Speech Emotion Recognition Using Spectral Entropy

Woo-Seok Lee, Yong-Wan Roh, Dong-Ju Kim, Jung-Hyun Kim,
and Kwang-Seok Hong

School of Information and Communication Engineering, Sungkyunkwan University, 300,
Chunchun-dong, Jangan-gu, Suwon, Kyungki-do, 440-746, Korea
grampus553@skku.edu, elec1004@skku.edu, raokdioguy@korea.com,
kjh0328@skku.edu, kshong@skku.ac.kr
http://hci.skku.ac.kr

Abstract. This paper proposes a Gaussian Mixture Model (GMM)–based speech emotion recognition methods using four feature parameters; 1) Fast Fourier Transform(FFT) spectral entropy, 2) delta FFT spectral entropy, 3) Mel-frequency Filter Bank (MFB) spectral entropy, 4) delta MFB spectral entropy. In addition, we use four emotions in a speech database including anger, sadness, happiness, and neutrality. We perform speech emotion recognition experiments using each pre-defined emotion and gender. The experimental results show that the proposed emotion recognition using FFT spectral-based entropy and MFB spectral-based entropy performs better than existing emotion recognition based on GMM using energy, Zero Crossing Rate (ZCR), Linear Prediction Coefficient (LPC), and pitch parameters.

1 Introduction

The ability to recognize, interpret and express emotions - commonly referred to as "Emotional Intelligence" [1] - plays a key role in human communication. Computers with emotion recognition and expression skills will allow a more natural and improved human computer interaction. An emotion-recognizing computer can "learn" during an interaction by associating emotional expressions (like pleasure or displeasure) with its own behavior, as a kind of reward and punishment [2]. In recent years, research that analyzed, human emotional information has been conducted, particularly in the fields of emotional speech, facial expressions, and body gestures etc. There are still few papers that address the problem of recognizing emotional states or human emotions contained in speech information, however and most of the papers are based on mathematical classification methods or pattern recognition techniques [3]. The most commonly analyzed aspects of emotion in speech are energy, tempo, duration, jitter, shimmer, LPC, Mel Frequency Cepstrum Coefficient (MFCC), pitch, and so on. The features that are most prominent in emotion recognition are pitch and energy [4], [5]. A probability model algorithm is used to classify speech and to derive the status of speaker's emotions. Generally, the emotion classifications are neutrality, anger, sadness, happiness, surprise and disgust [4].

This paper proposes to perform emotion recognition using speech information. Speech is able to deliver not only the meaning of speech but also information about

C. Xiong et al. (Eds.): ICIRA 2008, Part II, LNAI 5315, pp. 45–54, 2008.

emotions which is evident in aspects of speaker's speech that are able to be classified. The emotional speech database that is used in this paper consists of neutrality, anger, sadness, and happiness. The analytical methods for emotion recognition are FFT spectral entropy, delta FFT spectral entropy, MFB spectral entropy, delta MFB spectral entropy. We use a GMM algorithm as a pattern recognition algorithm for recognizing emotion in speech. The rest of this paper is organized as follows. Section 2 describes the previous variable emotion recognition methods. The proposed method is described in Section 3. Section 4 shows the experiment-based performance evaluation of the proposed method. Finally, conclusions are provided in Section 5.

2 Related Works

In this section, we describe work that has been done in emotion recognition. Kyung Hak Hyun [6] used non-zero-pitch for speech emotion recognition. Pitch information is quite useful in speech emotion recognition [6], and non-zero-pitch is the pitch contour that does not have a zero value. Because the zero value of the pitch causes some errors in the Gaussian distribution, it must be eliminated. Although the non-zero-pitch contour loses some information content such as unvoiced sounds, the emotion recognition result is improved. They have applied this concept to a Bayesian classifier and they obtained better results for emotion recognition than those obtained using the previous pitch contour.

Kwon et al. [7] used a CHMM (Continuous Hidden Markov Model) with a left-right topology and up to five states to model on the word level including a neutral style and three stress styles. Their paper showed an average emotion recognition rate of 70.1%, which was superior to the performance of a support vector machine (SVM) classifier that had a recognition rate of 67.1% [8]. For a second database containing short commands or greetings with several words using five basic emotions, an average recognition rate of 40.8% was achieved, but this time it was inferior to the SVM classifier's rate of 42.3%.

