An Approach for Emotion Recognition using Purely Segment-Level Acoustic Features

Hao Zhang[1], Shin'ichi Warisawa[2], and Ichiro Yamada[3]

[1] *School of Engineering, The University of Tokyo, Japan, zhanghao@lelab.t.u-tokyo.ac.jp*

[2] *School of Engineering, Graduate School of Frontier Science, The University of Tokyo, Japan, warisawa@k.u-tokyo.ac.jp*

[3] *School of Engineering, Graduate School of Frontier Science, The University of Tokyo, Japan, yamada@k.u-tokyo.ac.jp*

Abstract: A purely segment-level approach is proposed in this paper that entirely abandons the utterance-level features. We focus on better extracting the emotional information from a number of selected segments within utterances. We designed two segment selection approaches (miSATIR and crSATIR) for selecting utterance segments for use in extracting features that are based on information theory and correlation coefficients to create the purely segment-level concept of the model. We established a model using these selected segment-level speech frames after clarifying the time interval for the segments. Testing has been carried out on a 50-person emotional speech database that was specifically designed for this research, and we found that there were significant improvements in the average level of accuracy (more than 20%) compared to that using the existing approaches for all the utterances' information. The test results that were based on the speech signals stimulated by the International Affective Picture System (IAPS) database showed that the proposed method could be used in emotion strength analyses.

Keywords: Emotion recognition, Human-computer interface (HCI), Segment-level features, Probabilistic neural network (PNN), Emotion strength analysis.

1. INTRODUCTION

Interest in using speech for emotion recognition has recently increased because of the significant developments made in human-computer-interface (HCI) technology (Picard, 2000). Speech is the most natural and efficient type of human communication and this has inspired researchers to consider speech as an approach to human and machine interaction. However, simply adding

emotional intelligence to computers is a challenge, although it could lead to a meaningful evolution of the relationships between humans and automated systems, although these systems would largely benefit from knowing and adapting their operations according to the users' emotional states. Recognizing human emotion from speech also introduces some promising applications, such as in the emotion analysis of commercial conversations, virtual humans, emotion-based indexing, and in information retrieval, (Morrison, Wang, & De Silva, 2007) and (Gratch & Marsella, 2005).

The utterances (phrases, short sentences, etc.) referred to in current research papers are often considered the fundamental unit and are recognized based on the global utterance-wise statistics of the derived segment low-level descriptors (LLD), so the segment features are transformed into a single feature vector for each emotional utterance (Ververidis & Kotropoulos, 2006), (Kim, Hyun, Kim, & Kwak, 2009), (Qi-Rong & Zhan, 2010). However, many researchers have recently been focusing on an issue that questions whether or not the utterance level is the right choice for modeling emotions (Schuller & Rigoll, 2006). They are concerned with this because of the difficulties in using the utterance-wise statistics in avoiding the influence from spoken content, which requires accurate partitioning of an utterance for segmentation. Moreover, valuable but neglected information could be used in the segment-level feature extraction approach rather than calculating only the utterance-wise statistics. Many researchers also support this hypothesis (Schuller & Rigoll, 2006), (Yeh, Pao, Lin, Tsai, & Chen, 2011), based on the fact that improvements can be made by adding segment-level features to the common utterance-level features.

Motivated by these findings, we focus on developing a novel scheme for improving speech emotion recognition using segment-level features instead of using a strategy that includes ensemble learning from different classifiers using the same utterance-wise features (Morrison et al., 2007), (Schuller, Lang, & Rigoll, 2005). We took into consideration a purely segment-level strategy for speech emotion recognition and abandoned the utterance-wise features in order to reduce the noises from spoken content and use the neglected information in the calculation of the utterance-wise statistics in this study. The aim of this paper is to properly design an approach for utterance-level emotion recognition that is based on aggregating the segment-level labels without introducing more computational complexity.

Our experimental design is introduced in Section 2. In Section 3, the analytical method for emotion recognition using segment-level speech frames is described. The results are discussed in Sections 4 and 5. In Section 6, an application example for the proposed approach is introduced, and finally, in Section 7, our conclusion is drawn on the proposed research.

