SPEECH EMOTION RECOGNITION WITH ACOUSTIC AND LEXICAL FEATURES

Qio Jio[2-3]- Ci fohxio Li[2]- Si izif Ci fo[2]- Hvin io X v[2]*

[1] Computer Science Department, School of Information, Renmin University of China,
[2] Key Lab of Data Engineering and Knowledge Engineering of Ministry of Education,
Renmin University of China, Beijing 100872
{qjin, 2011202429, cszhe1, whmliu}@ruc.edu.cn

ABSTRACT

In this paper we explore one of the key aspects in building an emotion recognition system: generating suitable feature representations. We generate feature representations from both acoustic and lexical levels. At the acoustic level, we first extract low-level features such as intensity, F0, jitter, shimmer and spectral contours etc. We then generate different acoustic feature representations based on these low-level features, including statistics over these features, a new representation derived from a set of low-level acoustic codewords, and a new representation from Gaussian Supervectors. At the lexical level, we propose a new feature representation named emotion vector (eVector). We also use the traditional Bag-of-Words (BoW) feature. We apply these feature representations for emotion recognition and compare their performance on the USC-IEMOCAP database. We also combine these different feature representations via early fusion and late fusion. Our experimental results show that late fusion of both acoustic and lexical features achieves four-class emotion recognition accuracy of 69.2%.

Index Terms—Emotion recognition, Acoustic features, Emotion lexicon, Lexical features, Support vector machine

1. INTRODUCTION

Automatic emotion recognition from speech has been an active research area in past years, which is of great interest for human-computer interactions. It has wide applications ranging from computer tutoring applications [1] to mental health diagnostic applications [2-3].

Accuracy of automatic speech emotion recognition systems depends largely on two key factors: the choice of features and the choice of classifiers. Various classification techniques including Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), K-nearest neighbors (KNN), artificial neural networks (ANN), and support vector machines (SVM) etc. have been explored in the emotion recognition literature [4-8]. In these classifiers, SVM has been shown to provide a better generalization performance in different pattern recognition problems than traditional techniques [8]. In this paper, we use SVM as the classifier for speech emotion classification.

Acoustic features have been used as the dominant features in the speech emotion recognition literature. A number of acoustic features have been explored and used with SVM classifiers in different studies, including prosodic features (such as statistics of pitch, energy, duration and higher order formants etc. [9]), spectral features (such as spectrum centroid, spectrum cut-off frequency, correlation density and mel-frequency energy etc. [8]), and cepstral features (such as Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), and Perceptual Linear Prediction (PLP) etc. [10]). These features explored in the aforementioned different studies are related to the acoustic characteristics of frequency, energy, and spectral intensity. Most of them are frame-level features. Derivative features based on frame-level features using various statistical functions (such as mean, standard deviation, range etc.) are also commonly applied in emotion recognition systems [5]. Features that have been successfully used in speaker verification tasks such as i-vector features have also been used for emotion recognition tasks [11]. Recently, features learnt from deep neural networks (DNN) have shown great power in speech applications [12]. Linguistic features have also been applied in speech emotion recognition tasks, including various bag-of-words representations and n-grams [13-14] etc. Most of the work is based on reference transcripts, which we also use in this paper.

In this paper we study four-class (angry, happy, sad, neutral) utterance-level emotion recognition on a large emotion database, the University of Southern California's Interactive Emotional Motion Capture (USC-IEMOCAP) database [15]. We focus our efforts on the feature representation part. We extract both acoustic and lexical features. At the acoustic level, we first extract low-level features such as intensity, F0, jitter, shimmer and spectral contours etc. We then generate different acoustic feature representations based on these features, including statistics over these features, a new representation derived from a set of low-level acoustic codewords, and a new representation from Gaussian supervectors. At the lexical level, in addition to the traditional bag-of-words (BoW) features, we propose a new feature representation based on emotion lexicons. We name this new lexical feature representation as emotion vector (eVector). We apply these feature representations for emotion recognition and compare their performance on the USC-IEMOCAP database. We also investigate different ways to combine different features including late fusion at the score level and early fusion at the feature level.