Jian Zhou [9] proposed a speech emotion recognition method based on rough set (RS) theory and SNM. Rough Set (RS) is a valid mathematical theory to deal with imprecise, uncertain, and vague information. They used energy, pitch contour, and the first, second, and third formants as features for speech emotion recognition. The mean emotion recognition rate is 74.75%.

3 A Proposal for Speech Emotion Recognition

In this paper, we used spectral entropy and MFB spectral entropy to analyze the features for emotion recognition. Spectral entropy calculates each frame, and uses a combination of delta spectral entropy. MFB spectral entropy uses 27 filter banks. In this section, we first describe the compilation flow and the block diagram of the emotion recognition analysis.

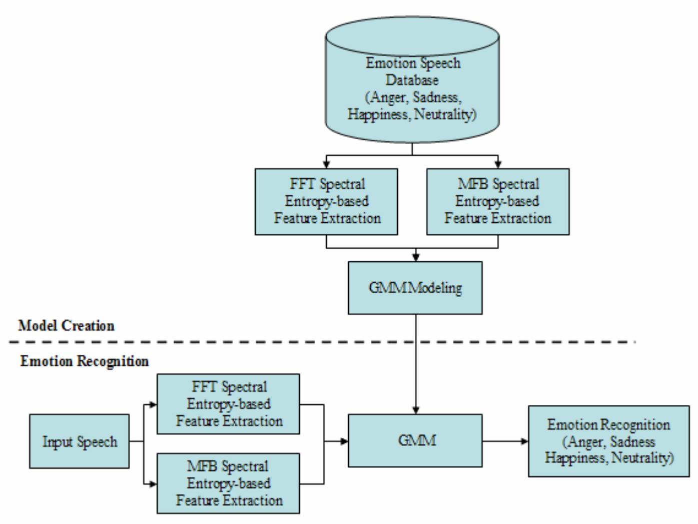

Fig. 1. Block Diagram of the Emotion Recognition Analysis

3.1 The Block Diagram of the Emotion Recognition Analysis

The emotion recognition system block diagram is shown in Fig. 1. In creating the model, we use the emotion speech database to perform FFT and MFB spectral entropy-based feature extraction using FFT spectral entropy, delta FFT spectral entropy, MFB spectral entropy, and delta MFB spectral entropy. The probabilistic model parameters are based on using the GMM. In recognizing emotions, we extract the FFT and MFB spectral entropy-based features and then the emotion recognition system compares the extracted parameters with the probabilistic models of the four emotions.

3.2 FFT Spectral Entropy

In this paper, we use FFT spectral entropy and delta FFT spectral entropy. Entropy is based on Shannon's information theory and it is the means of measuring the amount of information. In FFT spectral entropy, entropy theory is applied to speech emotion recognition.

The speech is first pre-emphasized using a pre-emphasis filter in order to spectrally flatten the signal, and then the pre-emphasized speech is separated into short segments called frames. This will significantly enhance the ability to identify the emotional aspects of speech in this paper. In order to reduce the familiar edge effects, a 512-point Hamming window is applied to each frame. The FFT of X(i,n) is given by equation (1).

$$X(i,n) = \sum_{m=1}^{M} x(m,n) e^{-j\frac{2\pi}{N}im} \tag{1}$$

Where X(i,n) is the nth frame and the ith frequency component. M denotes the number of points in the FFT. x(m,n) is the nth frame the mth sample. After the FFT block, the spectrum of each frame is filtered by a set of filters, and the power of each band is calculated. The power of each band S(i,n) is calculated using equation (2).

$$S(i,n) = | X(i,n) |^2 \tag{2}$$

The probability density of the spectrum is estimated by using a method that normalizes the frequency components. The FFT normalization P[S(i,n)] is defined in equation (3).

$$P[S(i,n)] = \frac{S(i,n)}{\sum_{m=1}^{M/2} S(m,n)} \tag{3}$$

The method of calculating the spectral entropy H(n) is shown in equation (4).

