A Study on Prosody Analysis

Padmalaya Pattnaik [1], Shreela Dash [2]
(Asst.Prof, C.V. Raman College of Engineering, Bhubaneswar, Odisha, India)

Abstract:
Speech can be described as an act of producing voice through the use of the vocal folds and vocal apparatus to create a linguistic act designed to convey information. Linguists classify the speech sounds used in a language into a number of abstract categories called phonemes. Phonemes are abstract categories, which allow us to group together subsets of speech sounds. Speech signals carry different features, which need detailed study across gender for making a standard database of different linguistic, & paralinguistic factors. Prosodic phenomena are specific to spoken language. They concern the way in which speech sounds are acoustically realized: how long they are, how high and how loud. Such acoustic modulations are used by human speakers to express a variety of linguistic or paralinguistic features, from stress and syntactic boundaries, to focus and emphasis or pragmatical and emotional attitudes. Linguistics and speech technology have approached prosody from a variety of points of view, so that a precise definition of the scope of prosodic research is not easy. A main distinction can be drawn between acoustic-phonetic analyses of prosody and more abstract, linguistic, phonological approaches When people interact with others they convey emotions. Emotions play a vital role in any kind of decision in affective, social or business area. The emotions are manifested in verbal, facial expressions but also in written texts. The objective of this study is to verify the impact of various emotional states on speech prosody analysis.

Keywords: *Duration, Emotion, Jitter, Prosody, Shimmer*

1. Introduction

No language is produced in a smooth, unvarying stream. Rather, the speech has perceptible breaks and clumps. For example, we can perceive an utterance as composed of words, and these words can be perceived as composed of syllables, which are composed of individual sounds. At a higher level, some words seem to be more closely grouped with adjacent words: we call these groups phrases. These phrases can be grouped together to form larger phrases, which may be grouped to form sentences, paragraphs, and complete discourses. These observations raise the questions of how many such constituents there are and how they are best defined. A fundamental characteristic of spoken language is the relation between the continuous flow of sounds on the one hand, and the existence of structured patterns within this continuum on the other hand. In this respect, spoken language is related to many other natural and man-made phenomena, which are characterized not only by their typically flowing nature but also by the fact that they are structured into distinct units such as waves and measures. Prosodic phonology is a theory of the way in which the flow of speech is organized into a finite set of phonological units. It is also, however, a theory of interactions between phonology and the components of the grammar. Although many speech interfaces are already available, the need is for speech interfaces in local Indian languages. Application specific Indian language speech recognition systems are required to make computer aided teaching, a reality in rural schools. This paper presents the preliminary work done to demonstrate the relevance of an Oriya Continuous Speech Recognition System in primary education. Automatic speech recognition has progressed tremendously in the last two decades. There are several commercial Automatic Speech Recognition (ASR) systems developed, the most popular among them are Dragon Naturally Speaking, IBM Via voice and Microsoft SAPI. Speech is a complex waveform containing verbal (e.g. phoneme, syllable, and word) and nonverbal (e.g. speaker identity, emotional state, and tone) information. Both the verbal and nonverbal aspects of speech are extremely important in interpersonal communication and human-machine interaction. Each spoken word is created using the phonetic combination of a set of vowel semivowel and consonant speech sound units. Different stress is applied by vocal cord of a person for particular emotion. The increased muscle tension of the vocal cords and vocal tract can directly or indirectly and adversely affect the quality of speech. We use emotions to express and communicate our feelings in everyday life. Our experience as speakers as well as listeners tells us that the interpretation of meaning orintention of a spoken utterance can be affected by the emotions that area expressed and felt.

2. Literature

According to the classic definition, prosody has to do with speech features whose domain is not a single phonetic segment, but larger units of more than one segment, possibly whole sentences or even longer utterances. Consequently, prosodic phenomena are often called supra-segmentals. They appear to be used to structure the speech flow and are perceived as stress or accentuation, or as other modifications of intonation, rhythm and loudness.