The remainder of this paper is organized as follows. Section 2 briefly introduces the USC_IEMOCAP database. Section 3 describes the different feature representations including acoustic and lexical features. Section 4 presents the experimental results. Section 5 concludes the paper and describes future work.

2. DATABASE DESCRIPTION

*Corresponding Author

In this paper we conduct experiments on the USC-IEMOCAP: University of Southern California (USC) Interactive emotional dyadic motion capture database [15]. USC-IEMOCAP is an acted, multimodal and multi-speaker database and contains approximately 12 hours of audio-visual data, including video, speech, motion capture of face, text transcriptions. 10 professional actors (5 male, 5 female) perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions. The speech sessions are manually segmented into utterances, and each utterance is annotated by at least three annotators into both categorical labels such as angry, happy, etc. and dimensional labels such as valence, activation and dominance. Our work only utilizes the categorical annotations and considers all utterances that have a majority voted ground truth. In order to balance data of different class labels, we merge the happiness and excitement categories as the happy category. Together with angry, sad, and neutral, we form a four-class emotion classification dataset. Table 1 presents the total number of utterances and durations in every emotion class. There are in total 5531 utterances. Every emotion class contains an approximately equal number of sample utterances.

Table 1: Number of utterances and durations per emotion class

Emotion	Anger	Happy	Sad	Neutral	Total
# Utterances	1103	1636	1084	1708	**5531**
Duration (min)	83.0	126.0	99.3	111.1	**419.4**

3. FEATURE REPRESENTATIONS

3.1 Acoustic Feature Representations

We first extract low-level acoustic features at a frame-level on each utterance and then generate different feature representations via applying operations on part of or the entire feature set. We utilize the openSMILE [16] toolkit for feature extraction and the configuration file is modified according to the configuration file "emobase2010.conf" based on the Interspeech 2010 Paralinguistic Challenge [17], which locates in the config directory of the openSMILE toolkit. The features extracted from openSMILE are grouped into three categories: continuous features, qualitative features and cepstral-based features (as shown in Figure 1). Both pitch and voice quality features are extracted with 40-ms frame window and 10-ms frame shift. Cepstral-based features are extracted with 25-ms frame window and 10-ms frame shift.

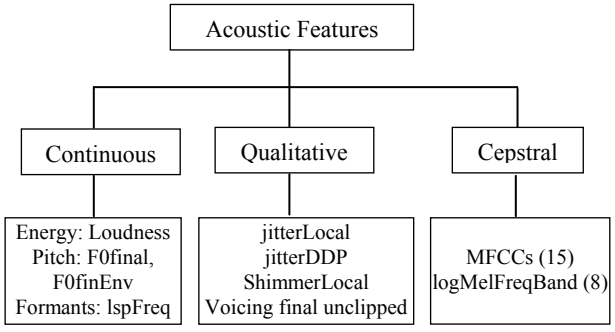

Figure 1: Categories of low-level acoustic features

3.2.2 Statistics pgBdpvtuid Gfawsft

We apply 21 commonly used statistical functions on aforementioned low-level frame-based features to turn a time series of variable length into a static feature vector. Functions include extremes like maximum and minimum position in frames, means, regressions, moments, percentiles and durations. Detailed description of the statistical functions can be found in studies [17].

3.2.3 Cah.pgx pset pgBdpvtuid Gfawsft

The codebook model is a common technique used in the document classification (bag-of-words) and image classification (bag-of-visual-words). The bag-of-audio-words technique has also been applied in multimedia event detection tasks [18]. An audio segment can be represented by features related to a histogram of codewords' counts via transforming low-level acoustic features based on codewords in the vocabulary (codebook). In this paper, we represent the emotion class in terms of a codebook or a bag-of-acoustic-words model. We generate the acoustic codebook via unsupervised clustering. Each utterance is then represented as a distribution over the codewords by using soft-assignment of low level acoustic features to these codewords. In our experiments, we set the codebook size as 4096 and generate codewords for low-level cepstral features using K-means clustering.

