2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Feasibility Analysis of Symbolic Representation for Single-Channel EEG-Based Sleep Stages

Zheng Chen[1], Pei Gao[1], Ming Huang[1,*], Naoaki Ono[1,2], MD Altaf-Ul-Amin[1], and Shigehiko Kanaya[1,2]

Abstract— **Sleep screening based on the construction of sleep stages is one of the major tool for the assessment of sleep quality and early detection of sleep-related disorders. Due to the inherent variability such as inter-users anatomical variability and the inter-systems differences, representation learning of sleep stages in order to obtain the stable and reliable characteristics is runoff for downstream tasks in sleep science. In this paper, we investigated feasibility of the EEG-based symbolic representation for sleep stages. By combining the Latent Dirichlet Allocation topic model and comparing with different feature extraction methods, the work proved the feasibility of multi-topics representation for sleep stages and physiological signals.**

I. Introduction

Sleep is the corner stone for healthiness and well-being throughout our life. Getting adequate sleep at nights can help protect our mental health, physical health, and quality of life [1]. Sleep screening based on sleep stages is one of the major tool in assessment of sleep-related disorders, such as sleep apnea syndrome, schizophrenia, depression, insomnia, narcolepsy, and other neural abnormalities. The gold standard for sleep construction is re-defined to five different stages, i.e., wake, rapid eye movement (REM) or non-REM where non-REM stage can be further divided into N1, N2, and N3 according to the American Academy of Sleep Medicine (AASM) [2]. Meanwhile, the stage scoring remains the multi-lead electroencephalogram (EEG) recording by overnight polysomnography (PSG) with manual labeling by sleep experts [3]. The sleep has informative frequency oscillation of EEG waves in 0.5 to 30-35 Hz range. Wakefulness is characterized by alpha (8-12 Hz) and beta frequency rhythms (16-30 Hz). The alpha frequency occupies more than 50% of the epoch for N1 while theta waves (4-8 Hz) are concomitant. N2 corresponds to the epoch in which the theta waves are also noticeable. Meanwhile the sleep spindles and K-complex appear in this stage. N3 refers to a deep sleep (or slow-wave sleep) interval that the presence of delta activity (0–4 Hz) for more than 20% of the epoch is classified as N3 [1]. In REM period, the epoch is scored when saw-tooth waves (or theta waves) and saccadic eye movements are evident. The alpha waves are also predominant during REM stage.

Numerous sleep-related studies are based on the assessment of sleep stages by using EEG recordings, for instance, analysis of insomnia disorder [4], modeling of transition mechanism [5], or developing an automatic system of sleep scoring [6], [7], [8]. In particular, the results in the literatures are promising with combining machine learning (or recent deep learning). The performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied [9]. Therefore, a large amount of the spur effort in deploying workflow of studies goes into the design of preprocessing pipelines, in order to obtain the stable and reliable characteristics, such as hand-crafted features [10], spectrogram [11], empirical mode decomposition [12], and feature mapping neural network [13]. Noteworthy, the large-scale patterns of synchronized neuronal activity (or EEG) are ever changing and thus exhibit a considerable variability over time [14]. This no-stationary nature in real EEG signals inevitably limits statistical data processing with time. In addition, the functional cooperative interaction of brain dynamics always has heterogeneous characteristics of inter-subject, even recording in different time for the same subject. As a consequence, exploring a dominant and reliable representation of EEG is central to understand the sleep construction and to making optimal data-driven strategies for downstream tasks.

