RESEARCH ARTICLE OPEN ACCESS

Enhancement of Speech Recognition Algorithm Using DCT and Inverse Wave Transformation

Sukhdeep Kaur, Er. Gurwinder Kaur
Research Student, Assistant Professor
(Electronics and Communication Engineering Department Yadavindra College of Engineering (YCoE) Talwandi Sabo, District Bathinda, Punjab, India.)
Electronics and Communication Engineering Department Yadavindra College of Engineering (YCoE), Talwandi Sabo, District Bathinda, Punjab, India.)

ABSTRACT
This research work focus on providing better performance in speech recognition algorithm by integrating digital signal transposition with speech recognition techniques. This is an approach for improving the performance of speech recognition algorithm using butterworth stopband filter and discrete cosine transformation based speech compression with inverse wave transformation. The main objective is to integrate filtering with speech recognition algorithm to improve the results when noise is present in the signal. In this work, the matching is done using inverse wave transformations which reduce the time for recognition of voices. Proposed algorithm is designed and implemented in MATLAB. The proposed algorithm has been tested on the given samples and evaluated using different recognizable and unrecognizable samples obtaining a recognition ratio about 98%. It is shown that the proposed algorithm provides better results than existing techniques. Proposed algorithm increase the accuracy of the speech recognition system.
Keywords: filter, inverse wave transformation, speech recognition, wave format.

I. Introduction

The speech recognition is defined as the process of considering the spoken word as an input speech and matches it with the previously recorded speeches on basis of various parameters. This can be done by various methods. It is a process of automatically recognizing who is speaking on the basis of features of speaker of the speech signal.

Speech recognition features has some of the advantages like speech input is easy to do because it does not demand a specialized skill as does typing or push button procedures. Information can be input even when the user is not constant or doing other activities including the hands, eyes, legs or ears. Since a telephone or microphone can be used as an input terminal [1].

Basically, speaker recognition is classified in to speaker identification and speaker verification. Wide application of speech recognition system includes control access to services such as database access services banking by telephone, voice dialing telephone shopping. Now speech recognition technology is the most desirable technology to create new services.

II. Classification of Speech Recognition Systems

Most speech recognition systems can be classified according to the following categories [2]:

- **Speaker Dependent versus Speaker Independent**

A speaker-dependent speech recognition system is one that is trained to recognize the speech of only one speaker. A speaker-independent system is one that is independence is difficult to attain, as speech recognition organizations tend to become adjusted to the speakers they are trained on, resulting in error rates are higher than speaker dependent system.

- **Isolated versus Continuous**

In isolated speech the speaker pauses shortly between every word, while in continuous speech the speaker speaks in a continuous and possibly long stream with little or no breaks in between. Isolated speech recognition systems are easy to build. Words spoken in continuous speech on the other hand are subjected to the co-articulation effect, in which the pronunciation of a word is modified by the words surrounding it.

- **Keyword based versus Subword unit based**

A speech recognition system can be trained to recognize whole words, like dog or cat. This is useful in applications like voice-command-systems, in which the system need only recognize a small set of words. This approach is simple but not scalable. As the dictionary of recognized words grow, so too the complexity and execution time of the recognizer.

III. Approaches to Speech Recognition

Basically there exist three approaches to speech recognition. They are [11]:

- **Acoustic Phonetic Approach**

Acoustic-phonetic approach assumes that the phonetic units are broadly characterized by a set of features such as format frequency, voiced/unvoiced and pitch. These features are extracted from the speech signal and are used to segment and level the speech.

- **Pattern Recognition Approach**

Pattern recognition approach requires no explicit knowledge of speech. This approach has two steps – namely, training of speech patterns based on some generic spectral parameter set and recognition of patterns via pattern comparison. The popular pattern recognition techniques include template matching, Hidden Markov Model

- **Artificial Intelligence Approach**

Knowledge based approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing and finally making a decision on the measured acoustic features. Expert system is used widely in this approach.

IV. Problem Definition

In problem definition define different problems in existing approaches and how these problems will be eliminated or reduced using improved audio recognition algorithm.