2. EXPERIMENTAL DESIGN

A well-annotated database is needed to construct a robust model for recognizing emotions using speech signals. Our experiment emphasizes "natural speech". The participants are prevented from becoming aware that they are in an experimental environment during the experiments, which is much more realistic than experiments that were conducted using scripted speech. Natural speech is difficult to analyze, but more suitable than scripted speech for validating the robustness of an emotion analysis method.

2.1. Experimental protocols

The experimental setup is composed of one instructor, one coordinator, and two participants. The coordinator cooperates with the participants in order to help better stimulate their emotions. However, the coordinator pretends to be one of the participants in the experiment to avoid being an

extra obstruction for the real participants. The stimulation process unfolds through conversations with the aid of videos. The steps are demonstrated as follows.

- The instructor sets up the experiment's environment, such as a projector for the videos and microphones for collecting the speech signals, and gives instructions to the participants.
- The instructor also explains the steps to the participants, including the coordinator, for freely providing their impressions related to the videos.
- Self-introductions are made to create an easy speaking atmosphere.
- After watching each emotion evoking video, which lasts several minutes, the speech signals are recorded from the impressions.

The emotion corresponding to each utterance from the recorded speech signals is not only self-assessed by the participants but also by ten other people after the experiment. Therefore, it is possible to evaluate the degree of reliability of the utterances emotion labeling.

2.2. Data information

Ninety six people participated in the experiments, which included 53 males and 43 females ranging from their early teens to their 40s. We provided the sample selections for obtaining reliable data in two steps. First, only the samples with the same label (pleasure or displeasure) based on the self-assessment and others-assessment were taken into consideration. Second, for maintaining a balance between the sample numbers for each label, we selected 300 utterances with higher rankings using the others-assessment, which consisted of 150 utterances as pleasure data and 150 utterances as displeasure data from the 50 participants. Ten specialists put a label on every utterance in the others-assessment, and the rank for each utterance was calculated based on the ratio of the numbers from the specialists who gave labels that were consistent with the label set from the self-assessment.

3. METHODOLOGY

The proposed methodology for emotion recognition is based purely on the segment-level speech frames, and the important issues for consideration here are the increased decline in the number of samples in the generalization ability of the classifier. In our work, we address the quantitative analysis of various analytical schemes related to segment-level speech emotion recognition, and we propose an automatic approach for decreasing the number of samples in order to reduce the computational complexity and improve the classifier generalization ability.

3.1. Feature extraction

We focused on a set of 162 acoustic features from speech signals, including 50 Mel-Frequency Spectral Coefficients (MFCC) (Davis & Mermelstein, 1980), 50 Linear Predictive Coefficients (LPC) (Atal & Hanauer, 1971), and 10 statistical features (mode, median, mean, range, interquartile range, standard deviation, variation, absolute deviation, skewness, and Kurtosis) calculated from each of the five levels of the detailed wavelet coefficients by using the Discrete Wavelet Decomposition (DWT) (Grossmann & Morlet, 1984), pitch, energy, zero-crossing rate (ZCR), the first seven formants, centroid, and 95%-roll-off-point from FFT-spectrum.

3.2. Segmentation approach

3.2.1. Existing segmentation schemes

Several segmentation strategies (Figure 1) were proposed in a previous study (Schuller & Rigoll, 2006), their concepts are introduced in the following.

- *GTI segmentation (Utterance-level segmentation).* The speech signals are segmented by pauses during the speech without word or syllable boundary detection.
- *ATI segmentation.* Different from utterance-level segmentation, the speech utterances are segmented at the same fixed time interval.
- *RTI segmentation.* Speech utterances are segmented at the fixed relative positions.
- *ATIR segmentation.* ATIR segmentation combines the ideas of ATI and RTI segmentation. Fixed-length segments are constructed at fixed relative positions, and this overcomes the drawback of different segment lengths and numbers obtained from different utterance lengths.

