Developing Syllable-Centric Speech Synthesis for Marathi Language Applications

Aniket D. Jadhav*[1] & S. B. Khadse[2]

*[1]National Institute of Electronics & Information Technology, Dr.Babasaheb .Ambedkar .
Marathwada.Universirty, Aurangabad, India
[2]Zeal College of Engineering & research, Pune .India

ABSTRACT

Speech synthesis is the most significant applications in linguistic communication process. The Text to Speech structure is the undertaking of accepts the input sentence and converts the audible speech as output. The Marathi language may be a syllable based language. A syllable is the unit of language, which may be spoken independent of the adjacent phones. It consists of an interrupted portion of sound, once the word is pronounced. The task of proposed Text to Speech System for Marathi language includes syllabication, Letter-to-Sound rules and concatenation. Syllabication is that the method of distinguishing the linguistic unit units, that is presented within the given input. The Trainable Text syllabication algorithm is employed for deriving the syllables. The Letter to Sound mapping technique is employed for changing the text to phonemes. These phonemes square measure mapped with the waveform that may be a recorded sound file, which can be a variety of wave files. The recorded sounds are concatenated by Unit selection Speech Synthesis algorithm, which uses the massive databases of recorded speech. The efficient joining cost is required to be calculated for locating the best sequence of speech as synthesized output. Java Media Framework speak engine is employed to synthesis the speech. The proposed text to speech system founded on syllable unit for Marathi language is employed to boost the excellence of speech.

KEYWORDS: Syllabification, Text to Speech synthesis, Letter to Sound conversion, Unit Selection Speech synthesis,

I. CONCATENATION COST, TARGET COST.INTRODUCTION

In India, the physically-impaired population has touched an alarming figure of 8.9 million of whom almost 15% suffer from speech and visual impairments. This section of the population depends solely on augmentative and alternative communication techniques for their education and communication skills. Different tools have been implemented for these people but, unfortunately, they are in the English language and are too costly for the Indian population. In response to their need, we have taken up the task of developing low-cost portable communication tools to aid the speech-impaired population in India. In this paper, we describe an Indian language text-to-speech system that accepts text inputs in Marathi, and produces near-natural audio output [10].

A large speech database is needed to achieve more natural synthesized speech. In most of the concatenative speech synthesis systems, search units are rather short such as syllables, phonemes and diaphone. A shorter unit, however, produces a larger number of candidates of voice waveform and a larger speech database cannot be used without narrow pruning for practical use, but narrow pruning impairs the quality of synthesized speech [1]. This method is expected to make synthesized speech more natural.

II. TEXT-TO-SPEECH SYSTEM

Text-to-Speech (TTS) System is a computer based system that should be able to read any text aloud, whether it was introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system [2]. The objective of a text to speech system is to convert an arbitrary given text into a spoken waveform.

III.SPEECH GENERATION COMPONENT

Given the sequence of phonemes, the objective of the speech generation component is to synthesize the acoustic waveform. Speech generation has been attempted by concatenating the recorded speech segments. Current state-of-art speech synthesis generates natural sounding speech by using large number of speech units. The approach of using an inventory of speech units is referred to as unit selection approach [12], [15]. The issues related to the unit selection speech synthesis system are Choice of unit size, Generation of speech database, Criteria for selection of a unit.

A. Concatenative Synthesis

In this approach synthesis is done by using natural speech. This methodology has the advantage in its simplicity, i.e. there is no mathematical model involved. Speech is produced out of natural, human speech [3]. Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. There are three main sub-types of concatenative synthesis: Unit selection synthesis, Diaphone synthesis, Domain-specific ssynthesis.unit selection speech synthesis system are choice of unit size, generation of speech database, criteria for selection 0f a unit.

