Analysis and Detection of Speech under Sleep Deprivation

Tin Lay Nwe, Haizhou Li and Minghui Dong

Institute for Infocomm Research, Republic of Singapore
tlnma@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg, mhdong@i2r.a-star.edu.sg

Abstract

Stress has effect on speech characteristics and can influence the quality of speech. In this paper, we study the effect of Sleep-Deprivation (SD) on speech characteristics and classify Normal Speech (NS) and Sleep Deprived Speech (SDS). One of the indicators of sleep deprivation is 'flattened voice'. We examine pitch and harmonic locations to analyse flatness of voice. To investigate, we compute the spectral coefficients that can capture the variations of pitch and harmonic patterns. These are derived using Two-Layer Cascaded-Subband Filter spread according to the pitch and harmonic frequency scale. Hidden Markov Model (HMM) is employed for statistical modeling. We use DCIEM map task corpus to conduct experiments. The analysis results show that SDS has less variation of pitch and harmonic pattern than NS. In addition, we achieve the relatively high accuracy for classification of Normal Speech (NS) and Sleep Deprived Speech (SDS) using proposed spectral coefficients.

Index Terms: Sleep-deprivation, Stressed speech, Normal speech, Spectral coefficients

1. Introduction

Studies on stressed speech have described changes in characteristics and patterns of speech. Hence, stress is one of the several factors that affect performance of speech recognition. One of the applications of speech recognition is in military aircrafts. Military operations are often conducted under conditions of stress, induced by physical or mental stressors [1]. Examples of stressors are high noise environments, g-force, physical workload, mental workload, fear and emotion, confusion due to conflicting information, psychological tension, pain, sleep deprivation and several others. These affect speech characteristics and the performance of communication equipments (example, low-bit-rate secure voice systems) and systems with vocal interfaces (example, advanced cockpits, command and control systems). Hence, the effect of stress on speech characteristics is important to study.

In this paper, we consider the effect of a stressor, Sleep Deprivation (SD), on the changes of speech characteristics. Then, we classify the Normal Speech (NS) and Sleep Deprived Speech (SDS).

Much work has been done to study the background noise, Lombard effect and work-load stress on speech characteristics [2, 3, 4, 5]. In [2], vocal changes under severe environmental effects such as noise and increased cognitive workload are examined. This study shows that stress causes changes in pitch, speech amplitude and amplitude variability and decrease in word duration. In [3, 4, 5], vocal tract and speech parameters are analysed under stressful conditions, including Lombard effect. These studies summarize that formant location for vowels and formant amplitude increase under Lombard condition.

Little research has been done to study the effect of SD. In [6], speech parameters such as pitch and word durations are studied under 36 hours sleep deprivation and workload stress. This study shows that pitch and word duration vary significantly when SD progresses. In general, stress and fatigue have effects on speech characteristics such as pitch, speech amplitude and duration.

In this paper, we study the characteristics of speech under 60 hours sleep deprivation on several speech attributes of pitch, harmonic patterns and duration. To investigate, we implement Two-Layer Cascaded-Subband Filter (TLCSF) that can capture the variations of the pitch and harmonic patterns in their output spectral coefficients. We employ these spectral coefficients and duration information for analysis.

The rest of the paper is organized as follows. In Section II, we discuss in details the procedures for acoustic feature extraction. In Section III, we discuss the duration parameters. In Section IV, we present the database, experiments and results. Finally, we conclude our study in Section V.

2. Acoustic Parameters

The expressiveness of the acoustic parameters has a direct impact on analysing the effect of fatigue on speech. We compute short-time spectral coefficients that capture the variations of pitch and harmonic pattern to analyse speech parameters.