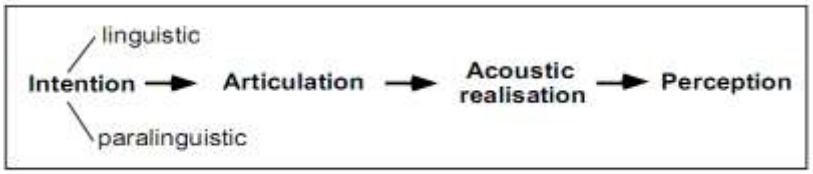

Fig: 1 Stages of oral communication

An emotion is a mental and physiological state associated with a wide variety of feelings, thoughts, and internal (physical) or external (social) behaviors. Love, hate, courage, fear, joy, sadness, pleasure and disgust can all be described in both psychological and physiological terms. An emotion is a psychological arousal with cognitive aspects that depends on the specific context. According to some researcher, the emotions are cognitive processes. Emotion is a process in which the perception of a certain set of stimuli, follows cognitive assessment which enables people to label and identify a particular emotional state. At this point there will be an emotional physiological, behavioral and expressive response. For example, the primordial fear, that alerts us as soon when we hear a sudden noise, allows to react to, dangerous situations and provides instantly resources to face them as escape or close the door. The emotional stimuli may be an event, a scene, a face, a poster, an advertising campaign. These events, as a first reaction, put on alert the organism with somatic changes as heart rate, increase of sweat, acceleration of respiratory rhythm, rise of muscle tensions.

Emotions give an immediate response that often don't use cognitive processes and conscious elaboration and sometimes they have an effect on cognitive aspects as concentration ability, confusion, loss, alert and so on. This is what is asserted in evaluation theory, in which cognitive appraisal is the true cause of emotions [2]. Two factors that emerge permanently are those related to signals of pleasure and pain and characterizing respectively the positive and negative emotions. It's clear that these two parameters alone are not sufficient to characterize the different emotions. Many authors debate on primary and secondary emotions other on pure and mixed emotions, leaving the implication that emotions can somehow be composed or added.

The systems based on the analysis of physiological response as blood pressure, heart rate, respiration change present an initial phase where the signals are collected in configurations to be correlated with different emotional states and a subsequently recognition basing on the measure of indicators. One of the interesting early works on the emotions was that one of Ortony [3]. From this work, through componential analysis, other authors constructed an exhaustive taxonomy on affective lexicon. According to Ortony, stimuli that cause emotional processes are of three basic types: events, agents and objects corresponding to three classes of emotions: satisfied/ unsatisfied (reactions to events), approve/disapprove(reaction to agents), appreciate/unappreciate (reaction to objects). According to Osgood [4] an emotion consists of a set of stages: stimulus (neural and chemical changes), appraisal and action readiness. Continuing the studies of Charles Darwin, the Canadian psychologist Paul Ekman [5] has confirmed that an important feature of basic emotions is that they are universally expressed, by everybody in any place, time and culture, through similar methods. Some facial expressions and the corresponding emotions are not culturally specific but universal and they have a biological origin. Ekman, analyzed how facial expressions respond to each emotion involving the same type of facial muscles and regardless of latitude, culture and ethnicity. This study was supported by experiments conducted with individuals of Papua New Guinea that still live in a primitive way.

3. Emotions

Human emotions are deeply joined with the cognition. Emotions are important in social behavior and to stimulate cognitive processes for strategies making. Emotions represent another form of language universally spoken and understood. Identification and classification of emotions has been a research area since Charles Darwin's age. In this section we consider facial, vocal and textual emotional expressions.

Emotion classifications of the researchers differ according to the goal of the research and the field .Also the scientist's opinion about the relevance of dividing different emotions is important. There is no standard list of basic emotions. However, it is possible to define list of emotions which have usually been chosen as basic, such as: erotic(love)(shringar), pathetic (sad) (karuNa), wrath (anger)(roudra), quietus (shAnta), normal(neutral).

The objective of this study is to analyze impact for different emotions on vowels in terms of certain parameters for stage prosody analysis.