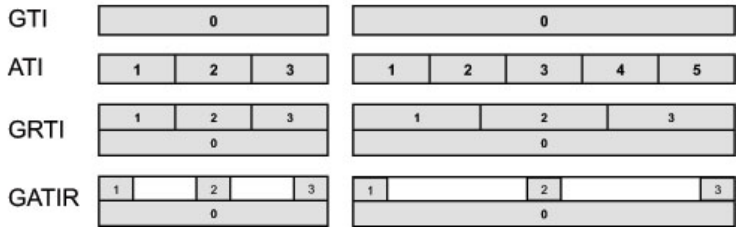

Figure 1: Illustrations of segmentation schemes. GTI: global time intervals (Utterance-level segmentation); ATI: absolute time intervals; RTI: relative time intervals; ATIR: absolute time intervals at relative positions; GRTI: combination of utterance-level and segmental features extracted using RTI segmentation; GATIR: combination of utterance-level and segmental features extracted using ATIR segmentation.

3.2.2. Proposed segmentation approaches

The following difficulty must be overcome in order to carry out segment-level based speech emotion recognition, which concerns defining the labels of the segments representing the classifiers. We propose the novel segmentation strategies, entropy-based ATIR (eATIR), mutual information-based ATIR (miATIR), and correlation coefficients based ATIR (crATIR) inspired from the ATIR segmentation method because of its advantages of getting a smaller and fixed number of segments from an utterance. We adopt a more efficient way for finding more informative segments by minimizing the amount of mutual information between feature vectors. Fixed-length segments are constructed in this study at selected positions based on the designed indexes. So, not all the segments of an utterance are used in the analysis. Framing is used to deal with the utterance signal after dividing the utterance into several sections. A 10-ms window with no overlap is used for calculating the ranking of the fixed-length segment. The proposed segmentation methods are illustrated in Figure 2.

We use several measuring indexes on the extracted features to get the most representative ones from among all the segments, and the average of the index values within each segment is used for the selection. The indexes are introduced in the following paragraphs.

Entropy index: Entropy is a measure of the information content, which is introduced as "a measure of how much 'choice' is involved in the selection of an event" (Shannon, 2001). We have to determine how to obtain the best approximation for creating the model in order to model an emotion with segments that are represented by an unknown probability distribution. One approach is to have the distribution with the maximum entropy ensure that the approximation satisfies and subjects to any constraints on the unknown distribution (Jaynes, 1957). A number of top-ranked segments are selected in our study after calculating the entropy of the segment features.

Mutual information index: The mutual information (Shannon, 2001) measures the "lumpiness" of the joint distribution, and several of the most informative segments can be selected by minimizing the redundant information.

Correlation coefficient index: The correlation coefficient (Pearson, 1895), which is also known as the Pearson product-moment correlation coefficient, is a measure of the linear dependence between two feature vectors. This index shares the same concept with the mutual information for reducing the redundancy.

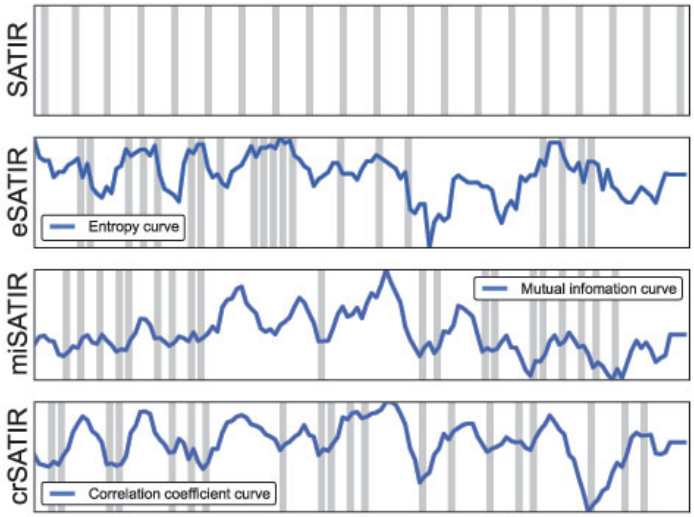

Figure 2: Fixed-length segment positions illustration using proposed segmentation approaches (20-segment selecting situation is shown and the positions are represented using grey lines). 'S' is included in the abbreviation to represent the purely segment-level concept

3.3. Decision model

The decision for determining the emotion of an utterance is based on the prediction of its segments from a classifier. We simply use the efficiency approach called the majority vote, which determines the label of the utterance from the label in majority so that we can pay more attention to examining the effectiveness of the proposed segment-level approaches for speech emotion recognition. The decision model is shown in Figure 3. Our decision model is based on a classifier called the probabilistic neural network (PNN) (Specht, 1990).