- The methods developed so far fail or not give efficient results when there exist noise in the audio signal.
- Noises add too much disturbance on the given signal and which decreases the audio recognition algorithm accuracy rate.
- Noise leaves bad effects on the bandwidth of a given network so it becomes major issue in audio signals.
- Exiting networks such as neural networks take much time in training period

V. Proposed Work

In the proposed method, the goal is to detect the speaker from the previously recorded wave samples. The main concentration is on accuracy and speed. The proposed method is implemented using MatLab.

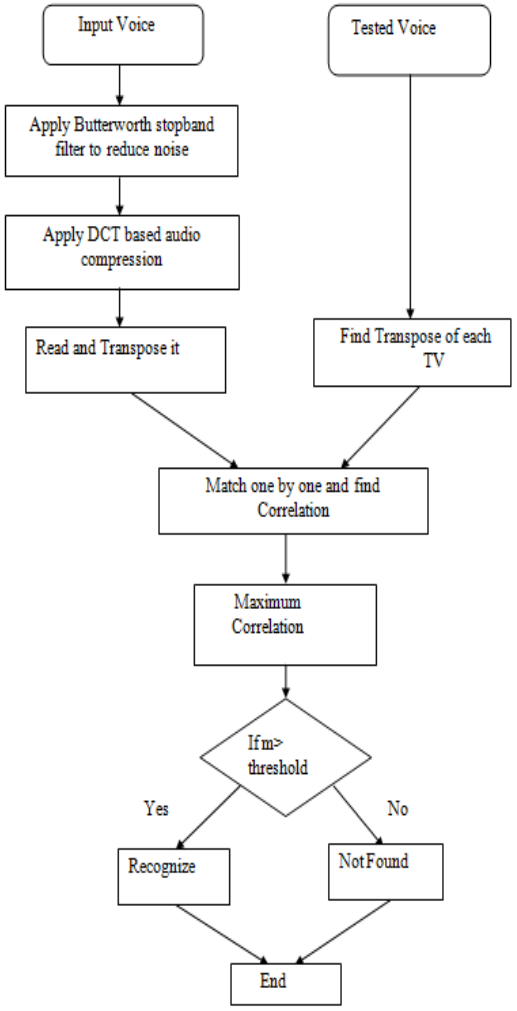

Fig.1. flow-chart for speech recognition algorithm

In the proposed algorithm samples of dummy speeches are taken for experiments and a database is prepared in which these speeches are saved. The speech samples are taken in the wave format. The input samples are taken from various persons by recording their voices. These are the tested voices TV_i for the given system. In the testing phase test voice is match with the stored input voice and recognition decisions are made.

Butterworth filter is used to remove the noise from the system. Speech signals degrade due to the presence of background noise and noise reduction is an important field of speech processing. Butterworth stopband filter is used to minimize the disturbance from the speech signals.

Discrete cosine transforms (DCT) based speech compression is used to reduce the size of the speech information. It is used to speed up the system by remove the redundancy from audio information. Compression is the process of elimination of redundancy and duplicity. The DCT is very common when encoding video and speech tracks on computers. The DCT is very similar to the DFT but the output values of DCT are real numbers and the output vector is approximately twice as long as the

DFT output. It shows a sequence of finite data points in terms of sum of cosine functions.

The 1-d discrete cosine transforms

$$y(k) = w(k) \sum_{n=1}^{N} x(n) \cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right)$$

k=1, 2,...N

(1)

Where

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \sqrt{\frac{2}{N}} & 2 \le k \le N \end{cases}$$

(2)

N is the length of x, and x and y are the same size. If x is a matrix, DCT transforms its columns. The series is indexed from $n = 1$ and $k = 1$ instead of the usual $n = 0$ and $k = 0$ because MATLAB vectors run from 1 to N instead of from 0 to N- 1. The transpose of each of the tested voices TV is taken. The transpose is taken so that the speech frames that are generally shown horizontally could be converted in vertical form so that it can be easy to understand and recognize the speech frames easily.

Take the input sample of voice that has to be tested. Read the wave input and take its transpose. The input sample taken is in wave format and it is tested among the database that contains previously recorded voice sample.

Match the input voice with tested database one by one and find correlation. Correlation computes a measure of similarity of two input signals as they are shifted by one another. The correlation result reaches a maximum at the time when the two signals match best. The correlation value is denoted by M. The highest the value of correlation, more the two voices are similar.