One representation that the data mining community has been considered transforming real valued data into symbolic representations, noting such representations would potentially allow researchers to avail of the wealth of data structures and algorithms from the text processing and the machine learning [15]. Moreover, such studies have more recent attention in the sleep stage analysis. Herrera et al., proposed the application of a novel method for symbolic representation of the EEG and evaluated its potential as information source for a sleep stage classifier [16]. To meet the criticism and reveal the latent sleep states, Koch et al., utilized symbolic aggregate approximation (SAX) to transform the sleep epoch of EEG to a mixture of probabilities of latent sleep states and developed an automatic sleep classifier using the Latent Dirichlet Allocation (LDA) topic model [17]. Christensen et al. inspired the idea of Koch et al. and used the same method to analyze the sleep EEG of people with insomnia disorder with a frequency-based sleep analysis procedure, which is describing each epoch as a mixture vigilance states [18]. However, the proposed SAX

This research and development work was supported by a Grant-in-aid for Young Scientists of the Japan Society for the Promotion of Science (JSPS) #20k19923

Zheng Chen, Pei Gao, Ming Huang*, Naoaki Ono, MD Altaf-Ul-Amin, and Shigehiko Kanaya are with Graduate School of Science and Technology, Nara Insitute of Science and Technology, Takayamacho 8916-5, Ikoma, 6300192 Japan. (e-mail: {chen.zheng.bn1, gao.pei.gi3, alex-mhuang, nono, amin-m, skanaya}@is.naist.jp)

Naoaki Ono and Shigehiko Kanaya are with Data Science Center, Nara Insitute of Science and Technology, Takayamacho 8916-5, Ikoma, 6300192 Japan.