4. Prosody

In linguistics, prosody is the rhythm, stress, and intonation of speech. Prosody may reflect various features of the speaker or the utterance: the emotional state of the speaker; the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast, and focus; or other elements of language that may not be encoded by grammar or choice of vocabulary. Prosody has long been studied as an important knowledge source for speech understanding and also considered as the most significant factor of emotional expressions in speech[16]. Prosody gives

naturalness and message intelligibility to speech Emotional prosody is the expression of feelings using prosodic elements of speech. Linguistically relevant prosodic events concur to express sentence structure: they highlight linguistic units by marking their boundaries and suggesting their function. Several types of prosodic units (differing mainly in their scope) have been proposed: paragraphs, sentences, intonation groups, intermediate groups, stress groups, syllables etc. Although prosody is by definition suprasegmental, prosodic analyses take often the phoneme as their minimal unit, where to measure rhythmical variations and locate intonation events. The family of prosodic phenomena includes the suprasegmental features of intonation, stress, rhythm and speech rate, whose variations are relevant to express the function of the different prosodic units: the prominent syllable in the word will be marked by stress, a falling intonation contour will mark the conclusion of a sentence, a faster speech rate and lower intonation characterize a parenthetical phrase.Such prosodic features are physically realized in the speech chain in terms of variations of a set of acoustic parameters. Acoustic-phonetic analyses identify the following 'phonetic correlates of prosody': fundamental frequency/pitch (f0), length changes in segmental duration, pauses, loudness, voice quality.

Prosody is the combination of voice's pitch, duration and energy variation during speech. It provides an additional sense to the words, which is extraordinarily important in natural speech. For example, interrogative and declarative sentences have very different prosody (especially intonation). Besides, the prosody of a sentence is one of the factors that make a speaker seem happy, sad, angry or frightened. We can even decide from prosody if the speaker is an energetic person or, on the contrary, a lazy one. When singing, intonation and timing evolution characterize melody. But prosody is not only important in natural speech but also in synthetic speech. Prosody is crucial in order to achieve an acceptable naturalness. If a TTS system does not have a good prosodic treatment, its output speech sounds completely monotonous and, moreover, it won't be able to distinguish between sentences of different kinds. The two main parameters of prosody are-

4.1 Intonation

In a first approximation, sounds can be classified invoiced and unvoiced sounds. Voiced sounds, unlike unvoiced ones, are produced making the vocal chordsvibrate. These vibrations provoke some periodicitiesin the speech signal and therefore a fundamental frequency (F0). This value is inversely proportional tothe distance between periodicities and it makesspeech sound with higher or lowers frequency. It iscommonly called pitch. On the contrary, as unvoiced sounds do not have any periodicities (vocal chords donot vibrate) and can be modeled as a filtered noise signal. So if we detect the pitch curve of a speech signal it will only exist in the voiced segments. Pitch is not constant but its value changes during asentence. That is called intonation. Thanks tointonation we can distinguish, for example, between adeclarative and an interrogative sentence or identifyfocused words inside a sentence.

4.2 Duration

The duration of speech segments is the other main parameter of prosody. The timing structure of a sentence is extremely important to give naturalness to speech. Phone duration depends on a great number of parameters, such as its phonetic identity, surroundingphones, level of stress or position in the sentence or in the word. What's more, duration of a word also depends on its importance in the sentence. For example, a focused word will generally have longer duration.

5. Prosody Analysis

TTS systems generate speech from a text. There is a need of prosodic assignment to phones to produce high quality speech. Once the phones to synthesize are determined, it is necessary to know the pitch and duration yet achieved the required quality needed for most applications. This is the main reason why the tool was created: the prosodic module requires models of prosodic patterns and these patterns have to be studied and tested before the application to TTS.The goal of the prosodic analysis is the generation of the pitch contour[17].

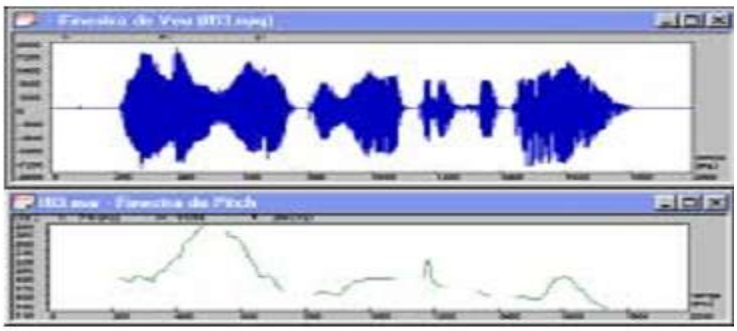

Fig: 2 Speech Segment & it's pitch contour

6. Parametric Measurements of Acoustics Signals for Prosody Analysis

Four acoustic signal features such as Duration of speech segment, Pitch, Jitter and Simmer were used to parameterize the speech.