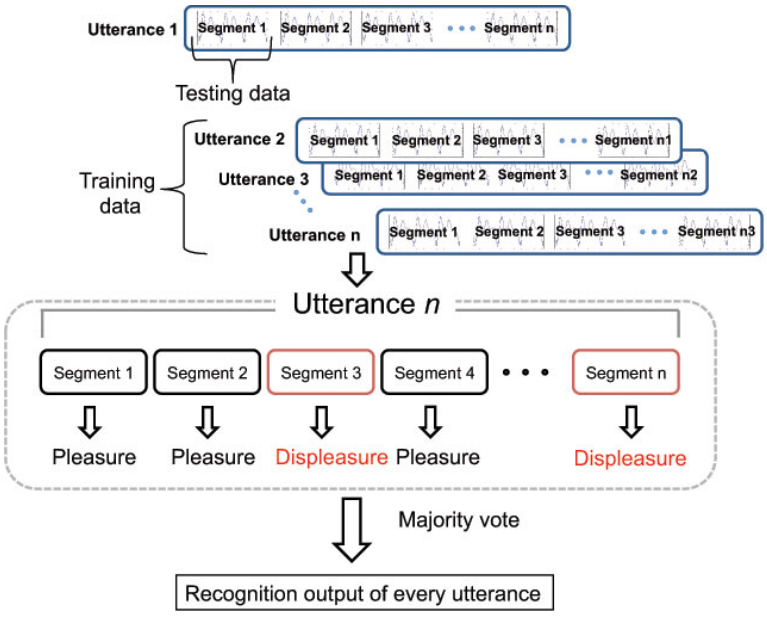

Figure 3: Illustration of segment-level classification concept for decision model.

4. RESULTS

A 10-fold cross validation is used to evaluate and test our proposed approaches as well as make comparisons with previous researches because it is used in many other emotion recognition researches for validating general models (Schuller & Rigoll, 2006), (Yeh et al., 2011), (Yu & Xu, 2007). We reviewed all the most recent research on the aspect of classifiers for creating a solid illustration and found that the support vector machine (SVM) is one of the most robust and popular classifiers in the field of affective researches, and it beats out many other kinds of classifiers in terms of the recognition accuracy (Morrison et al., 2007), (Schuller & Rigoll, 2006), (Chandaka, Chatterjee, & Munshi, 2009). Thus, our evaluation results based on PNN are compared with those based on SVM. Table 1 classifies the results when using our 162 proposed acoustic features applied to the existing schemes, such as the GTI and segment-level features that include schemes such as GRTI and GATIR. 500-ms segments are constructed at fixed relative positions for GATIR (Schuller & Rigoll, 2006).

Table 1: Comparison of emotion recognition accuracy between existing segmentation schemes using segment features with global features

Feature groups	**Accuracy (%)**	
	SVM	**PNN**
GTI	67.0	70.0
GRTI (utterance + 2 segments)	69.0	80.0
GRTI (utterance + 3 segments)	69.7	83.3
GATIR (utterance + 2 segments)	68.0	76.7
GATIR (utterance + 3 segments)	69.3	80.0

As listed in Table 1, the emotion recognition accuracy is higher when using PNN then for SVM in every case, and therefore, we chose PNN for our purposes. These results are also consistent with other similar researches (Schuller & Rigoll, 2006), (Yu & Xu, 2007), which show there is a better potential for use as segment features for emotion recognition extracted by using RTI segmentation. Hereafter, we begin our pure segment features with a name beginning with *S* in order to discriminate those from the global features or both proposed in previous researches. We use a SATI (ATI) with different time intervals to conduct a first scenario that is different from that in the utterance-level analysis and finally choose 50 ms as the segment length for our analysis due to the performance.

We compare our proposed segment-level feature extraction strategies, which generate feature groups called eSATIR, miSATIR, and crSATIR in the sentences and paragraphs that follow. We also compare them with a segment-level feature extraction approach called SATIR, which is directly inspired from previous research. We have to define the number of segments we want to generate from each utterance before applying the proposed strategies. Since we use a majority vote in the decision model, we considered 10 segments as reasonable for use as the smallest number for voting. We also take note of the fact that some utterance lengths are as short as one second, and thus, 20 segments is considered to be the largest number for a majority vote. Table 2 provides a comparison of the emotion recognition accuracy between different segmentation schemes using purely segmented features. According to these comparison results, it is clear that 20 segments

better captures useful information and the voting from 20 segments provides for more fault-tolerance. The results also show that miSATIR and crSATIR can greatly increase the utterance-level speech emotion recognition accuracy, where crSATIR leads to the best results.

