# MULTIMODAL AROUSAL RATING USING UNSUPERVISED FUSION TECHNIQUE

*Wei-Chen Chen*[1], *Po-Tsun Lai*[1], *Yu Tsao*[2], *Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan

## ABSTRACT

Arousal is essential in understanding human behavior and decision-making. In this work, we present a multimodal arousal rating framework that incorporates minimal set of vocal and non-verbal behavior descriptors. The rating framework and fusion techniques are unsupervised in nature to ensure that it can be readily-applicable and interpretable. Our proposed multimodal framework improves correlation to human judgment from 0.66 (vocal-only) to 0.68 (multimodal); analysis shows that the supervised fusion framework does not improve correlation. Lastly, an interesting empirical evidence demonstrates that the signal-based quantification of arousal achieves a higher agreement with each individual rater than the agreement among raters themselves. This further strengthens that machine-based rating is a viable way of measuring subjective humans' internal states through observing behavior features objectively.

***Index Terms***— behavioral signal processing, affective computing, arousal rating, multimodal signal processing

## 1. INTRODUCTION

Emotion is a fundamental attribute that governs human's behavioral production/perception and decision-making process. In the past decade, affective computing [1], a blooming field in engineering dedicated to compute human emotion from measurable signals, has made significant progresses along with the various simultaneous advancement in the human-machine interface design, e.g., virtual human [2], natural dialog interface [3], intelligent tutoring system [4], etc. Furthermore, researchers, from fields in psychology, psychiatry, and behavioral science, have also long been interested in understanding the behavioral manifestation of human emotion. In fact, different psychologically and clinically validated instruments (e.g., PANAS [5], BIS/BIA[6]) are often utilized to assess the internal emotional state of a person in order to aid the psychologists or clinicians to analyze behaviors or diagnose symptoms of interest.

The emerging interdisciplinary field of behavioral signal processing [7] , which models human behaviors using quantitative computational framework, aims at transforming different domain experts' decision-making process. In this work, we concentrate on computing a particular aspect of emotion description, arousal (a.k.a., activation), i.e., a dimensional representation of emotional state corresponding to how activated a person is. For example, angry and excitement are both high *arousal* emotional state. Significances in understanding arousal have also been greatly substantiated in studies across various domains; for example, interpersonal intimacy is theorized to be a function of arousal [8], infants' arousal is related to their mothers' symptoms of depression [9], the relationship between verbal synchrony and emotional state in couple therapy [10], etc. .

Previous engineering works on computing arousal from behavior signals have been concentrated on utilizing supervised machine learning techniques (e.g., [3, 11, 12]), . This approach has resulted in a large body of works and has demonstrated the effectiveness of training machines to automatically recognize arousal states. Yet, a key missing component that many domain experts rely on remains to be a lack of readily-applicable and interpretable computational framework to quantify affect (arousal) robustly. Recently, Bone et al. presented a vocal-based affect scoring method using unsupervised (rule-based) technique that incorporates a minimal set of knowledge-inspired vocal features [13]. That framework was interpretable, scale-continuous (as opposed to usual discrete *n*-class arousal states), and operational without much manual human labeling involved; the framework has further been demonstrated to achieve cross-domain robustness.

Numerous past studies have demonstrated that different communicative modalities of behavior all contribute to the encoding of emotion information (e.g., [14, 15]). Hence, we extend upon that previous work done by incorporating multimodal behavior features, i.e., nonverbal features including head movements and facial expressions, into an unsupervised arousal rating fusion framework. The core ideas behind the proposed framework in this work are the following:

- **Interpretable features**
- **Scale-continuous scoring**
- **Unsupervised technique**

In this work, we demonstrate that our proposed multimodal arousal rating framework improves over vocal-only arousal rating framework. Furthermore, additional experiments show that the unsupervised fusion technique performs equally-well compared to the supervised linear regression-based fusion

Thanks to Ministry of Science and Technology for funding (103-2218-E-007-012-MY3).

technique. Lastly, our experiment further implicates that the signal-based arousal rating achieves a higher agreement with each individual subjective arousal rating than the agreement among evaluators; an initial empirical evidences indicating that objective unsupervised method for modeling emotion is potentially viable eliminating the issue around subjectivity.

The rest of the paper is organized as follows: section 2 describes about research methodology, section 3 details the experimental setup and results, and section 4 concludes with discussion and future works.

## 2. RESEARCH METHODOLOGY

### 2.1. Database

We use the USC IEMOCAP database for the present study [16]. The USC IEMOCAP was collected for the purpose of studying different modalities in expressive spoken interactions. The database was recorded in five dyadic sessions, and each session consists of a different pair of male and female actors engaging in spoken interactions. The database included a total of 10 actors with approximate 12 hours of data.

