Using Neutral Speech Models for Emotional Speech Analysis

Carlos Busso[1], *Sungbok Lee*[1,2], *Shrikanth S. Narayanan*[1,2]

Speech Analysis and Interpretation Laboratory (SAIL)
[1]Electrical Engineering Department, [2]Department of Linguistics
University of Southern California, Los Angeles, CA 90089
busso@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Since emotional speech can be regarded as a variation on neutral (non-emotional) speech, it is expected that a robust neutral speech model can be useful in contrasting different emotions expressed in speech. This study explores this idea by creating acoustic models trained with spectral features, using the emotionally-neutral TIMIT corpus. The performance is tested with two emotional speech databases: one recorded with a microphone (acted), and another recorded from a telephone application (spontaneous). It is found that accuracy up to 78% and 65% can be achieved in the binary and category emotion discriminations, respectively. Raw Mel Filter Bank (MFB) output was found to perform better than conventional MFCC, with both broad-band and telephone-band speech. These results suggest that well-trained neutral acoustic models can be effectively used as a front-end for emotion recognition, and once trained with MFB, it may reasonably work well regardless of the channel characteristics.

Index Terms: Emotion recognition, Neutral speech, HMMs, Mel filter bank (MFB), TIMIT

1. Introduction

Detecting and utilizing non-lexical or paralinguistic cues from a user is one of the major challenges in the development of usable *human-machine interfaces* (HMI). Notable among these cues are the universal categorical emotional states (e.g., angry, happy, sad, etc.), prevalent in day-to-day scenarios. Knowing such emotional states can help adjust system responses so that the user of such a system can be more engaged and have a more effective interaction with the system.

For the aforementioned purpose, identifying a user's emotion from the speech signal is quite desirable since recording the stream of data and extracting features from this modality is comparatively easier and simpler than in other modalities such as facial expression and body posture. Previous studies on automatic categorization of emotional speech have shown accuracy between 50% and 85% depending on the task (e.g. number of emotion labels, number of speakers, size of database) [1]. A comprehensive review of the current approaches is given in [2].

However, such emotion categorization performance is largely specific to individual databases examined (and usually off-line) and it is not plausible to easily generalize the results to different databases or on-line recognition tasks. This is due to inherent speaker-dependency in emotion expression, acoustic confusions among emotional categories, and differences in acoustic environments across recording sesions. It is also fairly difficult, if not infeasible, to collect enough emotional speech data so that one can train robust and universal acoustic models of individual emotions, especially, if one considers that there exist more than dozen of emotional categories and their possible combinations that we can use to differentiate affective states or attitudes [3].

As a possible way to circumvent the fundamental problem in emotion categorization based on speech acoustics, this study tests a novel idea of discriminating emotional speech against neutral (i.e., non-emotional) speech. That is, instead of training individual emotional models, we build a single, neutral speech model and use it for emotion evaluation either in the categorical approach or in the dimensional approach [3] based on the assumption that emotional speech productions are variants of the non-emotional counterparts in the (measurable) feature space. For example, it has been shown that speech rate, speech duration, fundamental frequency (F0), and RMS energy are simultaneously modulated to convey the emotional information [4]. Also in the articulatory domain, it has been shown that the tongue tip, jaw and lip kinematics during expressive speech production are different from neutral speech [5, 6]. Hence, modeling the differential properties with respect to neutral speech is hypothesized to be advantageous. In addition, because there are a lot more neutral speech corpora, robust neutral acoustic speech models can be built. This paper presents our first attempt to examine the aforementioned idea.

In this preliminary report, the TIMIT database is used to train neutral acoustic models and two emotional speech databases are probed. *Hidden Markov Models* (HMMs) are trained with two different acoustic feature sets, *Mel Filter Bank* (MFB) and *Mel-Frequency Cepstrum Coefficients* (MFCCs), and their behaviors are examined in a broad phonetic-class recognition experiment setting based on recognition likelihood scores. Emotional discrimination performance by the two feature sets are also investigated and compared using a discriminant analysis. The results show that using only these features, accuracies up to 78% can be achieved for the binary emotion recognition task.

This paper is organized as follows. Section 2 describes the proposed approach. Section 3 analyzes the likelihood scores from emotional and non-emotional speech obtained with the neutral models. Section 4 provides a discriminant analysis of the likelihood scores. Finally, Section 5 gives a discussion and future direction of this work.

2. Methodology

The proposed approach to segment emotional speech has two steps. In the first step, neutral models are built to measure the degree of similarity between the input speech and the reference neutral speech. The output of this block is a *fitness measure* of the input speech. In the second step, these measures are used as features to infer whether the input speech is emotional or neutral. The primary focus of this paper is placed on the first block.