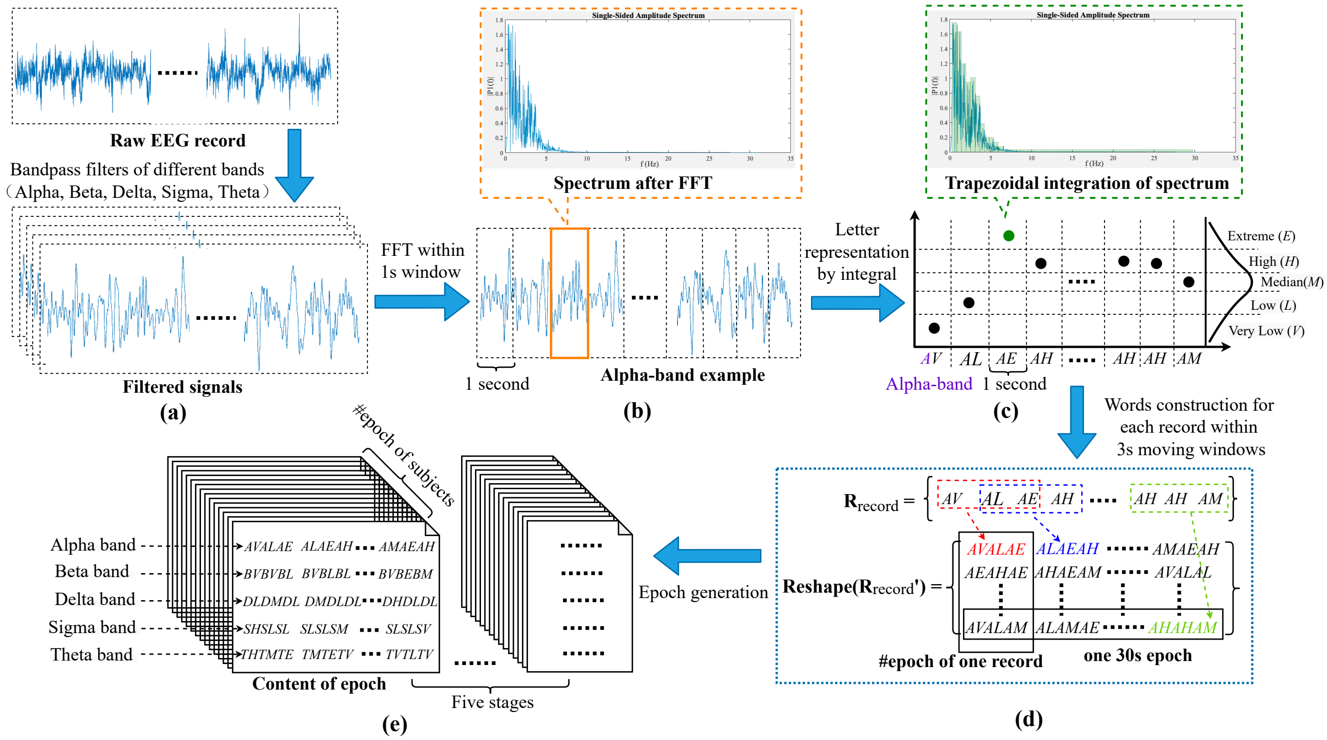

Fig. 1. The framework of symbolization: (a) illustrates the generation of the filtered signals within five classical frequency bands. The FFT is implemented to generate the spectrum for each 1 second showed by (b). Then, each 1s spectrum is to calculate different features (Mean, Trapezoidal integration, and spectral entropy), subsequently, the numerical values of each record are converted to a word sequence by categorizing the relative distribution. (c) illustrates an example for the Trapezoidal integration process within α-band. A 3 s moving window with one stride is utilized to generate the epochs from each record as shown in (d). Finally, the content of one sleep epoch containing 150 words and five sentences (five frequency bands) is illustrated by (e).

has a major limitation, that is, symbols are mapped from the average values of segments. Different segments with similar average values may be mapped to the same symbols, and the represented distance between them is 0 [19]. The topic model is competent in discovering latent, multi-faceted summaries of documents or symbolic data in the Natural Language Processing (NLP) community. Therefore, this paper further investigated that the multifold symbolic representation with LDA model from frequency domain is appropriate for representing and manifesting the latent characterization of sleep stages. The study attempted to explore the representation capability of the different statistical methods to evaluate relevant transformation and capture the spontaneous characteristics of different stages. Evaluating by the data-driven method, the EEG-based symbolic representation of sleep stages can be further incorporated into the pipeline of sleep studies.

II. Material and Method

A. Dataset and Preprocessing

The dataset used in this work is from the Sleep Heart Health Study (SHHS). Access of the SHHS was permitted via the National Sleep Research Resource[1]. The SHHS database consists of two rounds of at-home PSG recordings (SHHS Visit 1 and SHHS Visit 2). Due to the unscored epochs and misaligned records, here, we used only the SHHS Visit 1 containing two channel EEG records (C4-A1 and C3-A2) from 5736 subjects sampled at 125 Hz.

The EEG records have been typically contaminated by various types of artifacts, a 8th order Butterworth bandpass filter with cutoff frequencies between 0.5 and 30 Hz was implemented over all records. These cutoff frequencies for band pass filtering were selected since the brain activities have significant information in 0.5 to 30-35 Hz range during sleep [20]. The records of the SHHS were manually annotated into six classes (W, S1, S2, S3, S4, REM). Noteworthily, we merged S3 and S4 stages into one deep-sleep class referring to the AASM standards.

B. Symbolization of sleep epoch

As shown in Fig. 1, to reveal the inherent characteristics of sleep stages, five filtered signals for each EEG record were generated by clinical frequency bands (δ, θ, α, σ, β). The single-sided spectral analysis was then implemented into each band-pass filtered signal by fast Fourier transform within 1 second/bin window and no-overlapping. Three feature extraction methods (trapezoidal integration, spectral entropy and mean) performed the numerical representation of cumulative spectrum on each bin, in order to create subject-wise distribution of spectral feature within each classical frequency band. Since the generated distributions were normalized to $N(0, 1)$, each subject a histogram was produced for each frequency band and the four boundaries producing five equal proportion bins derived. Meanwhile, five quantification