Find the maximum value of correlation and let Correlation be M. If the value of M is greater than the threshold then recognized else not found. If the correlation is more than the threshold value then it must be the same audio else the input audio is not detected in the database. The threshold value is set first. If the correlation value is greater than the threshold value then voice input sample will be called as recognized otherwise the tested voices will not contain the input sample of voice.

5.1 Database

The database consists of recorded voices which are used for matching with the input samples. This database consists of 100 different voices. These are different images with different poses from different type of sources**.**

Various samples are taken in which 100 set of recognized persons and 100 set of unrecognized persons is taken and then tested with the previously stored 10 certified recognized voices.

Table 1 shows the number of certified audio samples that are saved according to which the inputs can be matched. For the experimental setup, 10 samples of certified audio clips are taken. 100 samples of unrecognized inputs are taken which have to be matched with certified audio samples. Again 100 samples of unknown person audio recordings are taken.

Table No. 1. Set of Samples

Type	Sample Size
Certified	10
Input Recognized	100
Unknown Person Audio	100

The 200 samples in total are taken for testing. For the samples are taken then hits are calculated which gives the number of times the correct recognition is done when the input is present in the previously saved set of certified voices. Recognition is done by matching the input samples voices having wave format with the previously certified set of samples. Similarly, number of miss are evaluated accordingly voices when sample is not found in the record of certified voices.

VI. Results

The input audio signal waveform is shown in figure 2. It represents an audio signal or recording. The amplitude of the signal is measured on the y-axis and time is measured on the x-axis. This figure shows that background disturbance is present in the speech signals. The disturbance in signals degrades the performance. The accuracy rate slightly decreases as errors are introduced.

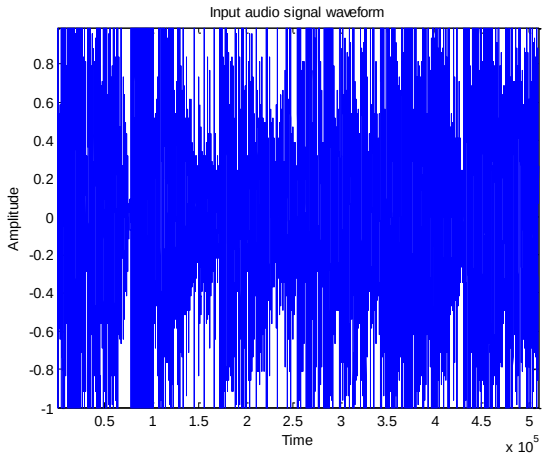

Fig. 2. input audio signal waveform

The spectrogram generated for a wave input is shown below in figure3. Spectrogram is Time-dependent frequency analysis. In the spectrogram the time axis is the horizontal axis and frequency is the

vertical axis. The input audio specgram shows that background disturbance is present in the speech signals and the accuracy rate slightly decreases.

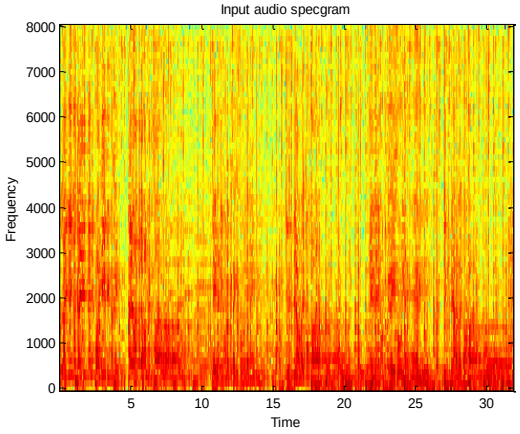

Fig. 3. input audio specgram

The filtered audio signal waveform is shown in figure 4. The background disturbance in the signals degrades the performance and noise reduction is an important field of speech processing. The filtered audio signal waveform shows that background noise was successfully removed from a signal by using butterworth stopband filter. In this matching is done using inverse wave transformations which reduce the time for recognition of voices. The filtered audio waveform shows that it has more accuracy.

