Stress Level Detection Using Double-Layer Subband Filter

Tin Lay Nwe, Xu Qianli, Guan Cuntai and Bin Ma

Institute for Infocomm Research, Singapore
tlnma,qxu,ctguan,mabin@i2r.a-star.edu.sg

Abstract

Stress level detection is important for human error prevention and health care services. Speech based stress level detection is the most effective as speech data can be obtained in non-intrusive and inexpensive ways. In this paper, we explore the features that use Double-Layered Subband (DLS) filter for detecting stress level from speech. Spectral Centroid Frequency (SCF) and Spectral Centroid Amplitude (SCA) are acoustic features that complement each other if these two are fused together using appropriate weighting coefficient. We extract SCA using DLS filters. And, we present how DLS filter integrates SCF information to SCA feature without actually computing SCF feature parameters. We investigate the effectiveness of proposed approach over combining SCA and SCF using weighting coefficient. We build user independent stress level detection system. Stress is detected using the scale of 0 to 1 (0' being index for no stress and 1' being index the highest stress level). The experiments show that the proposed system is able to detect the level of stress from speech with reasonably high accuracy.

Index Terms: stress level detection, spectral centroid amplitude, spectral centroid frequency, human stress, subband filters

1. Introduction

Stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activities, and deterioration of performance [1]. It is induced by physical or mental stressors [2]. Examples of stressors are high noise environment, physical taskload, mental taskload, etc... Stress level classification is useful in the applications such as call centers, vehicle driving, health care, etc... In call center applications, the system is to detect customer’s stress level (e.g., level of unhappiness). For the applications such as vehicle driving, the system can monitor variations of driver’s stress level which may impact the driver’s capability to control the vehicle [3].

A speech based stress level detection system includes two components: the front-end feature extraction and back-end stress level detector. Back-end stress level detector includes training and scoring against a stress model. Stress models are trained using hidden Markov models (HMMs) [3], [4] Neural Networks (NNs) [3], [5], Support Vector Machines (SVMs) [3], [5], [6], [7] and Gaussian mixture models (GMMs) [8]. In our experiments, we use GMM to train stress model.

For front-end feature extraction component, several features have been reported in literature for stress classification. The feature parameters generally considered include intensity, pitch, duration, vocal tract spectrum, glottal source and vocal tract articulatory profiles [2], [9]. In [8], delta cepstrum, acceleration (delta-delta) and shifted delta cepstra features are considered to classify three stress levels (low, medium and high). In [10], a method for indicating stress in speech is presented by using a weighted vibrato content of the speech. Although a number of features were investigated for stress classification systems, these systems are still limited in terms of performance for real life applications. There are systems used in real life applications. However, these systems need human operators to decide the final score for stress level. In [10], a trained person compares the measured vibrato contents with a vibrato content of known stress free speech to decide the stress level. In [11], a trained operator visualizes the parameters such as fundamental frequency, amplitude and slope to decide the final score.

Hence, the automatic stress detection system still has limitations on the practical applications. Front-end acoustic features have major contribution on system performance as the acoustic features need to characterize stress perfectly.

When people are under stress, muscle tension of vocal cords and vocal tract increases. This may directly or indirectly, adversely affect the quality of speech. Hence, features related to spectral energy distributions are good indicators of presence of stress [12]. In [12], SCA and SCF features that reflect spectral energy distributions, are used to classify three stress levels. GMM based stress level classifiers are build using the two features separately. As SCA and SCF features complement each other, the log-likelihood scores generated by the two classifiers are fused using weighting coefficients. This fusion needs proper selection of weighting coefficients which have major impact on classification accuracy. In this paper, instead of computing two parameters: SCA and SCF, we propose a method to extract only one parameter that includes information on both SCA and SCF. We employ DLS filter to include both SCA and SCF in one feature. Our proposed method does not require weighting coefficients.

In literature, stress level detection is performed on a database including stress speech utterances recorded under taskloads with different difficulty levels. For example, if database recorded under taskload with four difficulty levels, the system is designed to detect four stress levels. Each difficulty level corresponds to a stress level [4], [12]. In fact, people respond to stress differently. Due to a number of factors, such as domain or interface expertise, age, mental or physical impediments, different people are affected in different ways when performing the same task [8]. Considering this variation, detecting stress level experienced by individuals is important for human error prevention and self health care. In this paper, we are interested in detecting the level of stress perceived given by the same physical taskload.