$$H(n) = -\sum_{i=1}^{M/2} P[S(i,n)] \log_2 P[S(i,n)] \tag{4}$$

The delta FFT power spectrum of each calculation subtracts the nth frame and the ith frequency component from the n+1th frame and the ith frequency component. The delta FFT power spectrum S΄(i,n) is defined in equation (5).

$$S'(i,n) = S(i,n+1) - S(i,n) \tag{5}$$

We calculated the delta FFT power spectrum by taking the logarithm and the modulus, and then we calculated the spectrum normalization. P[S΄(i,n)] is the delta FFT normalization, which is defined in equation (6). In addition, the proposed delta spectral entropy H΄(n) is shown in equation (7).

$$P[S'(i,n)] = \frac{| S'(i,n) |}{\sum_{m=1}^{M/2} | S'(m,n) |} \tag{6}$$

$$H'(n) = -\sum_{i=1}^{M/2} P[S'(i,n)] \log P[S'(i,n)] \tag{7}$$

3.3 MFB Spectral Entropy

In this paper, we use MFB spectral entropy and delta MFB spectral entropy. The speech is first pre-emphasized using a pre-emphasis filter "1-αZ^{-1}" in order to spectrally flatten the signal, where "α" is in the range 0.9 ~ 1. The default value of "α" is 0.97 [10]. Then the pre-emphasized speech is separated into short segments called frames. The frame length is set to 32ms (512 sample) to guarantee that the signal inside the frame is stationary. There is an 16ms (256 sample) overlap between two adjacent frames to ensure that the signals from one frame to the next are stationary. A frame can be seen as the result of multiplying the speech waveform by a rectangular pulse whose width is equal to the frame length. In order to reduce the edge effects, a 512-point Hamming window is applied to each frame. After the FFT block, the spectrum of each frame is filtered by a set of filters, and the power of each band is calculated. Each signal frame is processed through a Mel-scale filter bank resulting in a vector that contains 27 energy coefficients.

The MFB multiplies the FFT by the Mel-scale filter. The MFB describes M(b,n), which is shown in equation (8).

$$M(b,n) = \frac{\sum_{i=L_b}^{U_b} V_b(i)S(i,n)}{\sum_{i=L_b}^{U_b} V_b(i)} \qquad (8)$$

where, M(b,n) is obtained by multiplying S(i,n) by the frequency response of the i^{th} Mel-scale filter as$V_b(i)$. L_b and U_b are the start frequency and the end frequency of the i^{th} Mel filter. The method of calculating the spectral entropy H(n) is shown in equation (9).

$$H_{MFB}(n) = -\sum_{b=1}^{B} P[M(b,n)]\log_2 P[M(b,n)] \qquad (9)$$

The proposed delta MFB is M´(b,n), which is defined by subtracting the n^{th} frame and the b^{th} Mel-frequency filter bank by the $n+1^{th}$ frame and the b^{th} Mel-frequency filter bank. The delta FFT power spectrum S´(i,n) is defined in equation (5). We calculated the delta MFB power spectrum by taking the logarithm and the modulus, and then we calculated the normalized delta MFB spectrum P[M´(b,n)] that is defined in equation (10).

$$P[M'(b,n)] = \frac{|M'(b,n)|}{\sum_{m=1}^{B} |M'(m,n)|} \qquad (10)$$

where, b is the number of Mel-filters that is used in this paper. The proposed delta MFB spectral entropy is shown in equation (11).

$$H'_{MFB}(n) = -\sum_{b=1}^{B} P[M'(b,n)]\log P[M'(b,n)] \qquad (11)$$

3.4 Emotion Recognition Using GMM

In this paper, we used a GMM to estimate the probability distribution instead of assuming that each class distribution is a normal distribution. A GMM uses various Gaussian distributions to model discrete probability distributions, so it can be optimized using the Expectation Maximization (EM) algorithm [11]. The modeled probability distribution using the GMM is defined in equation (12).

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(x) \qquad (12)$$

where, b_i is the GPDF (Gaussian probability density function), and p_i is the mixture weight. We need u_i, Σ_i and p_i to express the Gaussian mixture density, where u_i is the mean vector and Σ_i is the covariance matrix. It is a set of three parameters that express the probability model for a Gaussian distribution of the speaker's emotion. This set is defined by the GMM, and shown in equation (13).

$$\lambda = \left\{ \rho_i,\ \mu_{i,}\ \sum_i \right\} \qquad i = 1,2,3,\ldots,M \tag{13}$$

All emotional features are modeled using the GMM with diagonal covariance matrices measured over all frames of a speech.