Table 2: Comparison of emotion recognition accuracy between different segmentation schemes using purely segmented features

Feature groups	Accuracy (%)	
	10 segments	20 segments
SATIR	73.7	88.3
eSATIR	75.0	89.7
miSATIR	90.3	95.7
crSATIR	92.3	98.1

5. DISCUSSION

This research further develops this new train of thought to a level that totally abandons the global features from the utterances. The analytical results in Section 4 indicate the robustness of this advancement, which leads to a higher level of recognition accuracy by using only the segment-level features in the proposed decision model. 50 ms is chosen as the segment length not only because of the better level of accuracy, but also because of the emotional variability and content. We proposed several segmentation approaches in order to select the appropriate number of segments within an utterance. SATIR is simply extracted from the selected segments at fixed positions according to the length of the utterance, and the recognition results can be seen as a reference. We introduce two kinds of features from the segmentation approaches that are based on information theory including eSATIR and miSATIR, but the improvement when using miSATIR is much larger than that for eSATIR, where eSATIR can only slightly increase the level of accuracy compared to SATIR. Since eSATIR doesn't take into consideration the relationship between the utterance and its segments, the segments generated by considering only the maximum entropy have a higher chance to obtain confusing information such as noise or non-representative information for the utterance label and this might be the cause for differences. However, the miSATIR and crSATIR approaches generate segments for the decision model with less redundant information, which contributes to a better understanding of the utterance label and a better comprehensibility of the learned model.

As Table 2 specifies, the crSATIR results outperform those of miSATIR. This phenomenon seems hard to explain at first glance. In lots of situations, they share the same purposes and the mutual information makes it more powerful for detecting the complex non-linear relationships between two feature vectors, and this helps in problem solving for a lot of difficult issues such as in network analysis (Battiti, 1994), (Steuer, Kurths, Daub, Weise, & Selbig, 2002). However, the situation for feature extraction is different from that in our case, with the aim of better extracting information for representing an utterance label. The mutual information, which is the strategy adopted in our research in order to reduce the amount of redundant information when extracting features from segments, is able to detect these non-linear relationships and avoid them to a large extent. However, the correlation coefficients strategy investigates whether a linear relationship

exists between feature vectors and minimizes it. In our case, two feature vectors with a small or no linear relationship might have a strong non-linear relationship, and this dynamic relationship can be considered useful information for better representing the utterance labels, which might benefit the decision model learning, so using crSATIR leads to an improved level of accuracy compared with using miSATIR.

6. APPLICATION PERSPECTIVE: EMOTION STRENGTH ANALYSIS

A very interesting potential application area is emotion strength analysis using segment-level speech emotion recognition. We use majority voting for the utterance labels prediction with the assumption that the label with the most predicted segments represents the utterance label. For better understanding the segment labels, we further looked into the ratio of the most predicted label that can represent the strength of the utterance emotion. SATI is used because we want to examine all the segments in terms of the emotions.

6.1. Experimental data

We used the International Affective Picture System (IAPS) (Lang, Bradley, & Cuthbert, 1999) for evoking emotions with different strengths. The IAPS is an emotion stimulation system built from the results of many emotion experiments. The picture system is composed of about 1000 pictures labeled with a standard scale of valence (pleasure-displeasure) and arousal (exciting-sleepy). Therefore, it meets our requirement for stimulating emotions at different strengths. Figure 4 shows the four kinds of emotions we defined using the IAPS.