2.1. Effect of Sleep Deprivation (SD) on speech

Sleep deprivation has effect on the content and patterns of speech [7]. There are several indicators that represent SD. These include slurred speech, repetitive speech, slow or mumbled speech, flattened voice, decreased motivation [7, 8, 9] and little high frequency energy. Flattened voice results less variation of pitch and hence, harmonic pattern becomes stable. Furthermore, reduced muscle tone, resulting from the decreased motivation, may result in lowered pitch [8]. Repetition of words or slow speech results changes in word duration [10]. Hence, we implement the subband filters to capture variations of pitch and harmonic pattern as discussed in the following section.

2.2. Two-Layer Cascaded Subband Filters (TLCSF)

We propose two-layer cascaded subband filter shown in Figure 1 to capture the variations of pitch and harmonic pattern. The filter has two cascaded layers. The first layer has overlapped rectangular filters. For each filter in the first layer, there are 4 non-overlapped rectangular filters of equal bandwidth in the second layer. The first filter of first layer has a bandwidth spanned between 60Hz to 250Hz. This bandwidth covers pitch of male and female in general [11]. This filter is able to capture the variation in location of pitch. Details on how the filter captures pitch locations variations will be discussed in later paragraphs. Bandwidths of the following filters cover the harmonics of males and females. These filters capture harmonic locations variations.

The first layer has 7 subbands in total. Center frequencies and bandwidths of the 7 subbands in the first layer are given in Table 1.

Table 1. Center frequencies and bandwidths of the 7 subbands in the first layer

Subband No	Center Frequency (Hz)	Bandwidth (Hz)
1 (f0)	155	190
2 (2f0)	310	380
3 (3f0)	620	760
4 (4f0)	1240	1520
5 (5f0)	2480	3040
6 (6f0)	4960	6080
7 (7f0)	9920	12160

We have 7 subbands in the first layer. For each subband of first layer, we have 4 non-overlapped filters in the second layer. Hence, we have a total of 28 (7 X 4) subbands in the second layer. The range of our subband filters is between 60Hz ~ 16kHz

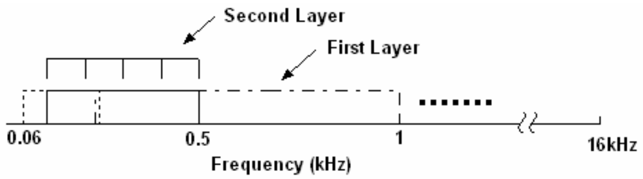

Figure 1 A bank of two-layer cascaded subband filters

Figure 2 (a) and (b) represents the variations of pitch and harmonic pattern between two consecutive frames for NS (before SD) and SDS (after SD) respectively. As can be seen in the Figure 2, pitch and harmonic pattern are varying between two consecutive frames of NS. However, for SDS, pitch and harmonic pattern become stable between two consecutive frames. Two-Layer Cascaded Subband Filter (TLCSF) captures this information as follows.

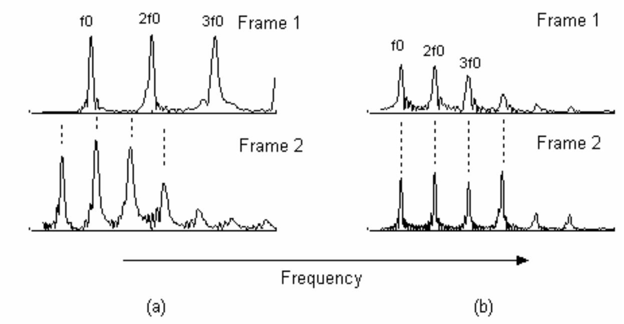

Figure 2 Variation of pitch and harmonic pattern between two consecutive frames of (a) Normal Speech (NS) and (b) Sleep Deprived Speech (SDS).