While the neutral models can be trained for any speech feature that shows emotional modulation, in this paper, we considered the conventional spectral features MFCCs and MFB outputs (prior work had demonstrated that spectral features carry significant emotional information [7]). Separate set of models for both types of features were built at the phonetic-class level. The English phonetic alphabet was aggregated in seven broad phonetic classes that share similar articulation configuration (Table 1).

HMMs were selected to build the neutral models, since they

Table 1: Broad phone classes

	Description	Phonemes
F	Front vowels	iy ih eh ae ix
B	Mid/back vowels	ax ah axh ax-h uw uh ao aa ux
D	Diphthong	ey ay oy aw ow
L	Liquid and glide	l el r y w er axr
N	Nasal	m n en ng em nx eng
T	Stop	b d dx g p t k pcl tcl kcl qcl bcl dcl gcl q epi
C	Fricatives	ch j jh dh z zh v f th s sh hh hv
S	Silence	sil h# #h pau

are suitable to capture the time series behavior of the speech, as widely demonstrated in *automatic speech recognition* (ASR). Here, an HMM of 3 states and 16 Mixtures of Gaussians was built for each broad phonetic category. The HTK toolkit was used to build these models, using standard techniques such as forward-backward and Baum-Welch re-estimation algorithm [8]. After a high-frequency pre-emphasis of the speech signal, MFCC and MFB feature vectors were estimated. For MFCCs, 13 coefficients were estimated with cepstral mean normalization (CMN) option. Likewise, the outputs of 13 filter banks were used as MFB features. In both cases, the velocity and acceleration of these coefficients were included forming a 39-feature vector.

An important issue in this approach is the selection of the *fitness measure* to assess how well the input speech fit the reference models. Here, the likelihood scores are used, which are provided by the Viterbi decoding algorithm. Since this likelihood depends on the length of the segment, the acoustic scores were normalized according to the duration of the phones (option $-o\ N$ in function *HVite*).

One assumption made in this approach is that the emotional corpus used to test this approach will have a set of neutral speech. The purpose of this assumption is to compensate different recording settings between the neutral reference and emotional databases. In particular, the speech files are scaled such that the average RMS energy of the neutral reference database and the neutral set in the emotional database are the same. This approach is similar to the normalization presented in [9].

3. Analysis of the likelihood scores

The popular read-speech TIMIT database was used as a reference for neutral speech [10]. This database contains 4620 sentences for the training set, and 1680 for the testing set, collected from 460 speakers. The nature, size, and the inter-speaker variability make this database suitable to train the proposed "neutral-speech" models. Two corpora are used as emotional databases. The first one is the EMA database [5], in which three subjects read 10 sentences five times portraying four emotional states: sadness, anger, happiness and neutral state. Although this database contains articulatory information, only the acoustic signals were analyzed. Notice that the EMA data have been perceptually evaluated and inter-evaluator agreement has been shown to be 81.9% [11]. The second corpus was collected from a call center application [12]. It provides spontaneous speech of different speakers from a real human-machine application. The data was labeled as negative or non-negative (neutral). Only the sentences with high agreement between the raters were considered (1027 neutral, 338 negative). This database is referred here on as *call center database* (CCD). More details of these two emotional databases can be found in [5, 12], respectively.

While the sample rate of the TIMIT database is 16KHz, the sample rate of the CCD corpus is 8KHz (telephone speech). To compensate this mismatch, the TIMIT database was downsampled to 8KHz to train the reference neutral models used to assess the CCD corpus. In contrast, for the EMA corpus the broad band TIMIT data was used for training, since its speech files were also recorded at 16KHz.

3.1. MFB-based neutral speech models

After the MFB-based neutral models were built with the TIMIT training set, the likelihood scores of the emotional testing corpora were computed. Figures 1 and 2 shows plots with the mean and standard deviation of the likelihood scores for the broad phonetic categories, obtained with the EMA and CCD databases, respectively. For reference, the likelihood scores for the TIMIT testing set were also plotted.