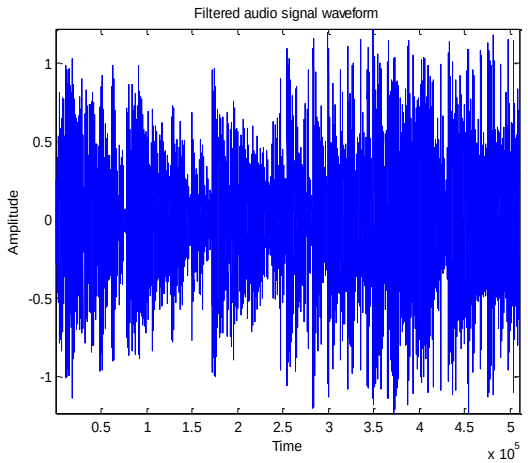

Fig. 4. filtered audio signal waveform

The spectrogram generated for a wave input is shown below in figure 5. The butterworth stopband filter is used to remove the disturbance from the speech signals and the accuracy is slightly increased. The matching is done using inverse wave transformations which reduce the time for recognition of voices. The filtered audio spectrogram shows that it has more accuracy than the other techniques.

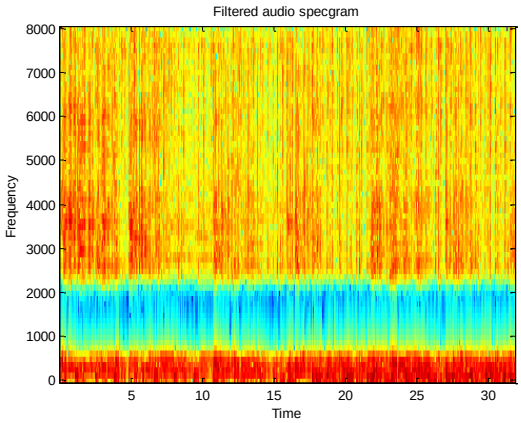

Fig. 5. filtered audio specgram

The compressed audio signal waveform is shown in figure 6. DCT based audio compression is used with butterworth stopband filter and inverse wave transformation to remove the disturbance from the speech signals and to speed up the process. DCT based data compression is the process of encoding information utilizing fewer bits than an un-encoded representation would use through use of particular encoding schemes. Background noise and redundancy are successfully removed from a signal by using Butterworth stopband filter and DCT. The compressed audio waveform shows that results are much better than exiting methods.

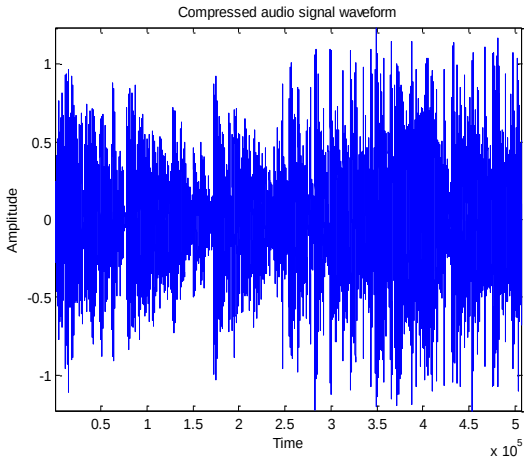

Fig. 6. compressed audio signal waveform

The spectrogram generated for a wave input is shown below in figure 7. DCT audio compression is used with butterworth stopband filter and inverse wave transformation to remove the disturbance from the speech signals and the accuracy is slightly increased. The compressed audio specgram shows that it has better results than the previous one.

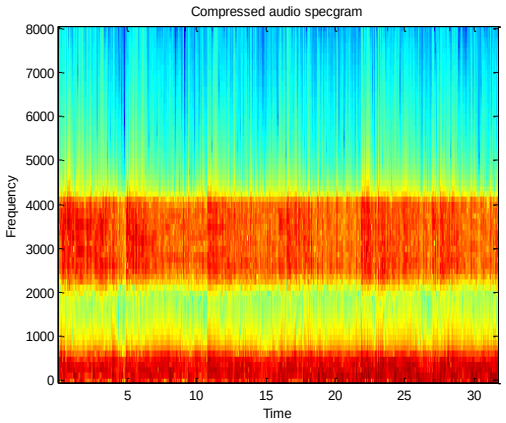

Fig. 7. compressed audio specgram

6.1 Comparison

Then according to formulas accuracy and error rates are calculated.