First, using all of the training data, a root GMM is trained with the EM(Expectation Maximization) algorithm with a maximum likelihood criterion, and then one GMM per class is adapted from the root model using the MAP(Maximum a Posteriori) criterion. MAP adaptation protects against overtraining and removes the need to optimize the number of Gaussians per class which may be necessary due to differences in the amount of available training data per class [12]. We use 32 Gaussians for spectral-based entropy and MFB spectral-based entropy.

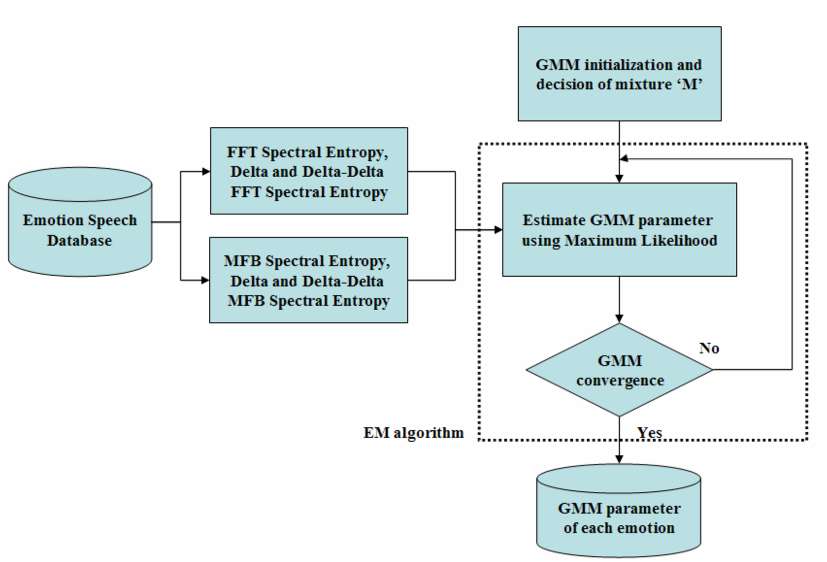
GMM initialization and decision of mixture 'M'
FFT Spectral Entropy, Delta and Delta-Delta FFT Spectral Entropy
Emotion Speech Database
MFB Spectral Entropy, Delta and Delta-Delta MFB Spectral Entropy
Estimate GMM parameter using Maximum Likelihood
GMM convergence
No
Yes
EM algorithm
GMM parameter of each emotion

Fig. 2. Block diagram of the process of training the GMM parameters

Figure 2 shows emotion recognition using the GMM. In the training step, we estimate the GMM parameters including the maximum Gaussian in each emotional data set using EM algorithm.

4 Experiments and Results

4.1 Emotional Speech Databases

Given that in many languages the fundamental tendencies of sounds are expressed in similar ways, our results in recognizing the emotions of Korean language speakers can generally be applied to speakers of other languages. For this reason, we used a database produced by Professor C. Y. Lee from Media and the Communication Signal Processing Laboratory of Yonsei University in Korea with the support of the Korea Research Institute of Standards and Science. This data covers the four emotions of neutrality, happiness, sadness, and anger; and its principles are as follows [13]:

- easy pronunciation in anger, sadness, happiness and neutrality
- 45 dialogic sentences that express natural emotion

The original data is stored in the form of 16kHz. To use the data in MATLAB, we transformed the data for the training and simulation into a 16bit format through quantization with a pulse code modulation filter.

There are fifteen male and fifteen female emotional speech samples in the database. In this paper, 30 subjects listened to the utterances of one speaker played back in random order.

4.2 Speech Emotion Recognition Using the Existing Methods

The table 1 shows that the result of the emotion recognizing experiment using general features. The general features define energy, log-energy, ZCR, LPC, and pitch. The existing method use combination of general features. The judgment currently uses four emotions. The experiment used four classifications by dividing the group into men and women, and then dividing each of these groups according to whether or not the speaker participated in the training.