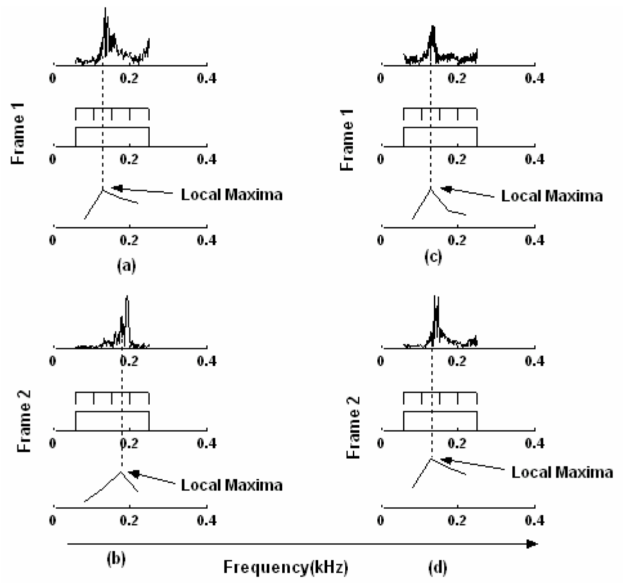

Figure 3 Pitch variation between two consecutive frames and TLCSF subband filtering: (a)Normal Speech (NS) for frame 1 (b) Normal Speech (NS) for frame 2 and (c) Sleep Deprived Speech (SDS) for frame 1 (d) Sleep Deprived Speech (SDS) for frame 2. For each figure, the upper panel shows the signal. The middle panel presents the frequency response of the TLCSF subband filters. The lower panel demonstrates the output of the TLCSF subband filters. The filters capture the position of pitch in the lower panel by locating the local maxima.

Location of pitch captured by TLCSF for NS and SDS are presented in Figures 3(a), (b) (c) and (d). In each figure, the signal in the upper panel is passed through the TLCSF filters shown in the middle panel. Then, output amplitudes of 4 subband filters are computed and shown in lower panel. As can be seen in the figures, the location of the local maximum in the lower panel is the location of pitch. Hence, TLCSF can captures variation of pitch position by tracking the local maxima of the subbands outputs in the second layer. Since TLCSF includes subbands for pitch and harmonic frequency ranges in the first layer, these subbands work together to capture pitch and harmonic pattern of the signal. Figures 3 (c) and (d) also shows that pitch is lowered and flat under fatigue and TLCSF is able to capture the flatness and location of pitch in its output.

The advantage of using two-layer cascaded subband filter over the triangular filters normally used in Mel Frequency Cepstral Coefficients (MFCC) computation, is that it tracks the local maxima of output subband energies. Thus, this filter is robust to amplitude variations which are normally associated with pitch.

2.3. Computation of subband coefficients

The 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, the audio frame is passed through a bank of cascaded subband filters and the spectral energy of each of 28 bands in the second layer is computed. In addition, we compute the derivative of spectral energy to integrate the speech rate into our speech parameters. Finally, a total of 56 Pitch and Harmonic frequency Spectral Coefficients (PHSC) are obtained for each audio frame.

3. Duration Parameters

In addition to the short-time spectral coefficients, we use duration information to analyse speech parameters. The need for microsleep and lapses increases when SD progresses [12]. As a result, the ability to perform the work as well as motivation is reduced [13]. This can be measured by Response-Time (RT) which is a measure of how fast Follower responds to the Giver's instruction. RT is the duration between ending point of the Giver's instruction and starting point of the Follower's response. Starting points of utterances for both Givers and Followers within the dialogues are provided in the database. Ending points of the utterances are obtained by using simple energy threshold method.

We use the statistical analysis method proposed by Elias [14] to examine the RT parameter distribution. This method compares the two parameter distributions provided these are Gaussians. This method uses the Elias Coefficient, *E*, to measure the degree of overlap between two Gaussian distributions. Mathematically, the Elias coefficient, *E*, is calculated as follows.

$$E = \int_{-\infty}^{+\infty} \left| p_1(x) - p_2(x) \right| dx \qquad (1)$$

where $p_1(x)$ and $p_2(x)$ are the probability densities associated with the two distributions. The difference between the two parameter distributions is considered the most significant if an Elias coefficient is 2 and the two parameter distributions are considered completely overlapped if an Elias coefficient is 0.