[1] www.sleepdata.org

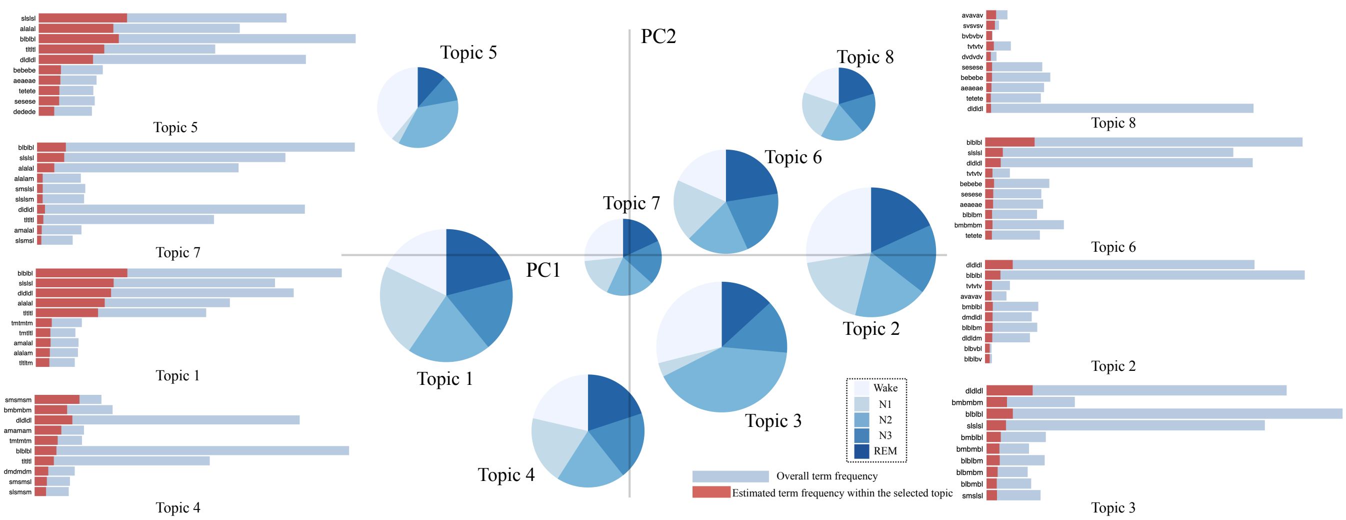

Fig. 2. The visualization of PCA of 8 topics. The histograms show the word distributions and estimated word frequency for each topic. The distribution of the pie chart illuminates the statistical contributions for five stages and the area of each pie chart is proportional to how many documents feature each topic.

categories were labeled as different symbols (Extreme, High, Median, Low, or Very low). Each 1 second numerical value of trapezoidal integration was converted to a word according to the cutoff boundaries and corresponding frequency band. That is, each record was represented into five symbolic sequences.

During the sleep cycles, idiopathic neural oscillatory activities are generated into different sleep stages. For instance, sleep spindles are derived by interplay of the thalamic reticular nucleus during stage 2 of non-REM sleep with a duration of 0.5–2 seconds. To facilitate the mining and representation of the exclusive patterns, a 3 second moving window with one stride (2 bins overlapping) was used to generate the epoch sample which contains 30 6-letter words for each frequency band. Each 6-letter word represents a 3 second spectral pattern of corresponding frequency band (e.g., $AHALAE$: $\alpha(Alpha) - high - low - median$). Consequently, one sleep epoch was described by 150 words and five sentences.

C. Latent Dirichlet Allocation Topic Model

The Latent Dirichlet Allocation is an NLP model which is based on the hypothesis that a document has certain topic or a mixture of different topics. When the documents with a similar topic, the topic can be reflected in the particular vocabulary and the probability distribution of words from the dictionary. The LDA facilitates the explanation of dataset by clustering the features of the data into latent unobserved sets. In this work, the symbolic represented epochs can be regards as a series of documents. Assuming the document (epoch) j has 150 words and w_{ij} is the observed value of word i in document j. LDA will traversal and cluster all the words from each document into T topics to label this document by derived the probability of each topic. This process of LDA is described as following:

- For a topic t, a multinomial parameter σ_t is sampled from Dirichlet prior $\sigma_t \sim D(\omega_1)$.
- For a document j, a multinomial parameter φ_j over the T topics is sampled from Dirichlet prior $\varphi_j \sim D(\omega_2)$.
- For a word i in document j, a topic label τ_{ji} is sampled from discrete distribution $\tau_{ji} \sim Discrete(\varphi_j)$.
- The value w_{ji} of word i in document j is sampled from the discrete distribution of topic $\tau_{ji}, w_{ji} \sim Discrete(\sigma_{\tau_{ji}})$.