Table 1. Emotion recognition results: (a) General features; (b) General and delta general features

	Participate in training		Not participate in training	
	Male	Female	Male	Female
Anger	78.0%	79.3%	74.3%	76.3%
Sadness	71.3%	75.0%	61.6%	74.6%
Happiness	42.0%	58.6%	40.3%	54.0%
Neutrality	51.0%	60.3%	47.6%	52.3%
Average	60.6%	68.3%	55.9%	64.3%

(a)

	Participate in training		Not participate in training	
	Male	Female	Male	Female
Anger	80.3%	82.6%	74.3%	77.3%
Sadness	77.0%	78.0%	75.7%	76.3%
Happiness	53.3%	56.0%	50.3%	57.3%
Neutrality	51.6%	51.6%	46.0%	49.3%
Average	64.5%	67.1%	61.6%	65.1%

(b)

Table 1(a) is the result of the first experiment that used general features. Table 1(b) is the result of the second experiment that used general features and delta general features. In this experiment, the emotion recognition result using general features is 62.3%, and the emotion recognition result using general and delta general features is 64.6%. The experiment of emotion recognition using general feature showed a maximum recognize rate when using a combination of general features and delta general features.

4.3 Emotion Recognition Using the Proposed Method

In experimental environments, we used 12,600 sentences as training data, and we performed the context-independent speech emotion recognition using experimental data consisting of 3,600 sentences. The sentences used in the experiments were 200 randomly selected sentences for each emotion. The experiments consist of two methods of using FFTs for spectral entropy based emotion recognition. Table 2(a) shows the experimental results using the FFT spectral entropy based emotion recognition method. Table 2(b) shows the experimental results using the FFT spectral entropy method and the delta FFT spectral entropy based emotion recognition method. In the experimental results, the emotion recognition result using FFT spectral entropy is 65%, and the emotion recognition result using FFT spectral entropy and delta FFT spectral entropy is 72.75%. The experiment demonstrates that the emotion recognition using FFT spectral entropy, and delta FFT spectral entropy show the maximum recognize rate. The experiments consist of two parts for the MFB spectral entropy based emotion recognition methods.

Table 2. Emotion recognition results: (a) FFT spectral entropy; (b) FFT and delta FFT spectral entropy

	Participate in training		Not participate in training	
	Male	Female	Male	Female
Anger	78.0%	68.0%	72.0%	66.0%
Sadness	80.0%	74.0%	86.0%	78.0%
Happiness	52.0%	58..0%	46.0%	66.0%
Neutrality	60.0%	62.0%	52.0%	54.0%
Average	67.5%	65.5%	64.0%	66.0%

(a)

	Participate in training		Not participate in training	
	Male	Female	Male	Female
Anger	82.0%	84.0%	84.0%	78.0%
Sadness	78.0%	72.0%	44.0%	76.0%
Happiness	64.0%	52.0%	48.0%	54.0%
Neutrality	62.0%	58.0%	66.0%	50.0%
Average	71.5%	66.5%	60.5%	64.5%

(b)

Table 3. Emotion recognition results: (a) MFB spectral entropy; (b) MFB and delta MFB spectral entropy

	Participate in training		Not participate in training	
	Male	Female	Male	Female
Anger	78.0%	66.0%	74.0%	62.0%
Sadness	76.0%	84.0%	76.0%	86.0%
Happiness	54.0%	58..0%	42.0%	54.0%
Neutrality	68.0%	72.0%	72.0%	70.0%
Average	69.0%	70.0%	66.0%	68.0%

(a)

	Participate in training		Not participate in training	
	Male	Female	Male	Female
Anger	92.0%	80.0%	84.0%	78.0%
Sadness	72.0%	92.0%	72.0%	96.0%
Happiness	56.0%	58.0%	54.0%	52.0%
Neutrality	86.0%	74.0%	82.0%	74.0%
Average	76.5%	76.0%	73.0%	75.0%

(b)

Table 3(a) shows the experimental results using the MFB spectral entropy based emotion recognition method. Table 3(b) shows the experimental results using the MFB spectral entropy and delta MFB spectral entropy based emotion recognition method.