The behavior modalities in the database consist of: high-definition directional microphone recordings and motion captured data. During each dialog, 61 markers (two on the head, 53 on the face, and three on each hand) were attached to one of the interlocutors to record $(x, y, z)$ positions at 60Hz of each marker. Figure 1 illustrates the placement of the markers. The markers were then placed onto the other actor and recorded again with the same set of interaction to complete a session. The recorded speech data from both subjects were available for every dialog. The database was manually segmented by humans into utterances. Emotion attributes, e.g., activation (i.e., arousal) level (from 1 to 5), were evaluated by multiple humans at the utterance-level. In this work, out of the total 10039 available utterances, we utilize a subset, 2401 utterances. This set includes the utterances where markers' information are available and no overlapping talks occur.

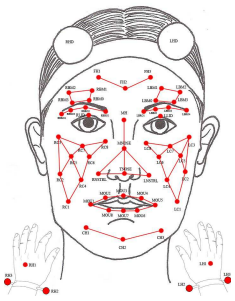

**Fig. 1**. Motion Capture Markers Placement.

### 2.2. Multimodal Behavior Features

The two modalities of behavior features that we extract correspond to vocal (prosody) and non-verbal (i.e., head and facial movements) information. Verbal behavior descriptors are exactly the same set as used Bone et al. i.e., median of pitch, intensity, and HF500 over an utterance [13].

The behavior manifestation of arousal can be imagined as the degree of *higher* or *calmer* movements. In this work, we construct the following minimal and interpretable set of 15 non-verbal features (7 from face, 8 from head) from the $(x, y, z)$ coordinate of the markers per frame, 60 Hz (please refer to Figure 1 for markers' naming reference, and Euclidean distance is utilized for all distance calculations):

**MOU**: Mouth opening quantified as the average distance of marker MOU1-8 to their center point

**LBRO**: Left eyebrow movement quantified as the average distances moved of marker LBRO2-3

**RBRO**: Right eyebrow movement quantified as the average distances moved of marker RBRO2-3

**LLID**: Left eye blinking movement quantified as the distance of marker LLID's moved

**RLID**: Right eye blinking movement quantified as the distance of marker RLID's moved

**LHD**: Left forehead movement quantified as the distance of marker LHD moved

**RHD**: Right forehead movement quantified as the distance of marker RHD moved

Note that the features above are computed by normalizing the coordinate with respect to the marker TNOSE to ensure that head movement does not contribute to the calculation. The rest of the eight features are computed after doing singular value decomposition (SVD) on the original raw coordinates to obtain the six degrees of head orientation and movement, i.e., translation in $x, y, z$ and angles of $pitch, roll, yaw$ [16].

**Head$_{\mathbf{M}}$**: Overall head movement quantified as the average distance of head moved in $(x, y, z)$

**Head$_{\mathbf{vM}}$**: Overall variation in head movement quantified as the distance of Head$_{\mathrm{M}}$ moved

**Pitch$_{\mathbf{M}}$**: Head movement quantified as the distance changed in the angle of pitch

**Pitch$_{\mathbf{vM}}$**: Head movement variation quantified as the distance changed in Pitch$_{\mathrm{M}}$

**Roll$_{\mathbf{M}}$**: Head movement quantified as the distance changed in the angle of roll

**Roll$_{\mathbf{vM}}$**: Head movement variation quantified as the distance changed in Roll$_{\mathrm{M}}$

**Yaw$_{\mathbf{M}}$**: Head movement quantified as the distance changed in the angle of yaw

**Yaw$_{\mathbf{vM}}$**: Head movement variation quantified as the distance changed in Yaw$_{\mathrm{M}}$

Note that translation can be imagined as instantaneous velocity, features such as Head$_{\mathrm{vM}}$ can be conceptualized as the degree of "irregularity" in the head movements. The above 15 features are computed per frame, and we take the *median* of each of the features over an entire utterance as a single measure to be used for arousal scoring (see section 2.3.1).

### 2.3. Unsupervised Arousal Rating Score

The overall procedure in computing our proposed multimodal arousal scoring for a given utterance is shown in Figure 2.