3.2.3 Havttiao Svpfswfdps pgBdpvtuid Gfawsft

In recent years, Gaussian supervectors (GSV) have proven to be extremely successful for a speaker verification task [19]. A Gaussian supervector is constructed by stacking the means or diagonal covariance or Gaussian weights of the adapted mixture components. We first train a Gaussian mixture universal background model (UBM-GMM) based on a pool of randomly selected speech data from different emotions.

$$g(X) = \sum_{i=1}^{M} \lambda_i N(X; U_i, \Sigma_i), \quad (1)$$

where λ_i are the mixture weights, $N(X; V_i, \Sigma_i)$ is a Gaussian, and V_i and Σ_i are the mean and covariance of the Gaussians, respectively. We assume the covariance matrix Σ is diagonal. For each emotion utterance, a GMM is built via MAP adaptation on the UBM-GMM using the features extracted on this utterance. In our setup, the relevant factor during MAP adaptation is 10. From this adapted GMM, we can form a Gaussian supervector by concatenating the means V_i or the diagonal covariance Σ_i or weights λ_i. The Gaussian supervector can be thought of as a transformation between an utterance and a high-dimensional vector. We then use the supervectors as input features to train the SVM emotion classifier. In this paper, we generate the GSV via stacking the means.

3.2 Lexical Feature Representations

3.3.2 Cah.pgx pset LfxidamGfawsft

The bag-of-words (BoW) model is widely used in text processing applications. It processes texts without considering the word order, the semantic structure or the grammar. The vector representation of BoW is a normally used feature representation for text document. The vocabulary is commonly selected using term-frequency-inverse document frequency (TF-IDF) theory. In this paper, we first perform stop-words removal and Porter stemming [20] on the transcripts and then use TF-IDF to select Top-K words. According to our experience on text classification, we find that a vocabulary size around 2000 is relatively appropriate. We therefore select around top 600 words from each of the four emotion classes respectively and merge them to form our basic word vocabulary of size 2000. The BoW feature representation can take different vector values. The first kind of vector representation consists of only 0 or 1 value for each dimension, where 1 stands

for the occurrence of a vocabulary word and 0 stands for non-occurrence. The second kind of vector representation consists of a real number for each dimension representing the frequency of one vocabulary word in a document. In this paper, we use the first type of BoW feature representation.

*3.3.3 Fn puipo Wfdups)f Wfdups*Lfxidam Gfawsft*

The vocabulary in the above BoW feature representation only collects frequent words that appeared in the texts. It cannot infer any quantitative correlation of each word to a certain emotion class. Inspired by the work in [21] which used mutual information framework to identify emotional words, we propose a new feature representation. We first construct a new lexicon for each emotion class in which each word has a weight indicating its inclination for expressing this emotion.

We observe that in texts, we can roughly categorize different words into three types. Taking the sentences with "anger" emotion as an example, the first type of words is "Emotional word" like "hell", which is a typical frequent word for expressing anger emotion. The second type of words is "Common word" like "real", which commonly appears in documents but does not usually contain emotion inclination. The third type of words is "Unemotional and Uncommon word" like "marry". Based on our intuition, we expect the three types of words may have the following distribution patterns as shown in Table 2, where N_{io} refers to the occurrence counts of a word in documents from a certain emotion class, N_{pvu} refers to the occurrence counts of this word in documents from other emotions, N_C refers to the number of emotion classes that this word appears in. We then use the following formula to compute the weight of every word and rank them in descending order:

$$weight = \frac{N_{in}}{\left(N_{out} * N_C + 1\right)} \quad (2)$$

Table 2: Expected distribution pattern of three types of words

Type	N_{in}	N_{out}	N_C
Emotional words	more	less	less
Common words	fair	more	more
Unemotional & uncommon words	less	less	less