In the experimental results, the emotion recognition result using MFB spectral entropy is 68.25%, and the emotion recognition result using MFB spectral entropy and delta MFB spectral entropy is 75.13%. The experiment demonstrates that emotion recognition using MFB spectral entropy and delta MFB spectral entropy show the maximum recognize rate.

5 Conclusions

This paper proposed human emotion recognition using FFT and MFB spectral-based entropy. For emotion recognition, the GMM was used to design the emotion probabilistic model. An emotion speech database was used that consists of four emotions: anger, sadness, happiness and neutrality. We trained and tested the emotion recognition using the collected FFT and MFB spectral-based entropy from the emotion speech database. In experiments using the FFT spectral entropy-based method, we attained a maximum recognition rate of 65.8% when we used FFT spectral entropy and delta FFT spectral entropy. In experiments using the MFB spectral entropy-based method, we attained a maximum recognition rate of 75.1% when we used MFB spectral entropy and delta MFB spectral entropy. We also experimented using the existing method, and we experimented using the proposed emotion recognition method, and then we compared the results of the experiment using the existing method with the result of the experiment using the proposed method. We attained a maximum recognition rate of 64.6% using general features, such as energy, ZCR, LPC, and pitch and delta using these general features. From the experimental results, the proposed FFT spectral entropy-based method improves the recognizing rate 1.2%, and the proposed MFB spectral entropy-based method improves the recognizing rate 10.5% compared to the emotion recognition method that uses general features.

Acknowledgment

This research was supported by MKE, Korea under ITRC IITA-2008-(C1090-0801-0046) and the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST)(No. R01-2008-000-10642-0).

References

1. Goleman, D.: Emotional Intelligence. Bantam Books, New York (1995)
2. Borchert, M., Dusterhoft, A.: Emotion in Speech-Experiments with Prosody and Quality Features in Speech for Use in Categorical and Dimensional Emotion Recognition Environments, Natural Language Processing and Knowledge Engineering. In: Proceedings of 2005 IEEE International Conference on IEEE NLP-KE 2005, 30 October-1 November (2005)
3. Kim, S.-i., Lee, S.-h., Shin, W.-j., Park, N.-c.: Recognition of Emotional states in Speech using Hidden Markov Model. In: Proceeding of KFIS Fall Conference, vol. 14(2) (2004)
4. Zhao, L., Cao, Y., Wang, Z., Zou, C.: Speech Emotional Recognition Using Global and Time Sequence Structure Features with MMD. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784. Springer, Heidelberg (2005)
5. Schuller, B., Rigoll, G., Lang, M.: Hidden Markov Model-based speech emotion recognition. In: Proc. ICASSP, HongKong, China, pp. 401–404 (2003)
6. Hyun, K.H., Kim, E.H., Kwak, Y.K.: Improvement of Emotion Recognition by Bayesian Classifier Using Non-zero-pitch Cencept, Robot and Human Interactive Communication. In: IEEE International Workshop on ROMAN 2005, 13–15 August (2005)

7. Kwon, O.-W., Chan, K.-L., Hao, J., Lee, T.-W.: Emotion Recognition by Speech Signals. In: Eurospeech, Geneva, Switzerland (2003)
8. Wagner, J., Vogt, T., Andre, E.: A Systematic Comparison of Different HMM Design for Emotion Recognition from Acted and Spontaneous Speech. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 114–125. Springer, Heidelberg (2007)
9. Zhou, J., Wang, G., Yang, Y., Chen, P.: Speech Emotion Recognition Based on Rough Set and SVM, Cognitive Informatics. In: 5th IEEE International Conference on ICCI 2006, July 17-19, 2006, vol. 1, pp. 53–61 (2006)
10. Young-Wan, R., Hong, K.-S.: Delta FBLC based Speech/Non-Speech Frame Decision in Real Car Environment. In: The 4th Conference on New Exploratory Technologies (Next 2007)
11. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning - data mining, inference, and prediction. Springer, Heidelberg (2000)
12. Reynolds, D., Quatieri, T., Dunn., R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)
13. Hyun, K.H., Kim, E.H., Kwak, Y.K.: Improvement of emotion recognition by Bayesian classifier using non-zero-pitch concept, Robot and Human Interactive Communication. In: IEEE International Workshop on ROMAN 2005, August 13-15, 2005, pp. 312–316 (2005)