The rest of the paper is organized as follows. Section 2 describes the proposed acoustic feature formulation method. Section 3 describes the stress speech database. Section 4 presents stress level detection method. Section 5 depicts experimental results, and Section 6 concludes the paper.
2. Acoustic parameters

Under a taskload, some form of pressure is applied to the speaker, which may result in a perturbation of the speech production process and hence of the acoustic signal. When a person is under a taskload, physiological changes occur in respiratory system. These include increase in respiration rate, irregular breathing, increase in muscle tension of vocal cords, etc. [2]. The increased muscle tension of the vocal cords and vocal tract makes changes to speech characteristics. Consequently, features that characterizing the spectral energy spread have been found to reflect the stress experienced by the speaker [3], [12]. Under a taskload, spectral intensity of speech concentrate more on high frequency regions compared to that of neutral speech [3]. An increase in perceived stress level is reflected by an increase in spectral energy spread and spectral center of gravity [3]. Hence, spectral energy spread and spectral center of gravity are good indicators to present the perceived stress level. The information on spectral energy spread can be obtained using Spectral Centroid Amplitude (SCA) features. As for spectral center of gravity, Spectral Centroid Frequency (SCF) feature can be used. Hence, we formulate acoustic features to include both information: SCA and SCF. The advantage of proposed method is that extracted features include both SCA and SCF information without weighting coefficients. We compute only SCA using Double-Layered Subband (DLS) filters. DLS filter is capable of integrating SCF information to SCA without actually computing SCF. We describe DLS filters in the following section.

2.1. Double-Layered Subband (DLS) Filters

To include SCA and SCF information in acoustic feature formulation, we employ a Double-Layered Subband (DLS) filter [13], [14], [15]. The structure of filter is shown in Fig. 1.

The filter has two layers of subbands. The first layer has 14 overlapped trapezoidal filters. The second layer has 5 nonoverlapped rectangular filters of equal bandwidths for each trapezoidal subband of the first layer. Trapezoidal filters are tapered on overlapped portions. There are a total of 70 (14 X 5) subbands in the second layer. The first layer of the filter bank was designed to mimic the varying auditory resolving power of the human ear for various frequencies, so divides the speech signal into 14 frequency bands that match the critical perceptual bands of the human ear. Subbands in the first layer are arranged using logarithmic scale. The filters in first layers spans 130Hz-20kHz. Although the useful bandwidth of a band-limited speech signal for speech recognition is between 200 Hz and 3.2 kHz [16], selecting this range for stress level detection may discard some useful information. Therefore, we select larger frequency range. In the following section, we present SCA feature computation using DLS filters and how DLS filters integrate SCF information into SCA.

We also use Single-Layered Subband (SLS) filter to extract individual SCA and SCF features to observe the effectiveness of the proposed features over fusing SCA and SCF in score level.

2.2. Features Based on Spectral Centroid

Spectral Centroid Amplitude (SCA) is the weighted average magnitude spectrum in the subband. And, the Spectral Centroid Frequency (SCF) is an estimate of the ‘centre of gravity’ within the spectrum [17]. We extract SCA features using DLS filters as follows. A speech signal is divided into frames of 20ms with 10ms overlapping. Each frame is multiplied by a Hamming window to minimize signal discontinuities at the end of each frame. Then, discrete spectrum, $S[f]$, of each frame is determined. $S[f]$, is divided into $M$ subbands with frequency responses, $W_m[f]$, where $m \in [1,M]$. Let $l_m$ and $u_m$ are the lowest and the highest frequencies of $m$th subband. SCA feature is calculated from $S[f]$ for $m$th subband using equation (1) [12].

$$ SCA_m = \frac{\sum_{f=l_m}^{u_m} f |W_m[f]|S[f]|}{\sum_{f=l_m}^{u_m} |S[f]|} \quad (1) $$

For each frame, a total of 12 SCA$_{DLS}$ parameters that use DLS filters are computed by taking Discrete Cosine Transform (DCT) [18] on log of the vector obtained by concatenating all $SCA_m$ of each subband.

To observe the effectiveness of the proposed SCA$_{DLS}$ features over fusing individual SCA and SCF features in score level, we extract SCA and SCF features using only the first layer of DLS filters as subband which is referred to as Single-Layer Subband (SLS) filters. SCA and SCF features extracted using SLS filter are referred to as SCA$_{SLS}$ and SCF$_{SLS}$ respectively. To extract SCA$_{SLS}$ features, we replace DLS filters with SLS filters in the SCA$_{DLS}$ feature extraction process. Equation (2) is to compute SCF feature from $S[f]$ for $m$th subband of SLS filter [12].