6.1 Duration

Utterance durations, vowel durations were measured from the corresponding label files produced by a manual segmentation procedure. On an average, utterance durations become longer when speech is emotionally elaborated.

6.2 Fundamental Frequency (pitch)

We calculated the pitch contours of each utterance using speech processing software. Global level statistics related to F0 such as minimum, maximum, mean were calculated from smoothed F0 contours.

6.3 Jitter & Shimmer

Jitter and Shimmer are related to the micro-variations of the pitch and power curves. In other words, Shimmer and Jitter are the cycle-to-cycle variations of waveform amplitudes and fundamental periods respectively. The Jitter & Shimmer occur due to some undesirable effect in audio signal. Jitter is the period frequency displacement of the signal from the ideal location. Shimmer is the deviation of amplitude of the signal from the ideal location.

7. Methodology and Experimentation

There are many features available that may be useful for classifying speaker affect: pitch statistics, short-time energy, long-term power spectrum of an utterance, speaking rate, phoneme and silence durations, formant ratios, and even the shape of the glottal waveform [8, 11, 12, 9, 10]. Studies show, that prosody is the primary indicator of a speaker's emotional state [1, 13, 12]. We have chosen to analyze prosody as an indicator of affect since it has a well-defined and easily measureable acoustical correlate -- the pitch contour. In order to validate the use prosody as an indicator for affect and to experiment with real speech, we need to address two problems: First, and perhaps most difficult, is the task of obtaining a speech corpus containing utterances that are truly representative of an affect. Second, what exactly are the useful features of the pitch contour in classifying affect? Especially as many factors influence the prosodic structure of an utterance and only one of these is speaker's emotional state [6, 7, 9].

The data analyzed in this study were collected from semi-professional actors and actress and consists of 30 unique Odia language sentences that are suitable to be uttered with any of the five emotions i.e., roudra, shringar, shanta, karuna, and neutral. Some example sentences are "Jayanta jagi rakhhi kaatha barta kaara", "*Babu chuata mora tinni dina hela khaaini*". The recordings were made in a noise free room using microphone. For the study, samples have been taken from three male & three female speakers. The utterances are recorded at the bit rate of 22,050Hz. The Vowels are extracted from the words consisting of 3 parts i.e.CV, V, VC. CV stands for Consonant to Vowel transition, V for steady state vowel, VC for Vowel to Consonant transition.

8. Experimental Results

For experimental analysis data samples were created from recordings of male and female speakers in various emotions (mood). Vowels are then extracted and stored in a database for analysis. From this database after analysis the result of the utterance. Duration, fundamental frequency & variations of pitch of every vowel are measured and compared to give following results.

8.1 Duration

It is observed from the duration Table 1 that speech associated with the emotion "love(Shringar)"has higher duration with the vowel /i/ gets elongated both for male and female speakers where as the emotion "sad (karuna)" for male speakers vowel(/a/,/i/) gets elongated whereas for female (/i/,/u/) gets elongated.

Table (1): Average duration of vowels for speakers in different emotions

Emotions	Duration in Millie Seconds (male-average)				
	/a/	/i/	/u/	/e/	/o/
Neutral	67	70	60	39	53
Santa	58	75	60	58	43
Karuna	**106**	**106**	61	56	54
Raoudra	64	50	43	49	56
Shringar	**100**	**113**	**90**	54	56
Emotions	Duration in Millie Seconds (female-average)				
	/a/	/i/	/u/	/e/	/o/
Neutral	50	64	66	40	35
Santa	53	**99**	79	43	50
Karuna	**83**	**105**	**101**	**86**	63
Raoudra	50	43	58	38	41
Shringar	**80**	**161**	**118**	55	54

8.2 Fundamental Frequency

Figure3 shows the analysis result that the mean pitch for male speaker associated with emotion karuna for vowel /i/ has dominance where as in female the speech associated with emotion shringar for vowel /a/ plays dominant role.