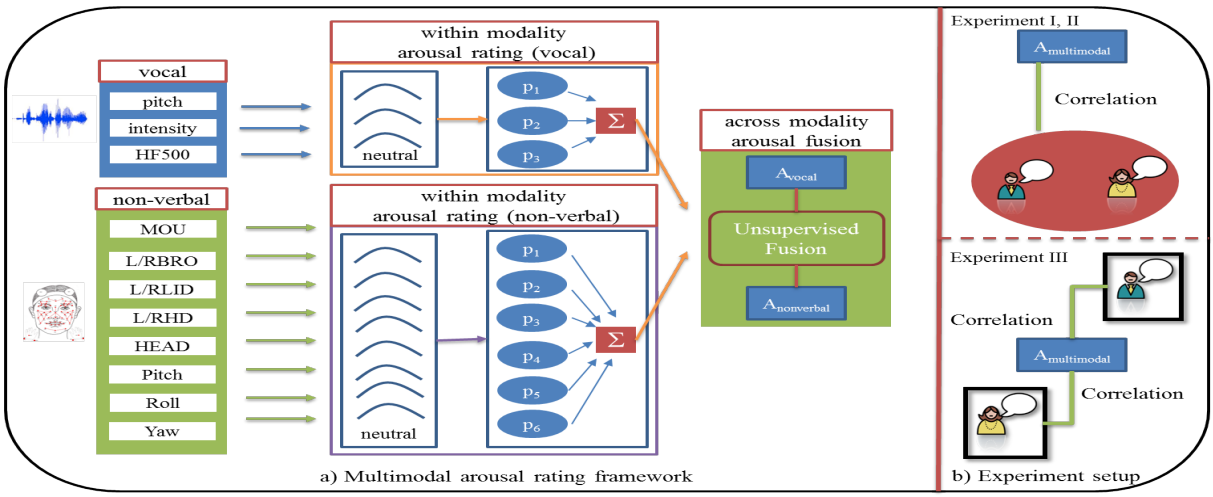


**Fig. 2**. a) A work flow on computing arousal by fusin intra- (within) and inter- (cross) modalities. b) A depiction on the evaluation setup of experiment I, II and experiment III, where experiment I, II assess the correlation of the framework with the average human raters, and experiment III assesses the correlation between the framework and individual rater

*2.3.1. Intra-modality Arousal Scoring*

Within modality arousal scoring framework is largely based on the methodology used by Bone et al. [13]. We compute arousal score for each feature within individual modality by first acquiring *neutral* samples to compute a baseline model, $N_i$, for each feature type $i$ (see section 2.2) for a given speaker in the database. Then, for a feature type $i$ in $j^{th}$ utterance, $x_{ij}$, we compute an unsupervised (rule-based) arousal scoring $p_{ij}$ as follows:

$$p_{ij} = 2 \times E[x_{ij} > N_i] - 1 \quad (1)$$

where $E[x_{ij} > N_i]$ is the percentage of neutral model ($N_i$) for which $x_{ij}$ is larger.

This creates a continuous-scaled arousal score bounded between [-1, 1] for each individual feature, and a combined arousal score within each modality, i.e., termed as $A_{\text{vocal}}$, $A_{\text{nonverbal}}$, is computed by summing individual features within each modality.

*2.3.2. Inter-modality Arousal Scoring Fusion*

In this work, instead of the common supervised fusion techniques, we utilize three different unsupervised fusion methods to fuse $A_{\text{vocal}}$, $A_{\text{nonverbal}}$. The unsupervised techniques are desired since the objective is to design a readily-applicable arousal quantification framework.

The three techniques are inspired from literature in information retrieval community, where different queries may result in different documents returned due to different scores generated; hence an unsupervised technique in fusion all likely returned documents is needed. The first two methods are "score-based' fusion techniques:

1. **CombSUM**: summation of arousal scores
2. **CombMAX**: taking the maximum of arousal scores

The third method is a method called reciprocal rank (e.g., [17, 18], which uses inverse of the ranking as scores that avoid the problems of skimming effect and score normalization. Since arousal scoring method itself is essentially a ranking-based method, the method is a natural fit to perform unsupervised fusion. The method is denoted as **CombRR**, where we compute the final score, $RR(i)$, by using the following formula:

$$RR(i) = \sum_{k=1}^{2} \frac{1}{R_k(i)}, \quad (2)$$

where $R_k(i)$ is the ranking of vocal and nonvocal arousal score respectively.

## 3. EXPERIMENTAL SETUP AND RESULTS

We conduct the following three experiments:

- **Experiment I**: Analyses of non-verbal unsupervised arousal scores with the average humans' ratings
- **Experiment II**: Unsupervised fusion of $A_{\text{vocal}}$ and $A_{\text{nonverbal}}$ for multimodal arousal rating
- **Experiment III**: Analyses of signal-based arousal rating with individual evaluator's subjective judgment

The depiction of Experiment I, II and III is also shown in Figure 2 (right portion). The first experiment aims at understanding whether there exists significant relationship between unsupervised facial/head signal-derived arousal scores and human perceptual arousal rating. The second experiment's goal is to form a better improved arousal rating score than a vocal-only arousal rating through multimodal fusion. Lastly, the third experiment demonstrates that the proposed multimodal arousal scoring has a higher agreement with each individual evaluators' agreement than the evaluators themselves. All correlation values are done using Spearman correlation.