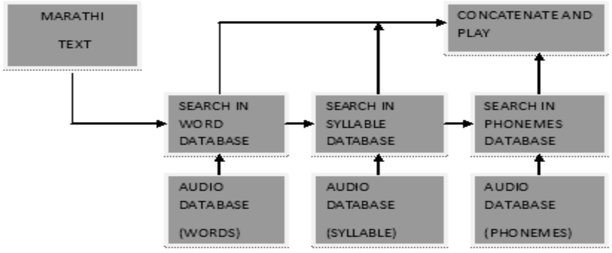

MARATHI TEXT
CONCATENATE AND PLAY
SEARCH IN WORD DATABASE
SEARCH IN SYLLABLE DATABASE
SEARCH IN PHONEMES DATABASE
AUDIO DATABASE (WORDS)
AUDIO DATABASE (SYLLABLE)
AUDIO DATABASE (PHONEMES)

Fig.1 Block diagram of speech synthesis system.

B. *Unit Selection Synthesis*

Unit selection synthesis uses large databases recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences [4], [8]. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some manual correction afterward, using visual representations such as the waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighbouring phones [22]. At runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). Unit selection provides the greatest naturalness, because it applies only small amounts of digital signal processing (DSP) to the recorded speech [13]. DSP often makes recorded speech sound less natural, although some systems use a small amount of signal processing at the point of concatenation to smooth the waveform. The output from the best unit-selection systems is often indistinguishable from real human voices, especially in contexts for which the TTS system has been tuned [11].

IV. INVENTORY DESIGN

TTS System is composed of two parts: A front-end that takes input in the form of text and outputs a symbolic linguistic representation. A back-end that takes the symbolic linguistic representation as input and outputs the synthesized speech in waveform. These two phases are also called as high-level synthesis phase and low-level synthesis phase, respectively. A recent trend in concatenative synthesis approach is to use large databases of phonetically and prosodically varied speech. The quality of the output speech primarily depends on the quality of the speech corpus [16].

V. SPEECH SYNTHESIS PROCESS

The text input is either non-standard words or standard words. If the input text is a number then it is handled by a digit processor. If input text is word then it searched in the word database. If the word does not exist in the database then it is cut into syllables and syllables are searched in the syllable database. If the corresponding syllable does not exist in the database then word is formed by concatenating barakhadi in the barakhadi database and played as shown in fig.2 [5]-[9].

A. *Database Creation and Searching*

Two databases are maintained viz. audio database that stores the audio files and textual database that stores the text files corresponding to audio files in the audio database. The textual database is required to search the index of the required word in the audio database [3], [5]. When the word does not exist then it is synthesized from syllables. The consonant vowel structure (CV) breaking of the word is performed.

B. *Cutting of the syllables*

While forming the new word that is not present in the database, we cut that word into syllables, then search the syllables into the database & concatenate them [14]. Thus we will have to cut the pre recorded words present in the database file into the syllables & select the particular syllable that we want to form the new word. For this purpose cutting of the word into the syllables must be very accurate.

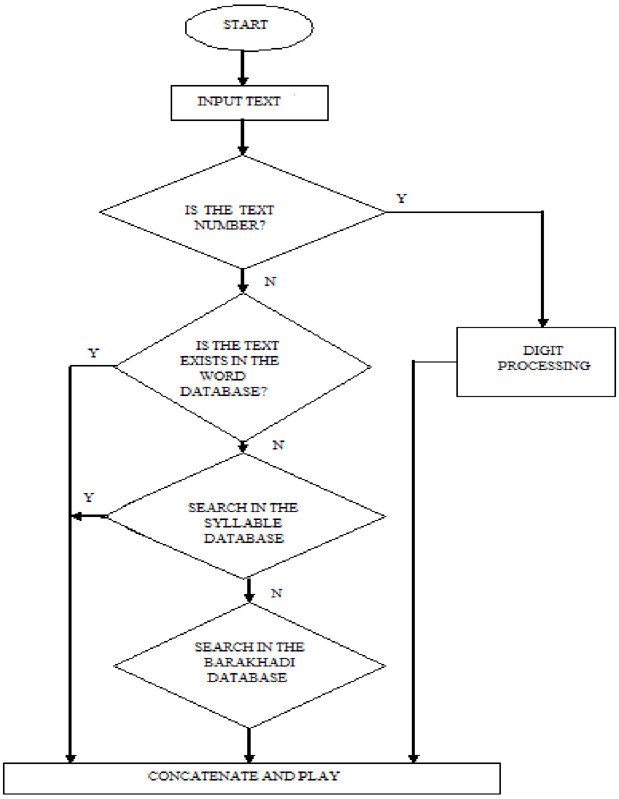
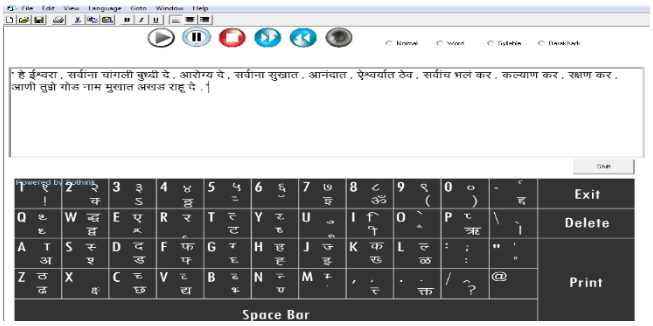