Figure 8 represents the hits analysis. It shows the difference between the number of hits of existing techniques and proposed technique. In order to distinguish them different colours has been used. In this fig X-axis represent number of tries and Y-axis represent number of hits. It has been plotted with different values of input samples and hits are calculated from the recognition of speeches. It is clearly seen that the hit ratio of old technique is very low as compared to new technique. As a result, more hits are obtained using DCT in speech recognition.

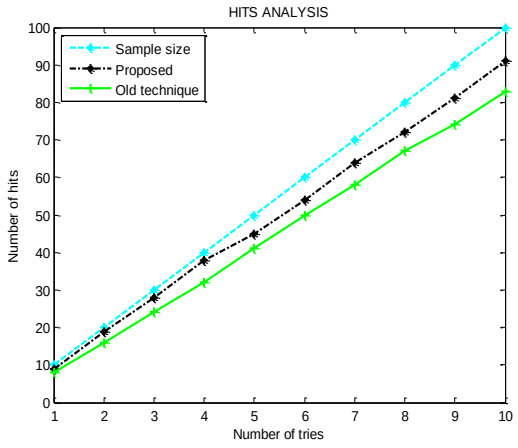

Fig. 8. hit analysis

Figure 9 represents the miss analysis. It shows the difference between the number of misses of existing techniques and proposed technique. In this fig X-axis represent number of tries and Y-axis represent number of misses. It is clearly seen that the miss ratio of new technique is very low as compared to old technique. As a result, less misses are obtained using DCT and filter in speech recognition. Miss ratio should be minimum and hit ratio should be maximum in order to achieve full accuracy.

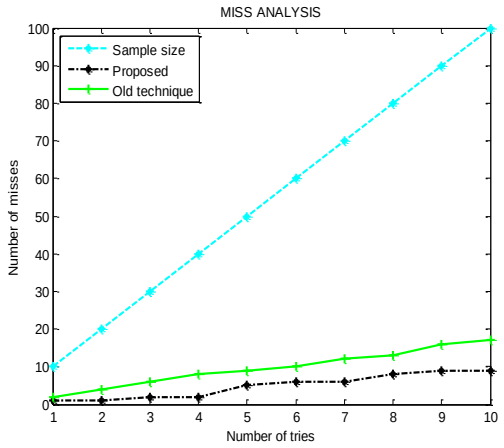

Fig. 9. miss analysis

Figure 10 represents the accuracy analysis. It shows the difference between the accuracy of existing techniques and proposed technique. In this fig X-axis represent number of tries and Y-axis represent accuracy in percentage. The maximum value of accuracy is 100 percent. It has been plotted with different values of input samples and accuracies are calculated from the recognition of speeches. It is clearly seen that the accuracy of new technique is very high as compared to old technique.

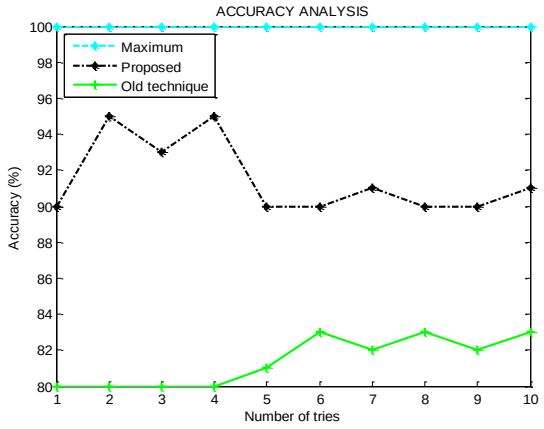

Fig. 10. accuracy analysis

Figure 11 represents the error rate analysis. It shows the difference between the error rate of existing techniques and proposed technique. In this fig X-axis represent number of tries and Y-axis represent error rate in percentage. The maximum value of error rate is 100 percent. It is clearly seen that the error rate of new technique is very low as compared to old technique. As a result, less error rate is obtained only when there are more hits and lesser misses. It is clearly seen that error rate of the old techniques is very high as close to the maximum value of error rate and the error rate of the new technique is very low as close to X-axis.