Where ω_1 and ω_2 are Dirichlet prior hyperparameters. σ_t and φ_t are hidden variables to be inferred while τ_{ji} can be sampled through a Gibbs sampling procedure.

D. Evaluation

A grid search [5:1:13] for the number of topics was first to find the optimal distributions of words based on the topic coherence metric within different feature engineering. The topic coherence measures score a single topic which is the artifact of statistical inference by measuring the degree of semantic similarity between high scoring words in the topic. This word utilized C_v measurement which is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that by using mutual information and the cosine similarity. Moreover, the LDA results (topics) had been decomposed via PCA to visualize the topic construction and distribution by using the pyLDAvis python packages. To map the topics to the sleep epochs, the probability of topics for epochs were resulted by trained LDA model, and the statistical distribution was generated to evaluate the co-occurrence and importance for each sleep stage.

III. RESULT

Tabel 1 illustrates the coherence metrics for different methods of feature extraction. Over all the grid search of topics, the trapezoidal integration of spectrum has a better performance than the common SAX (mean value) and the spectral entropy. Meanwhile, the fine-tuning of the number of topics is 8 in this word shows the highest coherence for trapezoidal integration.

TABLE I

A GRID SEARCH OF NUMBER OF TOPICS FOR DIFFERENT FEATURE PROCESS METHODS

#Topics	5	6	7	8	9	10	11	12	13
Mean Value	0.281	0.275	0.305	0.316	0.319	0.298	0.291	0.287	0.253
Trapezoidal Integration	0.313	0.306	0.327	**0.351**	0.316	0.295	0.315	0.314	0.314
Spectral Entropy	0.229	0.226	0.234	0.228	0.224	0.222	0.224	0.221	0.213

TABLE II

TOP 3 IMPORTANCE OF TOPICS FOR FIVE SLEEP STAGES

Stages	Wake	N1	N2	N3	REM
Top 3 topics	T2; T5; T7	T1; T8; T4;	T1; T5; T8	T5; T6; T3	T1; T8; T5

The PCA result of the optimal number of topics for trapezoidal integration has shown in Fig. 2. The topics (8-dimensional in our case) were flattened to be only 2-dimensional. There is no overlapping area among topics, that is, the topics and their word distributions are mutually independent. The center of pie charts represent the position of 8 topics in the latent feature space, while the distances between topics illustrate similarity or dissimilarity. Moreover, according to the word distributions, it can group topics 6 and 8 into one since the pie distribution are more similar, and have half estimated very low frequency of θ over all term frequency. Topics 2, and 7 are mainly contributed to wakefulness while these topics have a smaller metric distance. Topics 1 and 2 have high term frequency for σ and β bands according to Fig. 2, and the high term frequency of θ band can be reflected to N2 sleep. Topics 5 is related to N2 and wakefulness stages as shown in Table 2, while the amount of estimated words of θ band (e.g., $tltltl$: $\theta - low - low - low$) is represented to REM stages. Table 2 lists the top 3 importance of topics for different stages. The N1 and N2 have similar construction of topics since these two stages belong to light sleep. The most relevant term for topic 3 is the δ band word while the delta waves (more than 20%) are more related to N3 stage.

CONCLUSION

In this study, we investigated the symbolic representation of EEG records for sleep stages. Comparing with the different feature extraction methods, the work proved the feasibility of an improved representation method of sleep stages and physiological signals based on the data-driven. The further study will extend to systematically explore the sleep construction and to evaluate the downstream study in the sleep community.

REFERENCES

[1] P. Memar and F. Faradji, "A novel multi-class eeg-based sleep stage classification system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. PP, pp. 1–1, 11 2017.

[2] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, "The AASM Manual for the Scoring of Sleep and Associated Events," *American Academy of Sleep Medicine*, vol. 53, no. 9, pp. 1689–1699, 2013.