START
INPUT TEXT
IS THE TEXT NUMBER?
Y
N
DIGIT PROCESSING
IS THE TEXT EXISTS IN THE WORD DATABASE?
Y
N
SEARCH IN THE SYLLABLE DATABASE
Y
N
SEARCH IN THE BARAKHADI DATABASE
CONCATENATE AND PLAY

Fig.2 Design flow of TTS System

C. Front End

This TTS system is able to read any written text, even if it contains numbers, dates, time, addresses, telephone numbers and bank account numbers. This process is often called text normalization, pre-processing and tokenization. Front end is developed & coded in VB 6.0 as shown in fig. 3.



" हे ईश्वरा , सर्वांना चांगली बुध्दी दे , आरोग्य दे , सर्वांना सुखात , आनंदात , ऐश्वर्यात ठेव , सर्वांच भलं कर , कल्याण कर , रक्षण कर , आणी तुझे गोड नाम मुखात अखंड राहू दे . "

Fig.3 Text processing front end.

VI. PERFORMANCE EVALUATION

In order to evaluate the performance, the speech samples were synthesized by the proposed method and compared with those **made by the** conventional method using phonemes as a database [3].

Opinion Score	160 Min.	120 Min	80 Min	60 min	Natural speech
1	2	3	6	11	0
1.5	6.5	7	8	12	0
2	9	10	10	18	0
2.5	13.5	15	20	23	0
3	18	20	38	30	0

TABLE I

3.5	28	27	38	27	1
4	35	33	39	24	2
4.5	36	32	38	21	40
5	37	31	38	17	40

FIVE POINT MOS TEST

A. Paired Comparison Test

The listeners were five males and three females without any known hearing problems. The speech samples were presented through loud-speakers in a sound-proof room. The listeners were asked to listen to the speech samples only once because the mean length of one sentence was very long (about ten seconds) [5]. The listeners were asked to judge which of the two samples of the same target sentence they considered to be more natural. They were not allowed to judge both samples of the pair equally good. Each speech sample of a pair was arranged in random order, and the order of the sentence pairs was randomized, too [15]. The listeners took a rest intermittently. Fig.4. depicts the result of the paired comparison test.

Experimental result of paired comparison test reveals that 74% of synthesized speech by proposed method was evaluated as more natural speech than synthesized speech by the conventional method.

B. Five Point MOS Test

The perceptual scale for five-point MOS test conducted was 5. Natural, 4. Not natural but negligible, 3. Slightly noticeable, 2. Noticeable, 1. Very noticeable

System performance is evaluated using the proposed method with speech databases of different size. Forty speech samples were synthesized using the entire speech database of 160 min, three-fourth of 120 min, half of 80 min, one-eighth of 20 minute. Forty original speech samples were evaluated in the five point MOS test. The speech samples were presented through loud-speakers to the listener. They were asked to listen and rate them according to five-point rating scale [6].

Experimental result of five-point MOS test and opinion score for speech databases reveals that when the database is smaller, synthesized speech rated at 5 (natural) and 4 (not natural but negligible) decreases and synthesized speech rated at 3 (slightly noticeable), 2 (noticeable) and 1 (very noticeable) increases.

