded temporal segment of a specific base audio sequence, we calculate that participant's average quality rating. We then difference this from the quality rating given by the same subject to the same temporal segment of the same audio sequence which was presented at full (reference) quality. We do this to compensate for the fact that the participant may have heard things that were part of the reference audio sequence which he or she thought were degradations. This reference-normalization of the ratings data is applied to all of the quality classes, each subject, and each segment of audio separately. For example, for the segment from seconds 6 to 9 of the Pat Benatar sequence (labeled 'S' in Figure 6), we might have reference-normalized average quality ratings of 51 for 'low', 55 for 'medium', and 75 for 'high'. If the true class quality for a given trial is 'low' but our classifier selects 'medium', then we will still declare a correct classification if a perceptual uncertainty of 5% is allowed. This corresponds to the case in Figure 6 where the value on the x-axis is 5. On the other hand, if in the same example, the classifier selects 'medium' but the actual quality was 'high', we would need to increase the perceptual uncertainty to 20% for this to be considered a correct classification decision.

In Figure 4 we plot correct classification versus perceptual uncertainty for the two different distortion patterns, *d1* and *d2*, considered separately as well as for the combination of the two. Data from all of the test subjects, for all of the individual trials, over both distortion types, and all three base sequences is used here with the classifier being designed with and tested on orthogonal data sets. From the plot, we note that our classifier appears to be approximately 10% more accurate for distortion pattern *d2* than for *d1* across the uncertainty levels—this would appear to be a significant improvement. The distortion patterns for *d1* and *d2*, respectively, are (full, low, high, medium, low, full) and (full, medium, low, high, medium, full) where each quality level is held constant for 5 seconds and 1 second transitions are used. It is not clear to us at this time why there would be such a large difference in detection probability for these two distortion patterns, and it will probably be necessary to perform additional trials in which every time interval of audio is degraded to all three levels of impairment in order to fully understand this phenomenon.

Figure 5 performs the same comparison over the entire data set (separating again the test and training data) but compares results for the two types of distortion used. From the plot, it is very clear that our quality classifier is far more accurate for frequency truncation distortion than it is for scalar quantization. This is not particularly surprising. One notes that our classification approach was inspired by the observations and neuroscientific discussion presented in Section 2 which specifically refer to time-space-frequency energy distributions seen in the frequency truncation case. Similar time-space-frequency plots for the scalar quantization case have not been as easy to analyze. Consequently, it is quite possible that the feature vector set proposed here is not well suited to

the scenario in which audio is degraded by scalar quantization noise. Note that in the degraded audio presented here, scalar quantization was performed in the MDCT-domain uniformly across all frequencies (frequency truncation was also performed in this domain).

The plot of Figure 6 compares the results for the three different base audio sequences over the complete data set but with the test and training sets partitioned as before. We note that the approach appears to be most effective for the Michelle Branch sequence (labeled 'H') and less effective, at least for low levels of perceptual uncertainty, for the Pat Benatar sequence (labeled 'S'). For the third sequence Ode to Joy (labeled 'O'), our classifier appears to do well for low levels of perceptual uncertainty but does not improve very quickly as the allowable level of uncertainty is increased. These differences are probably related to the relative amounts of higher frequencies contained in the three base audio sequences (given that most of our classification accuracy appears to be derived from the effectiveness of the approach with respect to the frequency truncation): sequence 'H' appears to have a lot of relatively high power, high frequency harmonics while sequence 'S' appears to have very limited high frequency content. In sequence 'O', the impairment introduced by frequency truncation is far more subtle than for the other two sequences; it is therefore possible that subjects having more familiarity with classical music may be able to more readily identify (and more highly penalize) the impaired audio than the other subjects in the pool.

6 CONCLUSION

In this paper, we have considered the application of EEG for the purpose of evaluating the human perception of audio quality, considering in particular scalar quantization and frequency band truncation. Our work thus far has been highly preliminary and clearly incomplete—the proposed approach appears to be promising for frequency truncation distortion but less so for scalar quantization. Furthermore, the variations in its effectiveness for different audio sequences and different distortion patterns need to be more fully analyzed, and data for more test subjects needs to be considered. Finally, we believe that the feature vector set can be pruned and the effectiveness of the resulting classifier consequently improved by determining exactly which electrodes are and are not contributing information that is relevant to the audio quality classification problem. We feel that principle component analysis (PCA) methods applied in the spatial domain may be helpful in accomplishing this task.

7 References

[1] H. Fletcher, "Auditory patterns," *Rev. Modern Physics*, vol. 12, pp. 47–65, January 1940.

[2] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, February 1988.

[3] Timothy D. Griffiths, Sukhbinder Kumar, William Sedley, Kirill V. Nourski, Hiroto Kawasaki, Hiroyuki Oya, Roy D. Patterson, John F. Brugge, and Matthew A. Howard, "Direct recordings of pitch responses from human auditory cortex," *Current Biology*, vol. 20, no. 12, pp. 1128 – 1132, 2010.

[4] Siyi Deng and Ramesh Srinivasan, "Semantic and acoustic analysis of speech by functional networks with distinct time scales," *Brain Research*, vol. 1346, pp. 132 – 144, 2010.

[5] Boris Gourevitch, J. R. LeBouquin, G. Faucon, and Catherine Liegeous-Chauvel, "Temporal envelope processing in the human auditory cortex: Response and interconnections of auditory cortical areas," *Hearing Research*, vol. 237, no. 1-2, pp. 1 – 18, 2008.

[6] A. Roye, E. Schrger, T. Jacobsen, and T. Gruber, "Is my mobile ringing? Evidence for rapid processing of a personally significant sound in humans," *Journal of Neuroscience*, vol. 30(21), pp. 7310–7313, May 2010.

[7] Christophe Micheyl, Robert P. Carlyon, Alexander Gutschalk, Jennifer R. Melcher, Andrew J. Oxenham, Josef P. Rauschecker, Biao Tian, and E. Courtenay Wilson, "The role of auditory cortex in the formation of auditory streams," *Hearing Research*, vol. 229, no. 1-2, pp. 116 – 131, 2007.

[8] Shihab A Shamma and Christophe Micheyl, "Behind the scenes of auditory perception," *Current Opinion in Neurobiology*, vol. 20, no. 3, pp. 361 – 366, 2010.

[9] H. Hayashi, H. Shirai, M. Kameda, S. Kunifuji, and M. Miyahara, "Assessment of extra high quality images using both EEG and assessment words on high order sensations," in *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, 2000, vol. 2, pp. 1289 –1294 vol.2.

[10] A. Schnitzler and J. Gross, "Normal and pathological oscillatory communication in the brain," *Nat Rev Neurosci*, vol. 6, pp. 285–296, April 2005.

[11] Paul L. Nunez and Ramesh Srinivasan, *Electric Fields of the Brain*, Oxford University Press, New York, NY, 2006.

[12] V. Franc and V. Hlavac, "Statistical pattern recognition toolbox for matlab users guide," *Research report, Czech Technical University*, pp. 14–15, june 2004.