We expect "Emotional words" should be ranked in the top, "Common words" should be ranked in the bottom, and "Unemotional & Uncommon words" should be ranked in the middle. Some examples of such rankings for the emotion class "anger" are shown in Table 3. The result shows that our intuition is relatively correct. We then build the emotion lexicon for each emotion category by computing the word weights as in (2). Unlike the vocabulary in BoW feature representation, the new emotion lexicon not only collects words that appeared in one emotion class but also assigns a weight indicating its inclination for expressing this emotion. We then use this emotion lexicon to generate a vector feature representation for each utterance. We name such a feature vector as emotion vector (eVector). eVector is in 4 dimensions instead of hundreds of dimensions for BoW feature vectors. The eVector is in the format as follows:

$$eVector = \left(d_1, d_2, d_3, d_4\right) \quad (3)$$

$$d_i = \frac{1}{K} \sum_{k=1, word_k \in utt}^{K} weight(word_k) \quad (4)$$

where the four dimensions correspond to the four emotion classes of anger, happy, sad and neutral respectively. The value of each dimension is the average of all the words' weights in an utterance according to the emotion lexicon for emotion class *i*.

Table 3: Word examples in ranked list for anger class

"anger" emotion class	Word	Weight
The top part	BEAST	23.0
	HELL	16.0
	SNAP	14.0
	SHUT	13.0
	CRUEL	9.0
The middle part	ICE	0.3333
	MARRY	0.3292
	BUNCH	0.2857
	BURDEN	0.2857
	DEAL	0.2727
The bottom part	REAL	0.0172
	FUN	0.0149
	CAR	0.0131
	MOON	0.0117
	KIND	0.0112

4. EXPERIMENTS

4.1 Experimental Setup

All the experiments in this paper are conducted in a 10-fold leave-one-speaker out cross-validation scheme. The description of the terminologies used in the following experimental results is as follows:

- ACO: the utterance-level statistics of frame-level acoustic features as described in section 3.1.1 (excluding cepstral features, consisting of continuous and qualitative features as shown in Figure 1).
- Cepstrum: the utterance-level statistics of frame-level cepstral features.
- Cepstral-BoW: the bag-of-words feature representation based on frame-level cepstral features.
- GSV-mean: Gaussian supervectors generated by concatenating the MAP adapted GMM means. In this paper, we build the GMM with 128 mixture components.
- +: early fusion at the feature level (equals to feature concatenation operation). ACO+Cepstrum refers to the new fused features by concatenating ACO and Cepstrum features.
- (+): late fusion at the score level. ACO(+)Cepstrum refers to the fusion of the classification scores from the ACO-based system and the Cepstrum-based system.

The classifier is one of the most important components in emotion recognition systems. Various types of classifiers have been used in speech emotion recognition systems. The support vector machines have been recognized as one of the most effective classifiers in many applications and considered to be easier to use than neural networks [22]. We use the SVM as our emotion classifier with a linear kernel in this paper. We compare the performance with each type of features (single stream) in the SVM classifier and explore combining different types of features via early fusion.

4.2. Experimental Results

Table 4 shows the weighted emotion classification accuracies with different types of single stream features ranked in descending order. The weighted accuracy is a weighted mean accuracy over different emotion classes with weights proportional to the number of utterances in a particular emotion class. For acoustic features, we only show the top 4 best performed features. From the results we can see that the lexical features extracted from the transcripts achieve better performance than acoustic features under the single stream feature condition. The classification accuracy with eVector lexical feature representation of only 4 dimensions (57.44%) outperforms the accuracy with BoW lexical feature representation of 2000 dimensions (56%). Table 5 shows the top 5 pair-wise early fusion of acoustic features. Early fusion of cepstrum and GSV-mean features achieve best performance.

