Feature extraction for automatic interpretation of EEGs has been extensively studied. A number of commercial approaches use exotic feature sets such as wavelets or nonlinear statistical measures such as fractal dimension. These choices of features were the results of evaluations and optimizations conducted on small research databases often collected under very controlled conditions. These approaches have not been extensively evaluated on big data or clinical applications using state of the art machine learning technology. Therefore, in this study, we compare performance of a number of standard feature extraction techniques on the publicly available TUH EEG Corpus using a state of the art classification system.

The TUH EEG Corpus is the largest publicly available corpus of EEG data. It comprises over 28,000 sessions collected from 2002-2015 at Temple University Hospital. It is entirely composed of clinical data, which means the data is representative of all the problems typically encountered in clinical settings, such as patient movement, artifacts due to eye blinks, talking, etc. Such data poses a much different challenge for machine learning systems since rejection of background noise becomes a critical issue.

The classification system used for this study, known as AutoEEG™, automatically recognizes specific events in the EEG data and generates annotations. AutoEEG™ is based on a hidden Markov model (HMM) approach to modeling the temporal evolution of the spectrum. A maximum likelihood (ML) approach is used to train standard three-state HMMs consisting of 8 Gaussian mixtures per state and diagonal covariance matrices.

The system detects three events of clinical interest (PLED, GPED and SPSW) and three events used to model background noise (ARTF, EYEM and BCKG). The current system uses an enhanced feature extraction approach based on Mel Frequency Cepstral Coefficients (MFCC’s) together with differential energy, first and second derivatives. In this study, we evaluated 15 features, shown to the right, by augmenting the standard feature vector with one additional feature. These were evaluated on a subset of the TUH EEG Corpus designed to give rapid turnaround on experiments, yet correlate well with results on the full dataset. This particular set of 15 features was chosen based on analysis of historically significant publications in the field.

The MFCC approach has been in use for speech recognition applications for several decades and is known to provide a robust characterization of the temporal and spectral properties of the signal. In general, our findings indicate that any of these features individually influence performance very little. This contradicts findings previously published, but was not unexpected. Clinical data is extremely challenging and quite different from most published EEG corpora. Maximum Fractal Length (MFL) provided the greatest reduction in error rate, though the improvements were not statistically significant.

In related work, we demonstrate that wavelets, which are often proposed as an alternative to MFCCs, also provide no gain in performance. Though literature suggests that EEG signals can be viewed as chaotic time series with significant amounts of nonlinealities, the features we investigated, which are designed to characterize such properties, add little value to our standard feature extraction approaches. Future research will be focused on better time-frequency representations of the signal based on correlation and coherence.

1. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U01HG008468.
2. This research was also supported by the Brazil Scientific Mobility Program (BSMP) and the Institute of International Education (IIE).
Abstract

- The emergence of big data and deep learning is enabling the ability to automatically learn how to interpret EEGs from a big data source.
- The AutoEEG™ is a system that automatically annotates specific events in the EEG data and generates annotations.
- The system detects three events of clinical interest (PLEDs, GPEDs and SPSWs) and three events that model background noise (BCKG, ARTF and EYEM).
- The current system uses an enhanced feature extraction approach based on Mel Frequency Cepstral Coefficients (MFCCs), together with differential energy, first and second derivatives.
- This study evaluated a range of features by differential energy, first and second derivatives.
- An additional feature typically increases performance over the baseline MFCC approach.

Introduction

- Clinical EEG (cEEG) measures the electrical activity in the brain and is used to diagnose patients suffering from neurological disorders such as epilepsy and stroke.
- AutoEEG™ uses a speech recognition approach for classifying 1 second epochs of an EEG signal into one of over 20 events: generalized tonic-clonic seizures, partial seizures, sleep spindles, alpha (ARTF), eye movement (EYEM), and background activity (BCKG).
- A frame duration of 0.1 seconds is used to model 1 second epochs of the signal.
- An ML approach is used for classification.

Mel Frequency Cepstral Coefficients

- Maximum likelihood (ML) approach is used to train a hidden Markov model (HMM) using the temporal resolution of the data.
- A frame duration of 0.1 seconds is used to model 1 second epochs of the signal.
- An ML approach is used for classification.

Feature Extraction Methods

- Feature extraction reduces the sampled data sequence to a sequence of features that contain the most relevant information for classification.
- Derivatives accentuate spectral dynamics.
- Static coefficients can be greatly enhanced by adding time derivatives to the basic static parameters.
- Delta features (acceleration) are used to the basic static parameters.
- Derivatives are then computed over each window basis over each channel of the EEG signals and added to the respective HTK file immediately after the MFCC's.
- The derivatives are then computed over each window (feature vector) matching a time of 50 frames.

Baseline Performance

- An error correction matrix for the HMM-based system (MFCCs)

Experimental Design

- A pilot study was conducted on a small data set of 12 EEG sessions for training and an independent test set of EEG sessions for validation.
- The data was sampled at 250 Hz and analyzed using a frame duration of 0.1 seconds and an analysis window duration of 0.2 seconds (50 samples).
- The data was sampled at 250 Hz and analyzed using a frame duration of 0.1 seconds and an analysis window duration of 0.2 seconds (50 samples).

Preliminary Results

- A Detection Error Tradeoff (DET) curve:

Summary

- The results presented here were obtained using a small pilot corpus that is designed to examine the potential of a new feature set.
- The additional feature typically increases computation time by 10%.
- The system was validated using an additional feature set.
- Accurate detection of the SPSW class is most important since it is the most important indicator of a potential neurological disorder.
- Additional features can be applied to data labeled as PLED or GPED.
- Combining the background noise classes into a single class gives this confusion matrix for the MFCCs:

<table>
<thead>
<tr>
<th></th>
<th>BCKG</th>
<th>ARTF</th>
<th>EYEM</th>
<th>GPED</th>
<th>PLED</th>
<th>SPSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCKG</td>
<td>26.7%</td>
<td>40.8%</td>
<td>29.0%</td>
<td>26.9%</td>
<td>23.5%</td>
<td>17.8%</td>
</tr>
<tr>
<td>ARTF</td>
<td>29.0%</td>
<td>19.8%</td>
<td>7.55%</td>
<td>17.6%</td>
<td>18.6%</td>
<td>17.6%</td>
</tr>
<tr>
<td>EYEM</td>
<td>32.4%</td>
<td>23.88%</td>
<td>16.43%</td>
<td>19.2%</td>
<td>19.9%</td>
<td>19.2%</td>
</tr>
<tr>
<td>GPED</td>
<td>26.3%</td>
<td>19.6%</td>
<td>19.9%</td>
<td>23.5%</td>
<td>29.0%</td>
<td>29.0%</td>
</tr>
<tr>
<td>PLED</td>
<td>29.0%</td>
<td>29.0%</td>
<td>29.0%</td>
<td>29.0%</td>
<td>29.0%</td>
<td>29.0%</td>
</tr>
<tr>
<td>SPSW</td>
<td>26.3%</td>
<td>26.3%</td>
<td>26.3%</td>
<td>26.3%</td>
<td>26.3%</td>
<td>26.3%</td>
</tr>
</tbody>
</table>

Acknowledgements

- This research was also supported by the Brazil Scientific Mobility Program (BSMP) and the Institute of International Education (IIE).