We assume that RT parameter distribution is Gaussian. The parameter distributions (normalized histograms) of RT between NS and SDS are presented in Figure 4 (a). The figure shows that the Response-Time of SDS is longer than that of NS. We will integrate this information into HMM model scores for better classification accuracy.

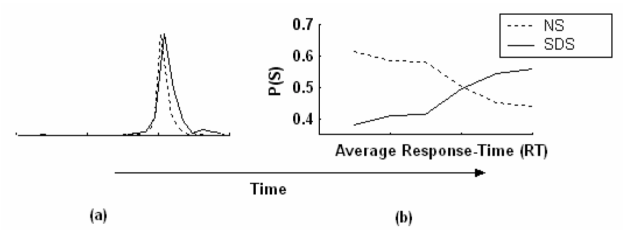

Figure 4 (a) Response-Time (RT) parameter distribution of NS and SDS (b) Probability, $P(S)$, of individual speech classes (NS and SDS) vs. average Response-Time (RT) over an individual dialogue.

4. Experiments and Results

The data used in this analysis is DCIEM Map Task Corpus [15] that includes 216 unscripted task-oriented dialogues spoken by 36 normal Canadian adults (34 males and 2 females) under 3 drugs conditions (placebo, d-amphetamine, modafinil). A total of 12 subjects is included in each drug group. A placebo is an inactive pill, liquid, or powder that has no treatment value. The lengths of the dialogues range from 3 mins to 13.5 mins. Each dialogue are recorded when speakers carry out route-communication task over 6 days which includes 60 hours sleepless period. A total of 72 dialogues are recorded for each drug group. Each speaker participated in 12 dialogues. Two speakers (instruction Giver and Follower) participated in each dialogue. Giver asks questions or gives instructions to the Follower to reach to the final destination. Details of the corpus can be found in [15].

We choose the speech data under 'placebo' drug for our experiments. We use the 6 dialogues before sleepless period as Normal Speech (NS) data and the 6 dialogues after 60 hour sleepless period as Sleep Deprived Speech (SDS) data in our experiments. Half of the data base is used as the training data and the system is tested on another half of the database.

Several experiments are conducted to evaluate the effectiveness of the proposed features. We use continuous density HMM with four states and two Gaussian mixtures per state for all HMM models in our experiments. Using training database, we train the two HMM models for Normal Speech (NS) and Sleep Deprived Speech (SDS). Then, the test sample is scored across the two models.

For classification using HMM, we usually formulate the maximum a posteriori (MAP) decision rule to find the most likely model which achieves the maximum posteriori probability $p(S \mid X)$ [16], i,e.,

$$\hat{S} = \arg\max_{S \in \Gamma} p(X \mid S).p(S) \qquad (2)$$

where X is the input feature vector, S is the speech classes (here, NS and SDS) and Γ is the set of all speech classes. The probability of observing X under the assumption that S is the underlying speech class for X, $p(X \mid S)$, is the score given by the HMM models. The probability of S , $p(S)$, is computed as follows. Using mean RT values of the dialogues from training data, we draw the $p(S)$ *vs. average RT* curve as shown in Figure 4(b). Then, we compute the $p(S)$ for individual dialogue of test data from the curve. We then, multiply $p(X \mid S)$ with $p(S)$ to find the best matched speech class.

Duration dialogues, Giver and Follower talks interchangeably. If their speech segments are longer than 2 seconds, the speech is segmented into 1 second segments. The classification decision is made on the test segments of one second or less than one second. We conduct the experiment to compare the system performance of three feature types, namely, PHSC, MFCC and LPCC. The results are given in Table 2.

Table 2. Classification accuracies of NS and SDS for three different features

Features	NS	SDS	Average
PHSC	93.4	79.6	86.5
MFCC	93.5	71.2	82.3
LPCC	93.5	77.4	85.4

Table 2 shows that PHSC feature, with an average classification rate of 86.5%, outperforms the tradition features, MFCC and LPCC. This result explains that PHSC feature can perform well to represent difference in speech characteristics between NS and SDS. In PHSC, two-layer cascaded subband filters captures information of pitch and harmonic variations of SDS and NS, and integrate this information into spectral coefficients. This is the advantage of PHSC feature over traditional MFCC.