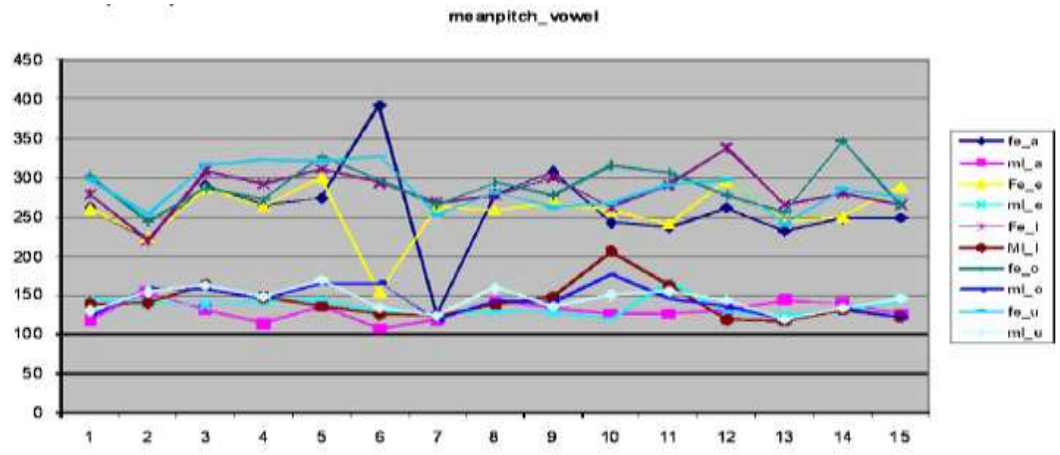

Fig: 3 mean pitch of vowels of speakers in different emotion

8.3 Jitter

Figure 4 show that the emotion "anger" of vowel /i/ is having a dominant role in jitter for male speaker. The jitter value for female speaker has dominant vowel /u/ for the emotion "love".

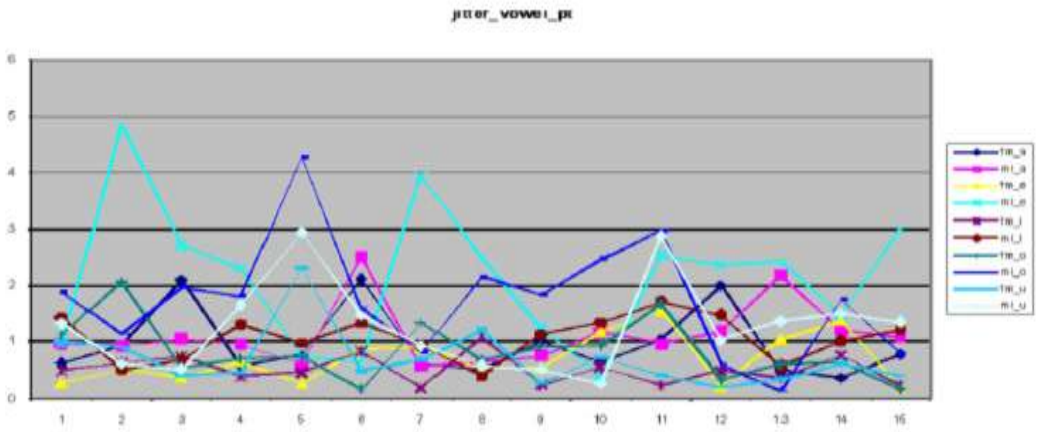

Fig:4 jitter of vowels of speakers in different emotion

8.4 Shimmer

It is observed from figure 5 that the shimmer in the emotion "anger" of vowel /o/ has dominant role for males & for female in emotion "anger" of vowel /i/ has dominant role.