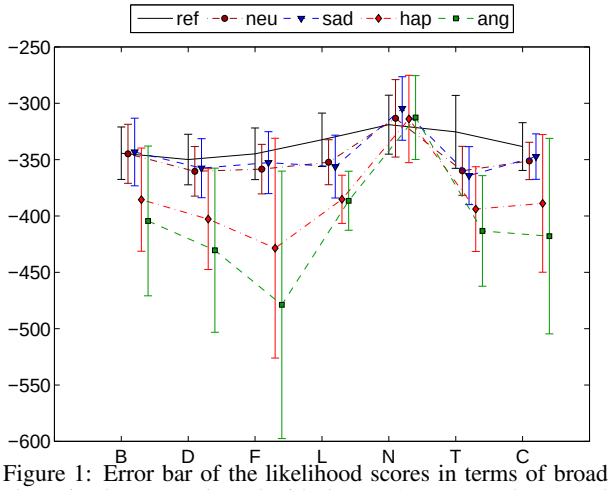


Figure 1: Error bar of the likelihood scores in terms of broad phonetic classes, evaluated with the EMA corpus. The neutral models were trained with MFB features.

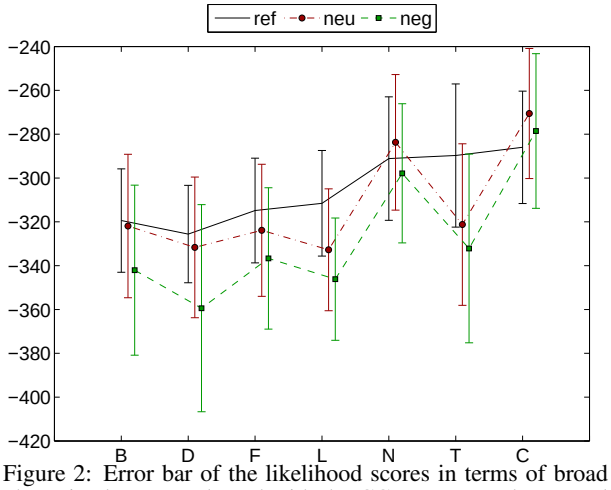
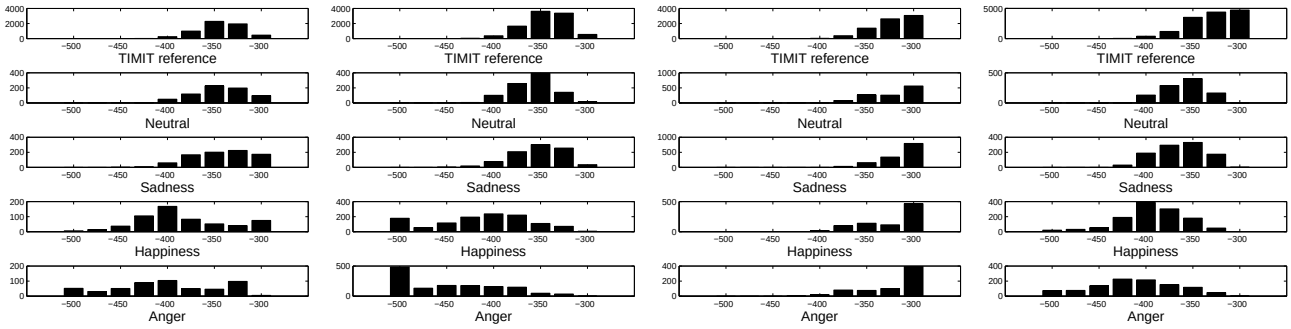


Figure 2: Error bar of the likelihood scores in terms of broad phonetic classes, evaluated with the CCD corpus. The neutral models were trained with MFB features.

These figures reveal that the mean and the variance of the likelihood score for emotional speech differ from the results observed in neutral speech, especially for emotion with high level of arousal such as anger and happiness. We also observed that some broad phonetic classes present stronger differences than others. For example, front vowels present distinctive emotional modulations. In contrast, the likelihood scores for nasal sounds are similar across emotional category (see Fig. 3-c), suggesting that during articulation there is not enough degrees of freedom to convey emotional modulation. These results agree with our previous work that indicated that emotional modulation is not displayed uniformly across speech sounds [4, 5].

Figure 3 gives the histograms of the likelihood scores of four broad phonetic classes for the EMA database. For reference, the likelihood scores for the TIMIT testing database are also included. This figure reveals that the results of the neutral speech from the EMA corpus are closer to the results of



(a) Class B (b) Class F (c) Class N (d) Class T

Figure 3: Likelihood score histograms for the broad phonetic classes B, F, N and T. EMA corpus is used (MFB-based neutral models).

the TIMIT references. It is also observed that the histograms for happiness and anger significantly differ from that of the references. Unfortunately, the results for sadness were similar to neutral speech, so it is expected that these classes will not be correctly separated with these chosen spectral features. Interestingly, similar confusion trends were observed in our previous work between these emotional categories with other acoustic speech features (spectral and prosodic features) [4, 11]. One possible explanation is that neutral and sad speech mainly differ in the *valence* domain. However, it has been shown that with speech features the *valence* domain is more difficult to recognize than the *arousal* or *activation* domain [11].