5. Conclusions

We have presented a method to analyse speech under Sleep Deprivation (SD) based on the speech characteristics of pitch and harmonic pattern variations. Further, we differentiate the Normal Speech (NS) from Sleep Deprived Speech (SDS) based on the differences in pitch and harmonic patterns between NS and SDS. Our contribution to this analysis method is the use of Two-Layer Cascaded Subband Filters (TLCSF) to capture the variation of pitch and harmonic pattern. The analysis results show that proposed spectral coefficients can represent the difference between NS and SDS. Furthermore, the results show that NS and SDS speech classes can be classified using these features with relatively high accuracy.

6. References

[1] Steeneken, H.J.M. and Hansen, J.H.L, “Speech under stress conditions: overview of the effect on speech production and on system performance,” IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 2079-2082, Mar 1999.

[2] Pisoni, D. B., “Speech perception and production in severe environments”. AAMRL-SR-90-507, 1990.

[3] Hansen, J.H.L., “Analysis and compensation of stressed and noisy speech with application to robust automatic recognition”, Ph.D. Thesis, Georgia Inst. of Tech., Atlanta, GA, 1988.

[4] Hansen, J.H.L., “Evaluation of acoustic correlates of speech under stress for robust speech recognition”, IEEE Proc. of the Fifteenth Annual Northeast Bioengineering Conference, (invited paper), pp. 31-32, Boston, Mass., March 1989.

[5] Hansen, J.H.L. and Clements, M., “Evaluation of speech under stress and emotional conditions,” Proc. of the Acoustical Society of America, 114th Meeting, H15, Miami, Florida, Nov. 1987.

[6] Whitmore, J. and Fisher, S., “Speech during sustained operations”, Speech Communication, 20, 55-70, 1996

[7] Harrison, Y. and Horne, JA., “Sleep deprivation affects speech.” 1997;20:871-877.

[8] Hockey, G. R., “Changes in participant efficiency as a function of environmental stress, fatigue and circadian rhythms”, in K. R. Boff, L. Kaufman and J. P. Thomas, Eds., Handbook of Perception and Human Performance, Wiley, New York, 00. 44-1 – 44-9, 1986.

[9] ____, Combat Stress. Department of the Navy, Headquarters United States Marine Corps, Washington, D.C, 2003.

[10] Fowler, C. and Housum, J., “Talkers signaling of new and old words in speech and listeners’ Perception and Use of This Distinction”, J. Memory and Language, Vol. 26, pp.45-49, 1987.

[11] Fant, G., Speech Sounds and Features. Cambridge: MIT Press, MA, 1973.

[12] Dinges, D. F. and Kribbs, N. B., “Performing while sleepy: Effects of experimentally induced sleepiness”, In T. H. Monk (Ed.), Sleep, Sleepiness and Performance. New York: Wiley, 1991.

[13] Horne, JA. and Pettitt, AN., “High incentive effects on vigilance performance during 72 hours of total sleep deprivation”, Acta Psychologica,58, 123–139, 1985.

[14] Elias, N. J., “New statistical methods for assigning device tolerances, Proc. IEEE Int.SYmp. Ccts. Sys., Newton, Mass., USA, pp.329-332, 1975.

[15] Bard, E. G., Sotillo, C., Anderson, A. H., Thompson, H. S., and Taylor, M. M., “The DCIEM Map Task Corpus: Spontaneous dialogue under sleep deprivation and drug treatment. Speech Communication 20:71-84, 1996.

[16] Jiang, H. and Lee, C. H., “A new approach to utterance verification based on neighbourhood information in model space,” IEEE Trans. Speech and Audio Processing, vol. 11, pp.425 – 434, 2003.