[3] S. Nasiri and G. D. Clifford, "Attentive adversarial network for large-scale sleep staging," in *Proceedings of the 5th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 126. Virtual: PMLR, 07–08 Aug 2020, pp. 457–478.

[4] J. A. E. Christensen, R. Wassing, Y. Wei, J. R. Ramautar, O. Lakbila-Kamal, P. J. Jennum, and E. J. Van Someren, "Data-driven analysis of EEG reveals concomitant superficial sleep during deep sleep in insomnia disorder," *Frontiers in Neuroscience*, vol. 13, no. JUL, pp. 1–12, 2019.

[5] A. Stevner, D. Vidaurre, J. Cabral, K. Rapuano, S. Nielsen, E. Tagliazucchi, H. Laufs, P. Vuust, G. Deco, M. Woolrich, E. J. W. Van Someren, and M. Kringelbach, "Discovery of key whole-brain transitions and dynamics during human wakefulness and non-rem sleep," *Nature Communications*, vol. 10, 03 2019.

[6] H. Phan, F. Andreotti, N. Cooray, O. Chén, and M. de Vos, "Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, pp. 400 – 410, 01 2019.

[7] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. PP, 03 2017.

[8] H. Phan, F. Andreotti, N. Cooray, O. Chén, and M. de Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, pp. 1285–1296, 05 2019.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 1798–1828, 08 2013.

[10] X. Li, L. Cui, S. Tao, J. Chen, X. Zhang, and G.-Q. Zhang, "Hyclasss: A hybrid classifier for automatic sleep stage scoring," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, pp. 1–1, 02 2017.

[11] Z. Chen, N. Ono, M. Altaf-Ul-Amin, S. Kanaya, and M. Huang, "ivae: An improved deep learning structure for eeg signal characterization and reconstruction," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Seoul, Korea (South): IEEE Computer Society, dec 2020, pp. 1909–1913.

[12] D. JIANG, Y.-n. LU, Y. MA, and Y. WANG, "Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement," *Expert Systems with Applications*, vol. 121, 12 2018.

[13] C. Sun, C. Chen, W. Li, J. Fan, and W. Chen, "A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, pp. 1–1, 08 2019.

[14] A. Kaplan, A. . A. Fingelkurts, S. Borisov, and B. Darkhovsky, "Nonstationary nature of the brain activity as revealed by eeg/meg: Methodological, practical and conceptual challenges," *Signal Processing*, vol. 85, pp. 2190–2212, 11 2005.

[15] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," 01 2003, pp. 2–11.

[16] L. Herrera, A. Mora, C. Fernandes, D. Migotina, A. Guillén, and A. Rosa, "Symbolic representation of the eeg for sleep stage classification," *International Conference on Intelligent Systems Design and Applications, ISDA*, pp. 253–258, 11 2011.

[17] H. Koch, J. Christensen, R. Frandsen, M. Zoetmulder, L. Arvastson, S. Christensen, P. Jennum, and H. Sorensen, "Automatic sleep classification using a data-driven topic model reveals latent sleep states." *Journal of neuroscience methods*, vol. 235, 07 2014.

[18] J. Christensen, R. Wassing, Y. Wei, J. Ramautar, O. Lakbila-Kamal, P. Jennum, and E. Someren, "Data-driven analysis of eeg reveals concomitant superficial sleep during deep sleep in insomnia disorder," *Frontiers in Neuroscience*, vol. 13, p. 598, 07 2019.

[19] Y. Sun, J. Li, J. Liu, B. Sun, and C. Chow, "An improvement of symbolic aggregate approximation distance measure for time series," *Neurocomputing*, vol. 138, p. 189–198, 08 2014.

[20] M. Nishida, J. Pearsall, R. Buckner, and M. Walker, "Rem sleep, prefrontal theta, and the consolidation of human emotional memory," *Cerebral cortex (New York: 1991)*, vol. 19, pp. 1158–66, 11 2008.