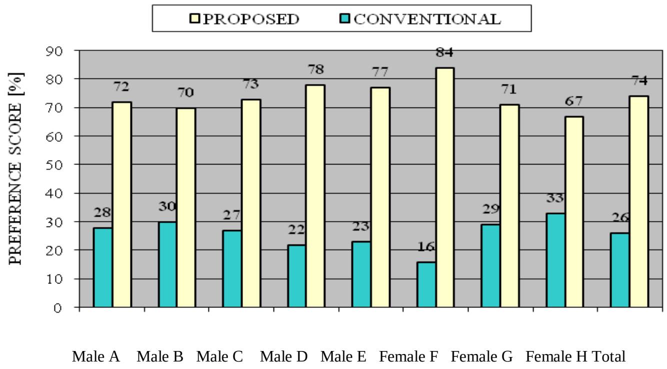


Fig. 4 Result of a paired comparison test.

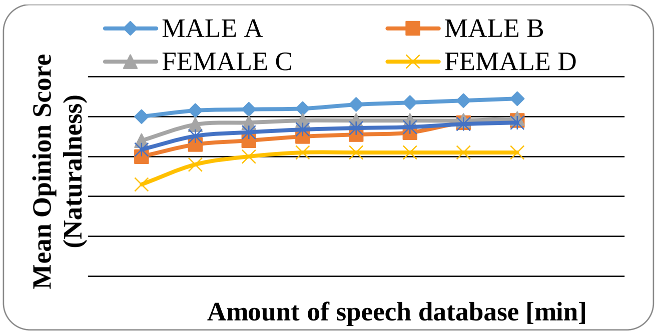


Fig. 5 MOS for various speech databases

TABLE II
OPINION SCORE FOR VARIOUS SPEECH DATABASES

DATABASE (MINUTES)	MALE A	MALE B	FEMALE C	FEMALE D	TOTAL
20	4	3	3.4	2.3	3.18
40	4.15	3.3	3.8	2.8	3.51
60	4.18	3.4	3.85	3	3.61
80	4.2	3.5	3.9	3.1	3.68
100	4.3	3.55	3.9	3.1	3.71
120	4.35	3.6	3.9	3.1	3.74
140	4.4	3.85	3.9	3.1	3.81
160	4.45	3.9	3.95	3.1	3.85

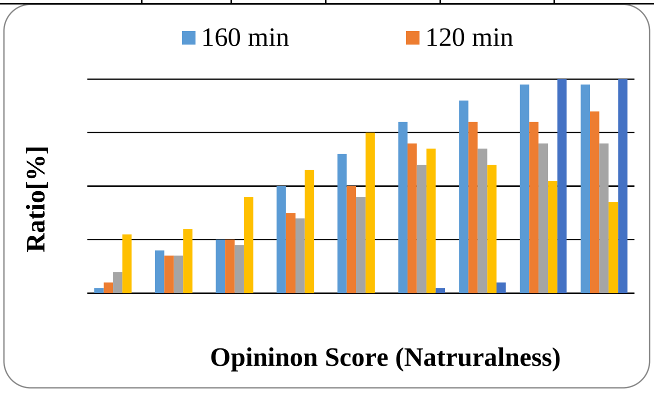


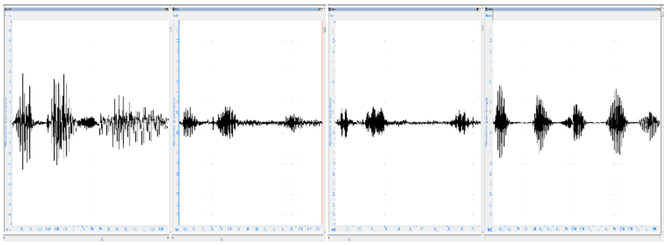
Fig. 6 Histogram for various speech databases

11.1% of synthesized speech with the 160 min speech database was rated at 2 and 1, so it is important to decrease the rate of such unnatural synthesized speech [15].

C. Spectrogram Analysis

Speech signal is represented as a sequence of spectral vectors. Time versus frequency representation of speech signal is referred to as spectrogram. Phones and their properties are better observed in spectrogram. Sounds can be identified much better by the Formants and by their transitions MAP spectral amplitude of speech signal. '0' represents black and '255'represents white. Higher the amplitude, darker the corresponding region. A high quality text to speech system should produce synthesized speech whose spectrograms should nearly match with the natural sentences [7]-[18].