Table 4: Weighted emotion recognition accuracies with different single stream feature representations

Feature	Accuracy
Lex-eVector	57.4%
Lex-BoW	56.0%
Cepstrum	53.5%
ACO	52.5%
GSV-mean	51.8%
Cepstral-BoW	49.0%

Table 5: Top 5 weighted emotion recognition accuracies via pair-wise early fusion of acoustic features

Feature	Accuracy	Relative improv.
Cepstrum+GSV-mean	55.4%	4.1%
Cepstrum+Cepstral-BoW	55.1%	3.4%
Cepstral-BoW+GSV-mean	55.0%	6.5%
ACO+Cepstral-BoW	54.2%	3.7%
ACO+GSV-mean	53.7%	2.7%

Table 6: Weighted emotion recognition accuracies via pair-wise late fusion of acoustic feature and lexical feature systems

Feature	Accuracy	Relative improv.
Lex-BoW(+)Lex-eVector	58.5%	2.4%
Lex-BoW(+)Cepstrum	63.1%	16.0%
Lex-BoW(+)ACO	62.1%	13.9%
Lex-BoW(+)GSV-mean	60.7%	10.7%
Lex-BoW(+)Cepstral-BoW	59.4%	7.6%
Lex-eVector(+)Cepstrum	63.9%	15.0%
Lex-eVector(+)ACO	63.9%	15.2%
Lex-eVector(+)GSV-mean	63.1%	13.4%
Lex-eVector(+)Cepstral-BoW	61.6%	9.7%

We then investigate pair-wise late fusion between the lexical and acoustic feature based systems at the classification score level. From Table 6 we can see that fusing the two lexical feature based systems achieves additional improvement. Fusion of the lexical feature based system with the acoustic feature based system achieves more gain. The fusion weights are learnt on held-out development data and applied to the test data. The weights are tuned per emotion class. The weights assigned to the acoustic feature based system are between 0.4 and 0.6 across different emotion classes. Table 7 presents the breakdown classification accuracies for each emotion class for the systems shown in Table 6.

Table 7: Per-class emotion recognition accuracies via pair-wise late fusion of acoustic feature and lexical feature systems

Feature	anger	happy	sad	neutral
Lex-BoW(+)Lex-eVector	57.6%	60.8%	42.0%	67.7%
Lex-BoW(+)Cepstrum	65.3%	65.0%	53.4%	66.0%
Lex-BoW(+)ACO	60.9%	65.1%	56.4%	63.4%
Lex-BoW(+)GSV-mean	63.1%	66.8%	44.0%	64.6%
Lex-BoW(+)Cepstral-BoW	58.9%	63.5%	53.4%	59.6%
Lex-eVector(+)Cepstrum	65.4%	65.6%	54.3%	67.4%
Lex-eVector(+)ACO	62.2%	66.2%	58.1%	66.6%
Lex-eVector(+)GSV-mean	65.5%	66.1%	50.2%	67.3%
Lex-eVector(+)Cepstral-BoW	62.4%	65.5%	54.9%	61.6%

We also conduct fusion among all acoustic and lexical features with permutation of all possible combinations. The system based on early fusion of Cepstral-BoW and GSV-mean acoustic features combined with ACO-based system, Cepstrum-based system, Lex-BoW-based system, and Lex-eVector-based system through late fusion achieves the best weighted emotion recognition accuracy of 69.2%. Table 8 shows the breakdown performance on each emotion class for this fused system.

Table 8: Per-class emotion recognition accuracies of the best fusion system

Feature	anger	happy	sad	neutral
ACO(+)Cepstrum(+) Cepstral-BoW+GSV-mean (+)Lex-BoW(+)Lex-eVector	70.4%	70.3%	61.7%	72.2%

5. CONCLUSIONS

In this paper we generate feature representations from both acoustic and lexical levels for emotion recognition. At the acoustic level, we first extract low-level features such as intensity, F0, jitter, shimmer and spectral contours etc. We then generate different acoustic feature representations based on these features, including statistics over these features, a new representation derived from a set of low-level acoustic codewords, and a new representation from Gaussian supervectors. At the lexical level, in addition to the traditional bag-of-words feature representation, we propose a new lexical feature representation named emotion vector (eVector). This new feature representation relies on the emotion lexicons which not only contain the vocabulary for a specific emotion but also assign weights to words showing their inclination of expressing certain emotions. We apply these different feature representations for emotion recognition and compare their performance on the USC-IEMOCAP database. We also combine these different feature representations via early fusion and late fusion. Our experimental results show that late fusion of both acoustic and lexical features achieves four-class emotion recognition accuracy of 69.2%.