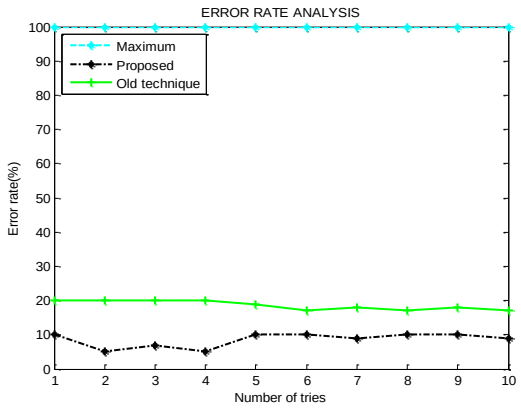

Fig. 11. error rate analysis

VII. CONCLUSION AND FUTURE WORK

7.1 Conclusion

The technique is able to authenticate the particular speaker based on the individual information that is included in the audio signal. The performance of the new technique is better than the existing techniques. Most of existing techniques does recognition of speech through the text wrapping. These techniques take too much time in recognizing the speech and are not accurate in the noisy environment. So the DCT and Butterworth stopband filter is used with inverse wave transformation to obtain the more accuracy and speed up the system. The DCT packs energy in the low frequency regions. The waveforms and specgrams showed that new technique provides high accuracy rate and consumes very less time in recognition. The error rate of the proposed algorithm is very low as compared to old algorithm.

7.2 Future Scope

In the future, the focus can be on reducing the noise or background disturbance that is introduced in the speech samples automatically while recording. Modified discrete cosine transform (MDCT) will be future compression algorithm, whether standalone or combination of speech and still or moving images. The various filtering techniques can be applied in order to reduce disturbance. By using these various filter techniques speech recognition will be more accurate and fast. The results can be further generalized if we are able to unite voice activation detection with this procedure we can perform speech recognition on live voices and speech. More research has to be done on this particular area to obtain more security.

References

[1] Susanta Kumar Sarangi, and Goutam Saha, "A Novel Approach in Feature Level for Robust Text-Independent Speaker Identification systemhigher", *IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction*, December 27-29, 2012

[2] Santosh K.Gaikwad, Bharti W. Gawali, and Pravin Yannawar, "A Review on Speech Recognition Technique*", International Journal of Computer Applications (0975 – 8887), Vol.10– No.3*, November 2010.

[3] K. Ramamohan Rao and P. Yip, "Discrete Cosine Transform: Algorithms, Advantages, Applications*", Academic Press Professional, Inc san Diego,* CA, USA, 1990.

[4] Swapnil D. Daphal, and Sonal K. Jagtap, "DSP Based Improved Speech Recognition System", *International Conference on Communication, Information and Computing Technology*, IEEE 2012.

[5] C. Y. Fook, "Malay Speech Recognition and Audio Visual Speech Recognition", *International Conference on Biomedical Engineering (ICoBE),* pp. 479-484, Feb. 2012.

[6] D. Addou, S.A. Selouani, M. Boudraa, and B. Boudraa "Transform-based multi-feature optimization for robust distributed speech recognition", *IEEE GCC conference and exhibition,* februaury, 2011.

[7] M.D. Pawar, and S.M. Badave,"Speaker Identification System Using Wavelet Transformation and Neural Network", *International Journal of Computer Applications in Engineering Sciences*, *Vol. 1*, special issue on CNS, July 2011.

[8] Yu Shao, "Bayesian Separation with Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition", *IEEE Transactions on Systems, Vol. 41*, pp. 284-294, March 2011.

[9] Ozlem Kalinli, Michael L. Seltzer, Jasha Droppo, and Alex Acero*,"* Noise Adaptive Training for Robust Automatic Speech Recognition", *IEEE Transactions On Audio, Speech and Language Processing, Vol. 18, No. 8*, November 2010.

[10] Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar1, and Elias Yaacoub, "Speech Recognition using Artificial Neural Networks and Hidden Markov Models", *IMCL 2008 Conference,* 16-18 April 2008.

[11] Patricia Scanlon, Daniel P. W. Ellis, and Richard B. Reilly, "Using Broad Phonetic Group Experts for Improved Speech Recognition" *IEEE Transactions on audio, speech and language processing, Vol. 15*, pp. 803-812, March 2007.