SPECTROGRAM ANALYSIS OF WORD 'AKASHWANI'



Original word 'Akashwani' Generated & recorded syllables barakhadi

Fig. 7 Speech signal of 'AKASHWANI'

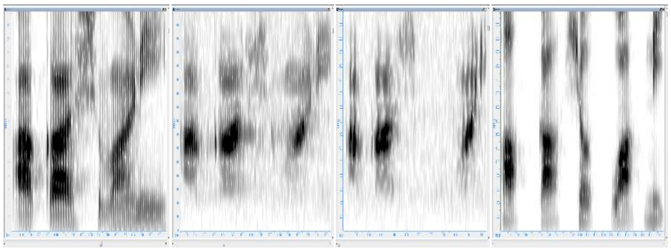


Fig. 8 Spectrogram of speech signal ‘AKASHWANI’

TABLE III
DATA OF SPEECH SYNTHESIZER FOR WORD ‘AKASHWANI’

Spectrogram analysis of the word ‘Akashwani’ speech signal is represented as a sequence of spectral vectors. Dark regions indicate peaks (formants) in the spectrum [19]. Spectrogram analysis is carried for recorded original word, TTS system generated word, syllables concatenation & barakhadi concatenation. From the spectrograms, it is observed that, dark regions and pattern of the pitch is almost same for original words and concatenated words with syllables & barakhadi. Naturalness hampers from word to barakhadi concatenation. Table III indicates that the number of samples and length increases from word to barakhadi concatenation. Samples and length parameters depend on the process of cutting of the syllables and barakhadi [20]. If cutting and concatenation is properly optimized then these parameters deviates less from the values of the original word & concatenated word. It is clear that samples and length variation not more than 10% to 25 % means cutting & concatenation were done properly which results more than 80 % naturalness in the speech output.

D. Comprehension Test

This is purely a test of intelligibility, concern with the ability to identify what was spoken or synthesized.

Parameter	Original	Generated	Syllable	Phonemes
Number of samples	43008	54400	74272	91648
Length	0.9755sc	1.1715 sec	1.3315 sc	1.5145 sec
Sampling Fc	44.1KHz	44.1K Hz	44.1KHz	44.1KHz
Bandwidth	22050Hz	22050 Hz	22050 Hz	22050 Hz
Storage format	16 bits stereo	16 bits stereo	16 bits stereo	16 bits stereo
Quantization size	16 bits	16 bits	16 bits	16 bits

Listeners of TTS output follow what is spoken but do not comprehend it. The following test is used to check whether the listener has comprehended what was heard. The better the intelligibility of the TTS output, the better would be the comprehension. A paragraph of TTS output was played to the subjects and they were asked five questions from the paragraph. Based on how many questions the subjects could answer correctly. Intelligibility of the TTS synthesizer can be verified [17]-[21].

TABLE IIIV
RESULT OF COMPREHENSION TEST

Subject	Number of correct answers for five questions
1	5
2	3
3	4
4	4
5	5
Average	4.2

From comprehension test it is clear that: Listener’s comprehension lies at 4.2 hence intelligibility of proposed system is quite good.

VII. CONLUSIONS

In this paper, we discussed the issues relevant to the development of unit selection speech systems for Marathi language. It was observed that when the coverage of units is small, the synthesizer is likely to produce a low quality speech. As the coverage of units increases, it increases the quality of the synthesizer.

A perceptual test reveals that the word unit performs better than the phoneme units, and seems to be a better representation for languages such as Marathi. The evaluation test revealed the effectiveness of this concatenative speech synthesis method, which was also shown to be very effective for reducing the speech synthesis runtime.

In order to achieve greatest naturalness, the areas that need more attention are text analysis, prosody and creation of big speech database for concatenation synthesis.

VIII. REFERENCES

[1] Paul Taylor, "Text-to-Speech Synthesis", University of Cambridge.

[2] T. Dutoit, An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, Dordrecht, 320 pp., ISBN 0-7923-4498-7. 1997.

[3] S. D. Shirbahadurkar, D. S. Bormane, R. L. Kazi, " Experiments with unit selection speech databases for TTS using concatinative synthesis for Marathi language" International Journal of Engg. Research & Industrial Applications, IJERIA, ISSN 0974-1518,vol.2, No. VII, pp 397-410, India.