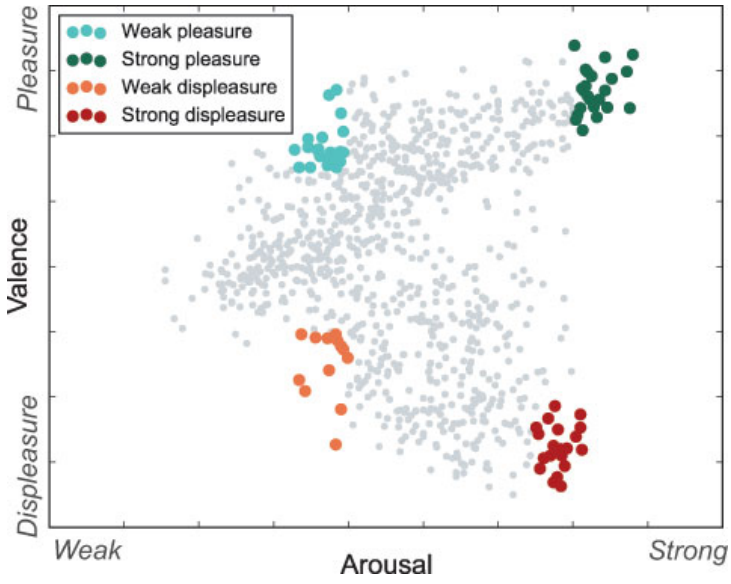

Figure 4: Defined emotion strength based on IAPS

The experimental approach was made up of four parts according to the pleasure and displeasure emotion stimulation, which includes the defined emotion strength (weak and strong). The detailed experimental protocol is demonstrated as follows. The pictures selected from the IAPS during the stimulation period (1 min.) were projected on a screen to evoke emotions. Participants were asked to read designed script with their evoked emotion while viewing each picture and their emotional speech signals were collected. They were requested to close their eyes to relax (2.5 min.) after viewing each picture and a control time of one minute while viewing a white screen was set before viewing the affective pictures. Seven Japanese males took part in the experiment. Data was

collected using the previously described procedures for estimating the emotional strength, which contains 312 samples including 156 pleasure (78 strong, 78 weak) and 156 displeasure (78 strong, 78 weak) data.

6.2. Results

We statistically analyzed the components of all the samples and then visualized the percentage of positive and negative emotion components using a bar chart with a standard deviation to illustrate the correlations between stimulations (Figure 4) and the output of the emotion components represented by using the segment-level predictions within an utterance using the proposed segment-level speech emotion recognition method. The results are shown in Figure 5.

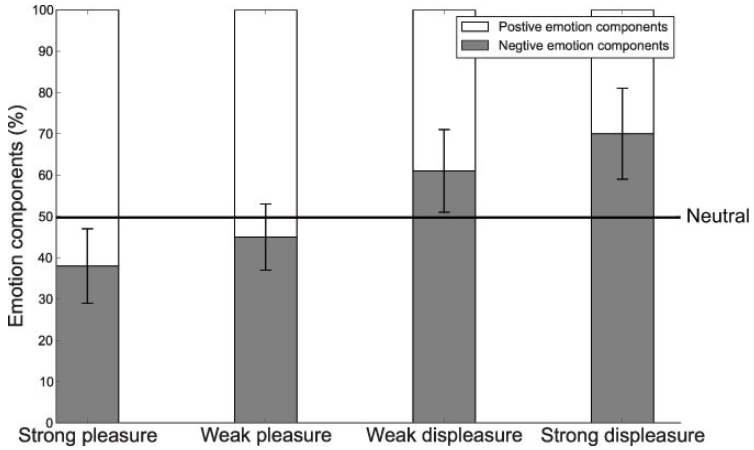

Figure 5: Statistical analysis for emotion components using segment-level speech emotion analysis for all speech samples

6.3. Discussion

The emotion recognition of utterances is one of the more attractive topics in speech analysis for HCI. However, emotion strength analysis has been a very essential but difficult research area. We discussed the potential for using segment-level frames for the emotion strength analysis within utterances. As shown in Figure 5, the proposed method can indeed reflect the strengths of emotions in utterance clusters for a number of spoken phrases or short sentences over a short period of time. However, difficulties still exist in applying it to a single utterance because of the variances in the emotional components regarding these utterances. Although further validation is necessary for collecting more solid findings in terms of the emotion strength analysis of utterances, segment-level speech emotion analysis creates a new focus for better recognizing human emotion strength using machines.