ACKNOWLEDGEMENTS

We would like to thank SAIL-USC for providing us access to the USC-IEMOCAP database. This work is supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), and the Beijing Natural Science Foundation (No. 4142029).

6. REFERENCES

[1] D. Litman and K. Forbes, "Recognizing emotions from student speech in tutoring dialogues." IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2003.

[2] D.J. France, R.G. Shiavi, S. Silverman, M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Trans. Biomedical Eng. 47(7) (2000) pp. 829-837.

[3] N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, W. Heinzelman, M. Sturge-Applie, "Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion", proceedings of the 4th IEEE workshop on Spoken Language Technology (SLT), Miami, Florida, December 2012.

[4] B. Schuller, G. Rigoll, M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", proceedings of the ICASSP, vol. 1, 2004, pp. 577-580.

[5] M. Ayadi, M. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, 44(2011) 572-587.

[6] Z. Zeng, M. Pantic, G. I. Rosiman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," Trans. on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[7] M. Kockmann, L. Burget, J. Cemocky, "Application of speaker and language independent state-of-the-art techniques for emotion recognition", in Speech Communication, Vol. 53, No. 9, pp. 1172-1185, ISSN 0167-6393, 2011.

[8] L. Chen, X. Mao, Y-L.. Xue, L.L. Cheng, "Speech Emotion Recognition: Features and Classification Models", Digital Signal Processing 22(6): 1154-1160, 2012.

[9] B. Schuller, S. Reiter, R. Mueller, M. Al-Hames, M. Lang, G. Rigoll, "Speaker-independent speech emotion recognition by ensemble classification", in Proc. ICME 2005, Amsterdam, Netherlands, 2005.

[10] T.L. Pao, Y.T. Chen, J.H. Ye, P.L. Li, "Mandarin Emotional Speech Recognition based on SVM and NN", in Proc. International Conference on Patter Recognition, vol. 1, pp. 1096-1100, 2006.

[11] R. Xia and Y. Liu, "Using i-Vector Space Model for Emotion Recognition", In INTERSPEECH, 2012.

[12] H. Lee, Y. Largman, P. Pham, and A.Y. Ng. "Unsupervised feature learning for audio classification using convolutional deep belief networks". Advances in Neural Information Processing Systems (NIPS) 22, 2009.

[13] B. Schuller, A. Batliner, S. Steidl, D. Seppi, "Emotion recognition from speech: putting ASR in the loop", in Proc. ICASSP 2009.

[14] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, A. Vembu, R. Prasad, "Emotion Recognition using Acoustic and Lexical Features", INTERSPEECH. 2012.

[15] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.

[16] F. Eyben, M. Wollmer, B. Schuller, "OpenSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), Florence, Italy, pp. 1459-1462, 2010.

[17] B. Schuller, A. Batliner, S. Steidl, D. Seppi, "Recognizing Realistic Emotions and Affect in Speech: State of the Art and Lessons Leant from the First Challenge", Speech Communication, 53(10), pp. 1062-1087, 2011.

[18] K. Lee, and D.P.W. Ellis, "Audio-Based Semantic Concept Classification for Consumer Video", IEEE Trans. Audio, Speech, and Language Processing, 18(6):1406-1416, 2010.

[19] W.M. Campbell, D.E. Sturim and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification", IEEE Signal Processing Letters, 2006, pp 308-311.

[20] M.F. Porter, "An algorithm for suffix stripping", *Qsphsan* , 14(3) pp 130−137, 1980.

[21] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs", Speech and Audio Processing, IEEE Transactions on, 13(2), 293-303. Chicago, 2005.

[22] C.W. Hsu, C.C. Chang, C.J. Lin, "A practical guide to support vector classification", Technical Report, Deparment of Computer Science, National Taiwan University, 2003.