3.2. MFCC-based neutral speech models

Neutral models were also built with MFCC features. Figures 4 and 5 show the results of the likelihood score in terms of the broad phonetic categories for the EMA and CCD databases, respectively. Although emotional modulation is observed, these figures show that the differences between the likelihood scores for emotional categories are not as strong as the differences obtained with the MFB-based models (note the values in the vertical axes). Interestingly, MFCCs are calculated from MFB, by applying the *Discrete Cosine Transform* (DCT) over the Mel log-amplitudes. This post-processing step seems to blur the acoustic differences between emotional and neutral speech.

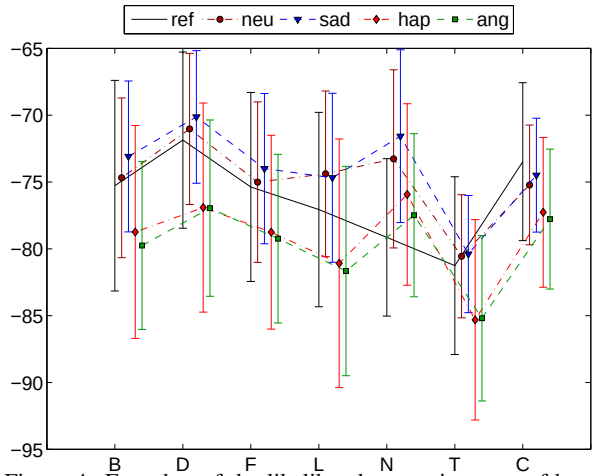


Figure 4: Error bar of the likelihood scores in terms of broad phonetic classes, evaluated with the EMA corpus. The neutral models were trained with MFCC features.

4. Discriminant analysis

This section analyzes the emotional information conveyed in the likelihood scores. It also discusses whether they can be used to segment emotional speech automatically.

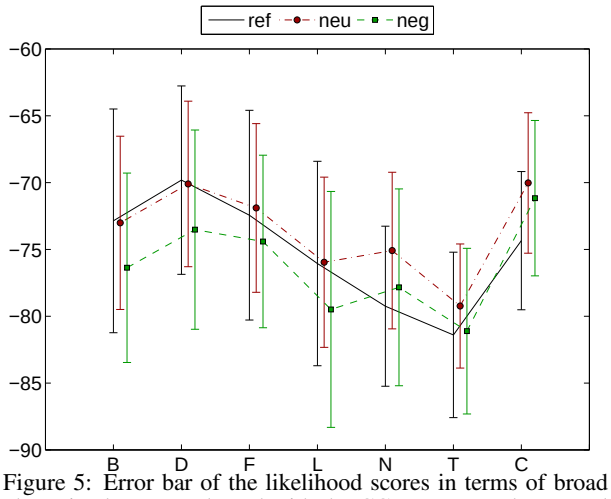


Figure 5: Error bar of the likelihood scores in terms of broad phonetic classes, evaluated with the CCD corpus. The neutral models were trained with MFCC features.

It can be observed from Figures 1, 2, 4, and 5 that the means and the standard deviations of the likelihood scores differ from the values obtained with neutral speech. In this experiment, the average of these measures at sentence level for each broad phonetic class was used as features for emotion recognition. If the emotional classes are denoted by C and the features for the phonetic class $i_j \in (F, B, D, L, N, T, C)$, are denoted by Lh_{i_j}, the classification problem can be formulated as:

$$P(C|Obs) \tag{1}$$

$$P(C|Lk_{i_1}, Lk_{i_2}, \cdots Lk_{i_N}) \tag{2}$$

where N is the number of different phonetic classes recognized on the sentence. Assuming independence between the results of the phonetic classes, Equation 2 can be rewritten as:

$$P(C|Obs) = \prod_{j=1}^{N} P(C|Lk_{i_j}) \tag{3}$$

In other words, only the probabilities $P(C|Lk_{i_j})$ from the phonetic classes detected in the sentences are combined. A linear discriminant classifier (LDC) was used in these experiments. Since the number of samples for each emotional category is different, the *prior* probabilities were set to equal values. The databases were randomly split in training (80%) and testing (20%) sets. The results reported here correspond to the average performance over 100 realizations.