### 3.1. Experiment I & II results and discussions

Table 1 describes summary results from these two experiments. First of all, while arousal has mostly been associated with behaviors in the vocal modality, we still observe

**Table 1**. *Summary of experiment I, II results (measured by Spearman correlation, all entries have p-value* $< 10^{-5}$

| **Type** | MOU | LBRO | RBRO | LLID | RLID | LHD | RHD | $\text{Head}_{\text{M}}$ | $\text{Head}_{\text{vM}}$ | $\text{Pitch}_{\text{M}}$ | $\text{Pitch}_{\text{vM}}$ | $\text{Roll}_{\text{M}}$ | $\text{Roll}_{\text{vM}}$ | $\text{Yaw}_{\text{M}}$ | $\text{Yaw}_{\text{vM}}$ | Voice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Corr. | 0.36 | 0.39 | 0.34 | 0.31 | 0.29 | **0.47** | 0.41 | 0.38 | 0.42 | 0.42 | 0.40 | 0.44 | 0.44 | 0.41 | 0.44 | **0.66** |

| | sum of all 15 non-verbal scores ($A_{\text{nonAll}}$) | sum of the 4 highest non-verbal arousal scores |
|---|---|---|
| Corr. | **0.51** | 0.48 |

| | $A_{\text{nonAll}}$ fused w/ $A_{\text{voice}}$ | | | $A_{\text{optimized}}$ fused w/ $A_{\text{voice}}$ | | |
|---|---|---|---|---|---|---|
| Fusion | CombSUM | CombMAX | CombRR | CombSUM | CombMAX | CombRR |
| Corr. | 0.66 | 0.64 | 0.64 | **0.68** | 0.65 | 0.65 |

that the movement activities of head or face each correlates well, i.e., in the range of 0.3 - 0.4, with the average human labels of emotion activation (the first row of table 1 refers to the arousal score generated by each individual non-verbal feature using equation 1). Furthermore, a summation of all 15 individual features' arousal rating results in a higher correlation, i.e., 0.51, than any individual non-verbal behaviors; it further depicts that the technical effectiveness of combining all features within each modality to generate a robust within-modality arousal score.

In order to construct a multimodal arousal rating system, we fuse the two modalities using the three techniques described in section 2.3.2. Results are presented in Table 1. First thing to note is that a direct fusion of $A_{\text{vocal}}, A_{\text{nonverbal}}$, where $A_{\text{nonverball}}$ includes scores from all 15 features, does not show an improvement over vocal-only arousal score. The vocal-only score alone is a very robust arousal indicator, and the constitution of this vocal score relies only on three very basic and robust features that overlap with much of the information presented in non-verbal channel. Hence, instead of summation all 15 features from non-verbal scores, we construct a $A_{\text{optimized}}$, that consists of arousal scores only derived from MOU, LBRO, LLID, LHD, $\text{Head}_{\text{M}}$, and $\text{Roll}_{\text{vM}}$ features. This constitution provides the best final resulting multimodal arousal rating framework (by using combSUM fusion method). It achieves a 0.68 correlation - higher than the vocal based arousal score. The result demonstrates that indeed there is still complementary information about arousal states in vocal and non-verbal communicative channels.

Lastly, since the best final unsupervised fusion framework, i.e.,combSUM, is a summation framework, we further conduct an experiment on learning the summation weight by using linear regression. The weights for each modality for one speaker are learned using the rest of nine speakers in the database. This supervised fusion framework results in correlation of 0.67, i.e., lower than the method of combSUM. This experiment further reinforces that robustness achieved in our proposed multimodal arousal rating framework is very competitive and comparable.