$$ SCF_m = \frac{\sum_{f=l_m}^{u_m} f |W_m[f]|S[f]|}{\sum_{f=l_m}^{u_m} |W_m[f]|S[f]|} \quad (2) $$

For each frame, a total of 14 SCF$_{SLS}$ parameters are obtained by concatenating all $SCF_m$ of each subband. DCT is not applied as SCF is frequency bused features.
The subjects stress levels were derived from the STAI-Y1 scores. These stress level parameters will be used as reference in our experiments. Cronbach's alpha for the three experiment stages are 0.923 (baseline), 0.899 (task load), and 0.904 (treatment), showing high internal consistency of the questionnaire items. Next, the normalized stress level (to be used as reference), \( s_{ij} \), for subject \( i \) is computed as,

\[
s_{ij} = \frac{S(s) - \min(S)}{\max(S) - \min(S)}
\]

where \( \max(S) \) and \( \min(S) \) refer to the largest and smallest STAI scores of all subjects in three stages. Thus, the stress indices are within the range of \([0, 1]\).

### 3. Corpus

We perform experiments to induce stress on subjects by carrying out the following activities: 1) initial rest (5min), 2) task load (5 min) and 3) recovery rest (15min). A total of 3 recordings are made for each subject after each activity. In initial rest (5 minutes), a subject is instructed to relax in a comfortable chair with eyes closed. Speech recorded after this initial rest is referred to as 'baseline'. Then, the subject performs a squat-exercise at a pace that is considered challenging to the subject. Speech recorded after this activity is referred to as 'taskload'. Finally, the subject rests in a comfortable chair for 15 minutes to recover from the possible stress due to the exercise. Speech recorded after this is referred to as 'recovery'. We recruited 24 subjects (23 males, 1 female, age 19-48), who were healthy adults without symptoms, signs, or history of respiratory disorders. Total duration of the corpus is 1 hour.

In the above process, the subjects’ speech was recorded and their self-reported stress was collected using the State-Trait Anxiety Inventory (STAI) for adults [19]. Specifically, upon completion of each activity, a subject read a 90-word essay in English. The speech was recorded with a sampling rate of 44.1kHz, mono channel, and 16-bit resolution. After reading the passage, a subject reported his/her subjective stress using the STAI form. This form provides a definitive instrument for measuring anxiety in adults. Since we were interested in the temporal status of the subjects, the questionnaire included 20 items for measuring the state anxiety.

The stress level is measured based on the distance between the speech and Stress Model (\( \lambda_{SM} \)). If a speech sample is very close to \( \lambda_{SM} \) the subject's perceived stress level is taken as 'high'. The reverse is true for the 'low' stress level. We use GMM loglikelihood scores as a 'distance' measure. In stead of using absolute distance, we employ Relative Distance (RD) in which distance is measured from a reference point. The reference point is defined using Baseline Model(\( \lambda_{BM} \)) which is trained using baseline samples from several speakers. We build Stress Level Indicator Table (SLIT) that presents RDs vs. stress levels. Given an RD value of an unknown speech sample, we use SLIT to obtain the stress level. The following is the process to build SLIT.

Let \( X_S = [X_1^s, X_2^s, ..., X_N^s] \) and \( X_B = [X_1^b, X_2^b, ..., X_N^b] \) to be \( N \) numbers of stress and baseline samples. Then, \( RD_i^s \) is computed from \( X_i^s \) as follows.

\[
RD_i^s = P(\lambda_{SM}|X_i^s) - P(\lambda_{BM}|X_i^s)
\]

We compute RDs for both \( X_S \) and \( X_B \) using (4). SLIT has 101 stress levels: '0' being no stress and '1' being the highest stress level with the step size of 0.01. A total of 101 stress levels are divided into three regions: low, medium and high. The first 33 levels: 0 ∼ 0.32, are for low stress. And, the second 33 levels: 0.33 ∼ 0.65, are for medium stress. Finally, the last 33 levels: 0.66 ∼ 1, are for high stress. Figure 4 shows SLIT. In Figure 4, \( RD_{0.0}, RD_{0.32}, RD_{0.66}, RD_1 \) are RD values at the boundaries of 3 regions. These RD values are defined as follows.
\[ RD_0 = \min[RD_1^0, RD_2^0, ..., RD_N^0] \] (5)

\[ RD_{0.32} = \max[RD_1^0, RD_2^0, ..., RD_N^0] \] (6)

\[ RD_{0.66} = \min[RD_1^0, RD_2^0, ..., RD_N^0] \] (7)

\[ RD_1 = \max[RD_1^1, RD_2^1, ..., RD_N^1] \] (8)

Once we define RD values for the boundaries, we obtain RD values for each of the 101 stress levels. We assume linear scale between RDs and stress levels within a region. And, we compute RD value for each level by increasing a linear step, \((\Delta RD)\) starting from the minimum boundary (example, \(RD_0\) for low stress region). The following is the example to compute \(RD_{0.01}\) which is RD value of the stress level 0.01 which is one level higher than ‘0’ stress level.

\[ RD_{0.01} = RD_0 + \Delta RD \] (9)

Using (9), we compute RD values for all the 101 stress levels and tabulated in SLIT. Once we have SLIT, stress level detection of an unknown speech utterance is performed by computing RD using equation (4) and using SLIT to obtain stress level.