[4] Jerneja Zganec Gros and Mario Zganec, "An Efficient Unit-Selection Method for Concatenative Text-to-speech Synthesis Systems", Journal of Computing and Information Technology- CIT 16, 2008,1, 69-78 doi:10.2498/cit.1001049.

[5] S. D. Shirbahadurkar, D. S. Bormane, R. L. Kazi, "Subjective and Spectrogram Analysis of Speech Synthesizer for Marathi TTS using concatinative synthesis" ITC, 2010, 978-0-7695-3975-1/10 DOI 10.1109/ITC.2010.76 ©2010 IEEE.

[6] Zeynep Orhan and Zeliha Görmez, "Evaluation of the Concatenative Turkish Text-to- Speech System", 978-1-4244-4131-0/09 © 2009 IEEE.

[7] Branislav Gerazov and Goce Shutinoski, " A Novel Quasi-Diphone Inventory Approach to Text-To-Speech Synthesis", 978-1-4244- 1633-2/08/.00 ©2008 IEEE.

[8] Weibin Zhu and Wei Zhan, "Corpus Building For Data-Driven TTS Systems", 0-7803- 7395-2/02©2002 IEEE.

[9] Min Chu and Chun Li, "Domain Adaption for TTS System", 0-7803-7402-9/02©2002 IEEE.

[10] A. Mukhopadhyay, S. Chakraborty, M. Choudhury, A. Lahiri, S. Dey and A. Basu, "Shruti: an embedded text-to-speech system for Indian languages" IEE Proc.-Softw., Vol. 153, No. 2, 2006.

[11] S.P Kishore, Alan W Black, Rohit Kumar, and Rajiv Sangal, "Experiments with unit selection Speech Databases for Indian Languages."

[12] Aniruddha Sen, "Speech Synthesis in India", IETE Technical Review, Vol 24, No 5, Sep-Oct 2007, pp 343-350.

[13] S. P.Kishore and A. W. Black, "Unit size in Unit selection Speech Synthesis", Proceedings of EUROSPEECH, Geneva, Switzerland, 2003.

[14] S. P. Kawachale and J. S. Chitode , " An Optimized Soft Cutting Approach to Derive Syllables from Words in Text to Speech Synthesizer", in proceedings Signal and Image Processing, 2006, pp 534.

[15] Hiroyuki Segi, Tohru Takagi and Takayuki Ito, "A Concatenative Speech Synthesis Method using Context Dependent Phoneme Sequences with variable length as a Search Units, Fifth ISCA Speech Synthesis Workshop- Pittsburgh.

[16] Kalika Bali, Partha P ratim Talukdar, N.Sridhar Krishna, A. G. Ramakrishnan, " Tools for the development of a Hindi Speech Synthesis System", Fifth ISCA Speech Synthesis Workshop- Pittsburgh.

[17] Eric Lewis and Mark Tatham, "Word and Syllable Concatenation in Text to Speech Synthesis", Proceedings of sixth European Conference of Speech Communication and Technology, pp615-618, ESCA, 1999.

[18] Dan Chazan and Ron Hoory, "Small Footprint Concatenative Text-to- Speech Synthesis System using
Complex Spectral Envelope Modeling", IBM Research Laboratory in Haifa, Israel.

[19] P. Prathibha, A.G. Ramakrishnan, R. Muralishankar, Thirukkural II- A text-to-Speech Synthesis System, Edition, 2003.

[20] E. Raghavendra, S Desai, B. Yegnanarayana, Alan W Black,, " Experiments on Unit Size for Unit selection Speech Synths", Blizzard 2008.

[21] S. P. Kishore, Rohit Kumar, and Rajeev Sangal, " A data – driven synthesis approach for Indian Languages using syllable as basic unit," in Proceedings of International Conference on National Language Processing (ICON), 2002.

[22] Eric Lewis, Mark Tatham and K Morton, " Syllable Reconstruction in Concatenated waveform Speech Synthesis", Proceedings of International Congress of Phonetic Sciences, pp2303-2306, ESCA, z1999.