### 3.2. Experiment III result and discussions

We set up experiment III to demonstrate that the use of signal-derived quantification scheme is an viable approach to provide a more consistence and more objective way of measuring arousal than judging by human evaluators. Instead of usual approach where we compute correlation of the proposed framework with respect to the average human evaluator's assessment, we compute correlations between our proposed framework to individual evaluator for a given subset of data samples in the database. The results are listed below:

**Table 2**. *Summary of experiment III results*

| | Evaluator 1 | Evaluator 2 | $A_{\text{vocal}}$ | $A_{\text{multi}}$ |
|---|---|---|---|---|
| Evaluator 1 | 1.0 | 0.44 | 0.48 | 0.53 |
| Evaluator 2 | 0.44 | 1.0 | 0.64 | 0.64 |

We observe that given the same samples (in this case, 1930 samples), the agreement between evaluator 1 and evaluator 2 is 0.44. However, our proposed arousal rating (note: imaging the framework is just as another *evaluator* since the rating is unsupervised with respect to these human evaluations) correlates 0.53 and 0.64 to each evaluator, respectively; it is a number that's much higher comparing to the number between the two humans. This experimental result implicates that there seems to be a higher variation in terms of human subjective evaluations on emotion than the "machine's signal-based" evaluation. Furthermore, when comparing between vocal arousal rating and multimodal arousal rating, it is interesting to observe that within these two people, there is already a discrepancy. Non-verbal behaviors provide additional information for evaluator 1 in terms of assessing arousal state but do not do so for evaluator 2.

## 4. CONCLUSIONS AND FUTURE WORK

In this work, we present an extension on unsupervised vocal arousal by incorporating multimodal behavior features, i.e., minimal set of descriptors involve movements in the face and head. By utilizing six features only, we obtain an improved correlation compared to vocal-based arousal rating using robust score summation. There are many threads of future works, on the engineering side, the first is to extend this framework to multi-corpora and implement the non-verbal features in terms of computer vision techniques. Second, to strengthen the results in experiment III, we plan to collect a large number of evaluators using crowd-sourcing methodologies to systematically study the subjective phenomenon of emotion evaluation. A computational framework that is readily-applicable to use can not only provide quantitative measure to the emotion of interest (arousal), but also could provide novel insights that were not accessible without availability of a consistent and objective of behavioral computing.

## 5. REFERENCES


[1] Rosalind W Picard, *Affective computing*, MIT press, 2000.

[2] Jonathan Gratch and Stacy Marsella, "Tears and fears: Modeling emotions and emotional behaviors in synthetic agents," in *Proceedings of the fifth international conference on Autonomous agents*. ACM, 2001, pp. 278–285.

[3] Chul Min Lee and Shrikanth S Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[4] K Dmello Sidney, Scotty D Craig, Barry Gholson, Stan Franklin, Rosalind Picard, and Arthur C Graesser, "Integrating affect sensors in an intelligent tutoring system," in *Affective Interactions: The Computer in the Affective Loop Workshop at*, pp. 7–13.

[5] David Watson, Lee A Clark, and Auke Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales," *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063, 1988.

[6] Charles S Carver and Teri L White, "Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the bis/bas scales.," *Journal of personality and social psychology*, vol. 67, no. 2, pp. 319, 1994.

[7] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language: Computational techniques are presented to analyze and model expressed and perceived human behavior-variedly characterized as typical, atypical, distressed, and disordered-from speech and language cues and their applications in health, commerce, education, and beyond," *Proc IEEE Inst Electr Electron Eng*, vol. 101, no. 5, pp. 1203–1233, 2013, Narayanan, Shrikanth Georgiou, Panayiotis G ENG R01 AA018673/AA/NIAAA NIH HHS/ 2013/09/17 06:00 Proc IEEE Inst Electr Electron Eng. 2013 Feb 7;101(5):1203-1233.

[8] P. A. Andersen and J. F. Andersen, "The exchange of nonverbal intimacy - a critical-review of dyadic models," *Journal of Nonverbal Behavior*, vol. 8, no. 4, pp. 327–349, 1984, Agn80 Times Cited:17 Cited References Count:77.

[9] Maria Hernandez-Reif, Tiffany Field, Miguel Diego, and Maxine Ruddock, "Greater arousal and less attentiveness to face/voice stimuli by neonates of¡ i¿ depressed¡/i¿ mothers on the brazelton neonatal behavioral assessment scale," *Infant Behavior and Development*, vol. 29, no. 4, pp. 594–598, 2006.

[10] Chi-Chun Lee, Athanasios Katsamanis, Matthew P Black, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan, "Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[11] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[12] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[13] Daniel Bone, C Lee, and Shrikanth Narayanan, "Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features," 2014.

[14] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, and Shrikanth Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. of the Int'l Conf. on Multimodal Interfaces*.

[15] Loic Kessous, Ginevra Castellano, and George Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 33–48, 2010.

[16] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008, 403XR Times Cited:36 Cited References Count:55.

[17] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 758–759.

[18] Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma, Zhe Jiang, Yijiang Jin, Yiqun Liu, Le Zhao, and S Ma, "Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments," *NIST SPECIAL PUBLICATION SP*, , no. 251, pp. 586–590, 2003.