7. CONCLUSION

We proposed a novel emotion recognition strategy that uses only the segment-level features instead of the entire utterance-level features or both. We first introduced advanced relative segmentation methods using mutual information (miSATRI) and correlation coefficients (crSARTI), which can greatly increase the emotion recognition accuracy to more than 95% and 98%, respectively, using a validation database of speech signals from 50 participates in order to make the proposed method more efficient and accurate.

The proposed method also showed effectiveness in determining the emotional strengths of utterances over a period of time. It can provide hints about the emotion strength information

according to our validation results using the IAPS database.

ACKNOWLEDGMENTS

The authors would like to acknowledge Guillaume Lopez at Aoyama Gakuin University, Masaki Shuzo at Kanagawa University, AGI Inc., and all the participants for their significant contributions to the experiments.

REFERENCES

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America, 50, 637.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. Neural Networks, IEEE Transactions on, 5(4), 537-550.

Chandaka, S., Chatterjee, A., & Munshi, S. (2009). Support vector machines employing cross-correlation for emotional speech recognition. Measurement, 42(4), 611-618.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech and Signal Processing, IEEE Transactions on, 28(4), 357-366.

Gratch, J., & Marsella, S. (2005). Evaluating a computational model of emotion. Autonomous Agents and Multi-Agent Systems, 11(1), 23-43.

Grossmann, A., & Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM journal on mathematical analysis, 15(4), 723-736.

Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical review, 106(4), 620.

Kim, E. H., Hyun, K. H., Kim, S. H., & Kwak, Y. K. (2009). Improved emotion recognition with a novel speaker-independent feature. Mechatronics, IEEE/ASME Transactions on, 14(3), 317-325.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Technical manual and affective ratings: Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. Speech communication, 49(2), 98-112.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58(347-352), 240-242.

Picard, R. W. (2000). Affective computing: MIT press.

Qi-Rong, M., & Zhan, Y.-z. (2010). A novel hierarchical speech emotion recognition method based on improved DDAGSVM. Computer Science and Information Systems/ComSIS, 7(1), 211-222.

Schuller, B., & Rigoll, G. (2006). Timing levels in segment-based speech emotion recognition. Paper presented at the INTERSPEECH, Pittsburgh, Pennsylvania, USA.

Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3-55.

Specht, D. F. (1990). Probabilistic neural networks. Neural networks, 3(1), 109-118.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. Bioinformatics, 18(suppl 2), S231-S240.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9), 1162-1181.

Yeh, J.-H., Pao, T.-L., Lin, C.-Y., Tsai, Y.-W., & Chen, Y.-T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. Computers in Human Behavior, 27(5), 1545-1552.

Yu, F. B. J. Y. Y., & Xu, D. (2007). Decision Templates Ensemble and Diversity Analysis for Segment-Based Speech Emotion Recognition. Paper presented at the ISKE 2007, San Diego, CA, USA.

BIOGRAPHY

Hao Zhang is working toward the PhD degree in the Department of Mechanical Engineering at the School of Engineering from The University of Tokyo. His research interests are about helping people improve their lifestyles and prevent lifestyle diseases using signal processing, machine learning, and wearable sensing technologies. He has worked on the research and helped the development of wearable dietary habit monitoring system, etc. He is currently conducting research on emotion recognition system based on speech and physiological signals.

Shin'ichi Warisawa received his B.Eng., M.Eng., and D.Eng. in Mechanical Engineering from The University of Tokyo in 1989, 1991, 1994, respectively. He is an associate professor of Graduate School of Frontier Science at The University of Tokyo. He is currently focusing on Nanomechanics that covers design, fabrication, measurement and application of Nano Electromechanical Systems (NEMS), and also promoting its application to wearable sensors utilized in a preventive healthcare service. He is a member of Japan Society of Mechanical Engineers, Japan Society of Precision Engineering, Japan Society of Applied Physics, Robotics Society of Japan and more.

Ichiro Yamada received his PhD in mechanical engineering from The University of Tokyo in 1985. He was a director of NTT (Nippon Telegraph and Telephone Corp.) Lifestyle and Environmental Technology Laboratories, and is presently a professor of Graduate School of Frontier Sciences, The University of Tokyo. He has worked on research and development of optical MSS (mass storage system), fuel-cell energy system, wearable sensors and their applications, etc. He is promoting the research of Human and Environmental Informatics, and is presently interested in wearable sensing systems for preventive healthcare monitoring.