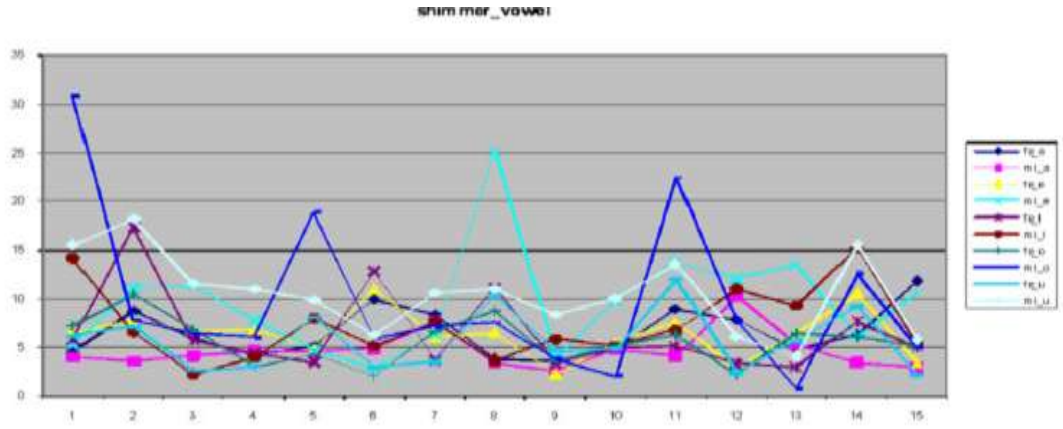

Fig:5 shimmer of vowels of speakers in different emotion

9. Conclusion

The great importance and complexity of prosody in speech makes this subject an important area of research in speech synthesis applications In this study, we investigate acoustic properties of speech prosody associated with five different emotions (love (shringar)), pathetic (sad) (karuNa), wrath (anger) (roudra), quietus (shAnta),normal (neutral) intentionally expressed in speech by male and female speakers. Results show speech associated with love (shringar) and sad (karuna) emotions are characterized by longer utterance duration, and higher pitch. However we observed that for jitter anger or love has dominance over others, whereas for shimmer the emotion anger plays a vital role. Future works of my research are the following. We have to collect synthetic speech and put emotion labels on them. We have to reconsider how to estimate emotion in speech using parallel programming.

References

[1] P. Olivier and J. Wallace, Digital technologies and the emotional family, International Journal of Human Computer Studies, 67 (2), 2009, 204-214.

[2] W. L. Jarrold, Towards a theory of affective mind: computationally modeling the generativity of goal appraisal, Ph.D. diss., University of Texas, Austin, 2004.

[3] C. G. Ortony and A. Collins, The cognitive structure of emotions, (Cambridge University Press: New York, 1990).

[4] M. M. C.E. Osgood and W.H. May, Cross-cultural Universals of Affective Meaning, (Urbana Chaimpaigne: University of Illinois Press, 1975).

[5] E. Paul, Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life, (NY: OWL Books, 2007).

[6] A. Ichikawa and S. Sato, Some prosodical characteristics in spontaneous spoken dialogue, International Conference on Spoken Language Processing, v. 1, 1994, 147-150.

[7] R. Collier, A comment of the prediction of prosody, in G. Bailly, C. Benoit, and T.R. Sawallis (Ed.), Talking Machines: Theories, Models, and Designs, (Amsterdam: Elsevier Science Publishers, 1992).

[8] H. Kuwabara and Y. Sagisaka, Acoustic characteristics of speaker individuality: Control and conversion, Speech Communication, 16(2), 1995, 165-173.

[9] K. Cummings and M. Clements, Analysis of the glottal excitation of emotionally styled and stressed speech, Journal of the Acoustical Society of America, 98(1), 1995, 88-98.

[10] D. Roy and A. Pentland, Automatic spoken affect classification and analysis, Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition, 1996, 363-367.

[11] A. Protopapas and P. Lieberman, Fundamental frequency of phonation and perceived emotional stress, Journal of Acoustical Society of America, 101(4), 1997, 2267-2277.

[12] X.Arputha Rathina and K.M.Mehata, Basic analysis on prosodic features in emotional speech, nternational Journal of Computer Science, Engineering and Applications (IJCSEA) , Vol.2, No.4, August 2012,99-107

[13] D. Hirst, Prediction of prosody: An overview, in G. Bailly, C. Benoit, and T.R. Sawallis (Ed.), Talking Machines: Theories, Models, and Designs, (Amsterdam: Elsevier Science Publishers, 1992).

[14] L. Rabiner and R. Shafer, Digital Processing of Speech Signals, (New York: Wiley and Sons, 1978)

[15] Wavesurfer, http://www.speech.kth.se/wavesurfer/

[16] Masaki Kuremastsu et eal, An extraction of emotion in human speech using speech synthesize and classifiers foe each emotion , International journal of circuits systems and signal processing, 2008.

[17] J .L. Navarro, I. Esquerra, A Time-FrequencyApproach to Epoch Detection, Proceedings ofEurospeech'95, 405-408, Madrid 1995.