Table 2 presents the recognition results for two experiments using the EMA database: binary emotion recognition, in

Table 2: Discriminant analysis of likelihood scores for EMA database (*Neu*=neutral, *Sad*=sadness, *Hap*= happiness, *Ang*= anger, *Emo*=emotional)

		Ground truth							
		MFB-based models				MFCC-based models			
		Sad	Ang	Hap	Neu	Sad	Ang	Hap	Neu
Classified	Emo	0.20	0.98	0.97	0.02	0.05	0.99	0.81	0.13
	Neu	0.80	0.02	0.03	0.98	0.95	0.01	0.19	0.87
	Sad	0.66	0.00	0.10	0.27	0.65	0.00	0.02	0.25
	Ang	0.00	0.73	0.32	0.01	0.01	0.65	0.44	0.03
	Hap	0.02	0.27	0.58	0.06	0.04	0.35	0.46	0.22
	Neu	0.33	0.00	0.01	0.66	0.30	0.00	0.07	0.50

which the emotional classes sadness, anger and happiness were grouped together versus neutral speech, and categorical emotion recognition, in which the labels of the four categories were classified. For the MFB-based neutral models, the binary classifier achieved an accuracy of 78% (chance is 50%). As can be observed from the confusion matrix, the classification errors were mainly between neutral state and sadness. The performance of the 4-label emotion recognition test was 65% (chance is 25%). In our previous work, an accuracy of 66.9% was achieved for a similar task by using many acoustic features [11]. These two experiments reveal that with this approach neutral speech can be accurately separated from emotional speech when there is a high level of arousal (i.e. happiness and anger). For the MFCC-based neutral models, the performance decreases measurably. These results agree with the analysis presented in Section 3.

For the CCD corpus, the emotional classes considered in the experiment were negative versus neutral speech. The results for the MFB-based neutral models were 42% for the negative class, and 84% for neutral class (63% is the average). The performance for MFCC-based models was slightly lower than with MFB: 38% for the negative class, and 85% for neutral class (61.5% is the average). As a reference, our previous work has reached approximately 74% accuracy for the same task, by using many different acoustic speech features, including prosodic features [12]. One reason that the performance is worse in the CCD data than in the EMA data is that most of the sentences are short with only one word (median duration is approximately 1.5 seconds). This issue affects the accuracy of the features extracted from the likelihood scores.

5. Discussion and Conclusions

This paper presented a new approach to classifying emotional versus non-emotional speech by using neutral reference models. The results show that this approach can achieve accuracies up to 78% in the binary emotional classification task. These results suggest that well-trained neutral acoustic models can be effectively used as a front-end for emotion recognition. Interestingly, the models trained with conventional MFCCs are found to perform worse than the models with the original MFBs for both emotional databases, suggesting that MFB-based models will achieve better performance regardless of the speech characteristics.

The proposed approach can be enhanced and expanded in different directions. First, the proposed framework can be applied to prosodic features such us pitch and energy, which have been shown to convey emotional information. A challenging question is how to normalize those features to remove inter-subject, inter-gender, and inter-recording differences, preserving inter-emotional discrimination.

Although this framework addresses binary emotion classification, the proposed scheme can be used as a first step, in a more sophisticated emotion recognition system. After detecting emotional speech, a second level classifier can be used to achieve a finer emotional description of the speech.

The results indicated that some broad phonetic classes present more emotional differentiation than others. In general, vowels seem to have more degrees of freedom to convey emotional information than phonemes such as nasal sounds. These observations can be used for recognition by weighting the features extracted from the likelihood scores according to the observed emotional modulation conveyed in them.

Notice that the detected phoneme classes are the ones that maximize the likelihood in the Viterbi decoding path. If the transcript of the sentences is known, which may not be true for all real time applications, the phoneme recognition accuracy could also be used as a *fitness measure* for emotional discrimination. The hypothesis here is that the performance will be higher for neutral speech, compared to emotional speech. This approach could be extremely useful to automatically find emotional speech portions in larger scripted emotional corpora.

Figures 1, 2, 4, and 5 reveal that likelihood scores for the neutral set in the emotional corpus is close to the results for the neutral references. To reduce the mismatches between the neutral corpora even further, the acoustic models in the HMMs can be adapted to the neutral set in the emotional speech. These are some topics that we are currently pursuing.

6. Acknowledgements

This research was supported in part by funds from the NSF, and the Department of the Army.

7. References

[1] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, September 2003.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[3] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

[4] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, 2004.

[5] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 497–500.

[6] S. Lee, E. Bresch, and S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 525–532.

[7] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, 2004.

[8] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, England, December 2006.

[9] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 2, San Diego, CA, USA, June 2005, pp. 967–972.

[10] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Timit acoustic-phonetic continuous speech corpus," 1993.

[11] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. In Press, 2007.

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