5. Experiments and Results

Experiments are conducted to evaluate the effectiveness of the proposed \(SCA_{DLS}\) feature. We also compute \(SCA_{SLS}\), \(SCFSLS\) and traditional Mel Frequency Cepstral Coefficient (MFCC) features for comparison with proposed feature. We use the Gaussian mixture models (GMMs) with 8 mixtures to train the models in our experiments. \(SCA_{SLS}\) and \(SCFSLS\) are combined at score level by using score weighting coefficient.

We divide corpus into train, development and test sets. Training data contains 7 speakers. Development data has 5 speakers. And, test set includes 12 speakers. As there are no speaker overlaps among training, development and test sets, our system is speaker independent. We build SLIT using training and development data as mentioned in Section 4. The models: \(\lambda_{SM}\) and \(\lambda_{BM}\) are trained using training data. Then, development data is used to compute RD values for 101 stress levels to build SLIT. Once we have SLIT, stress level detection is performed for all the speech samples of test set as mentioned in Section 4. We compute stress level detection error which is the difference between normalized stress indices, \(s_{ij}\) derived from STAI-Y1 scores (Section 3), and detected stress level obtained from SLIT. Mean stress level detection error rates are presented in Table 1.

Table 1: Mean Stress Level detection Error Rates for 3 feature sets. (BL=Baseline, TL=Taskload, RC=Recovery)

<table>
<thead>
<tr>
<th>Features</th>
<th>BL</th>
<th>TL</th>
<th>RC</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.25</td>
<td>0.29</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>(SCA_{SLS} + SCFSLS)</td>
<td>0.22</td>
<td>0.25</td>
<td>0.21</td>
<td>0.23</td>
</tr>
<tr>
<td>(SCADLS)</td>
<td>0.21</td>
<td>0.17</td>
<td>0.18</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Figure 5 shows detected mean stress levels for test set by all three feature sets. All features seem perform well. However, differences of stress levels between 1) baseline and taskload, as well as 2) recovery and taskload are significant when using \(SCA_{DLS}\) features. Hence, the results show that \(SCA_{DLS}\) feature outperform all other features. And, Figure 6 shows detected stress levels for all the speakers of test set using \(SCA_{DLS}\) feature. Stress level detected is the highest for taskload among all three types: baseline, taskload and recovery. This trend is the same for all the speakers except speaker 1. It can be observed that given the same taskload, stress gained by individual speaker is different.

Table 2: Classification accuracies (%) between baseline and taskload.

<table>
<thead>
<tr>
<th>Features</th>
<th>Baseline</th>
<th>Taskload</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>100</td>
<td>16.67</td>
<td>58.33</td>
</tr>
<tr>
<td>(SCA_{SLS} + SCFSLS)</td>
<td>75</td>
<td>83.33</td>
<td>79.17</td>
</tr>
<tr>
<td>(SCADLS)</td>
<td>75</td>
<td>91.67</td>
<td>83.33</td>
</tr>
</tbody>
</table>

To further confirm the effectiveness of \(SCA_{DLS}\) feature over other features, we perform stress classification experiments in which classification is done between baseline and taskload. We do not include recovery samples in this experiment. The average classification accuracy of test set is presented in Table 2. The results show that \(SCA_{DLS}\) feature performs the best among all the three features.

6. Conclusions

We have presented an approach to detect perceived stress level of different speakers in which stress is induced by same taskload. In this paper, we propose a feature formulation method to integrate Spectral Centroid Frequency(SCF) to Spectral Centroid Amplitude(SCA) to using Double-Layer Subband (DLS) filter. DLS filter integrate SCF information into SCA features without actually computing SCF values. This formulation has three advantages 1) this formulation does not need weighting coefficient. 2) computation expense is reduced to half as the formulation needs to build only one system using SCA feature (the system using SCF feature is not necessary), and 3) it gives better performance over the system fusing SCA and SCF together using weighting coefficient.
7. References


