PHONOTACTICAL CONSTRAINTS STRENGHT CHANGES AS FUNCTION OF INSIDE WORD SYLLABLE POSITION

ALESSANDRO FALASCHI

University of Rome "La Sapienza"
Information and Communication Department
Via Eudossiana 18, 00184 Roma, Italy

ABSTRACT

The aim of the work is to show how much and where the spoken chain is constrained by paradigmatic neighbouring effects. The constraints strength is evaluated in information theory terms by the ratio between the average mutual information of phonetic symbols pairs and the entropy of the phonetic sequence. Positional variation of the symbols predictability is evidenced by evaluating the strength measure along a probabilistic phonetic word model, which takes into account the phonological knowledge about the syllable structure. The parameters of the model are estimated from a phonetic sequence training corpus obtained by automatic translation and syllabic segmentation of a training text data base.

I. INTRODUCTION

Recent advances in automatic speech recognition technology [1][2] have lead to a statistical point of view of linguistic message analysis. Such approach can easily bring to new tools to be utilized in verifying linguistical hypothesis.

Along this study the phonetic message is regarded as a sequence generated by a symbolic information source whose alphabet consists of the italian language phonemes set. As the symbols emitted by a source are not completely predictable they convey information; such quantity will be so much greater as the symbols have little probability. The time average of the information is called entropy, and represents a global information about the source predictability and is measured in bits/symbol. This number represents in fact the average number of binary choices necessary to "guess" the next symbol once the past ones and the statistical behaviour of the source are completely known. It is clear that more constrained are the sequences emitted by the source and smaller will result the entropy.

The phonetic sequences are constrained by many low and high level factors like syllable structure, sintax and semantic: a perfect knowlege of them will bring us to the true entropy of the language. Ignoring some of them will produce higher entropy values, and the difference indicates how much the omitted knowledges conditionates the allowable symbol sequences.

Since our interest is to evaluate the neighbouring effects strength, a comparison between the entropy values obtained supposing known or not the last symbol emitted is done. In order to analyze if and how the phonotactical constraints varies with the position inside a generalized word model, the entropies will be evaluated for every morphological state of the model, giving rise to non-stationary entropy functions. It is to be highligthed that this approach results innovative with respect to a stationary source model [3].

The following section illustrates how the morphological knowledge is organized in a data structure representing the phonetic information source as a markov source. Section III will give the analitical formulas for computing the constraints strenght measure and section IV will illustrate how the phonetic conditional probabilities to be utilized in such formulas are estimated. Finally, the results of the analysis are presented in section V in form of istograms of the computed quantities.

II. THE WORD MODEL

The word source model (WSM) that will be now described is just an economic way to represent the phonetic non stationary conditional probabilities which are needed for evaluating the entropy functions briefly describer above. In oder to take into account the neighbouring effects and the morphological word structure, the phoneme probabilities will be treated as functions of the following four events: 1) Previous phoneme (pp); 2) Syllable number (sn); 3) Syllabic morphological state (ss) 4) Syllable position whit respect to the lexical stress (a).

The dependece upon the syllable number and position with respect to the stress can be modelled by the transition diagram of Fig.1, where $S^0{}_i$ represents a syllabic source model (SSM) for syllable number i which precedes or contain the stressed vowel of a word (a=0), $S^1{}_i$ represents SSMs (a=1) for syllables following the stressed vowel and the arcs between the models individuate the allowable transition between syllables. Note that transitions from $S^0{}_i$ to $S^1{}_{i+1}$ or to the lower bar (representing end-of-word symbol emission) may occur

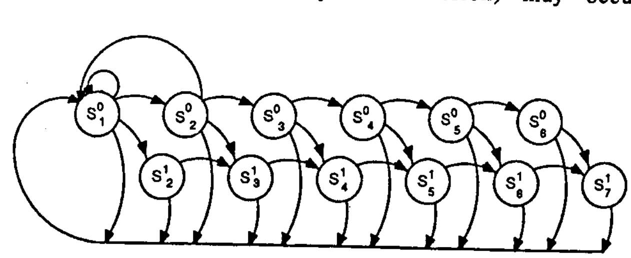

Fig. 1 - Word Source Model main structure

only after a stressed vowel, thus assuming the existence of only one stressed vowel for word. The only exception is for apostrophed words, which are treated as prefixes and modeled by means of the upper backward transition of Fig.1.

In order to represent the syllabic morphology every SSM is then substituted with another transition diagram which exploites the allowable phoneme class sequences as described in [4]; such graphs are drawn in Figg. 2 and 3 for the S^0 and S^1 type SSM respectively. These diagrams identify the italian syllabic structure by means of an outlined syllabe state number (indicated as ss before) and phoneme class label superimposed over the transitions; back transitions for apostrophed words are not shown.

The previous phoneme conditioning is also not represented in Figg. 2-3; the SSM adopted will then have, for any ss shown, as many states as the number of different phonemes (pp) which can be emitted by the WSM while it executes a transition ending in such position.

At this point it should be clear how the four conditioning events have been encoded inside the WSM. It remains to emphasize that the set (a,ns,ss,pp) is in fact a markov source state identifier, whose outgoing transitions are described not only by the phonetic symbol identity but also by the destination state to be reached. In the following the transition probabilities will be indicated as $P(t_i/a,ns,ss,pp)$ where i spaws from 1 to the state outgoing number of transition.

III. CONSTRAINTS STRENGHT MEASURE

As anticipated in Sect. I, the amount of symbol predictability due to the previous one is evaluated by means of the difference between the source entropy H_0 and the conditional source entropy H_1. This quantity is called average mutual information (I_m) of the symbols pairs emitted by the source and it also will be a function of the word morphological state, thus having

(1) $I_m(a,ns,ss) = H_0(a,ns,ss) - H_1(a,ns,ss)$

Let us now considerate the behaviour of I_m as a function of the entropy values. If the knowledge of the last symbol emitted by the source do not increase the sequence predictability the conditional entropy value H_1 results equal to H_0, and the zero value of the mutual information can be regarded as <u>no constraints</u> on the symbols sequence. On the other hand, I_m assumes the maximum value H_0 when H_1 results zero, indicating that the previous symbol knowledge impose an <u>unique choiche</u> for the next one, i.e. a maximum phontactical constraint strength condition. In oder to derive an omogeneous strenght measure along the WSM states, the I_m is normalized by the entropy H_0 thus having the normalized average mutual information

(2) $I_N(a,ns,ss) = I_n(a,ns,ss)/H_0(a,ns,ss)$

whose values spawn between zero and one for the cases of null or absolute sequence predictability.

As the entropy is the expected value of the information emitted by the source and having defined the information received after a transition in the model has taken place as the base 2 logarithm of the inverse of the transition probability, the entropies computation formulas once given the parameters of the model are

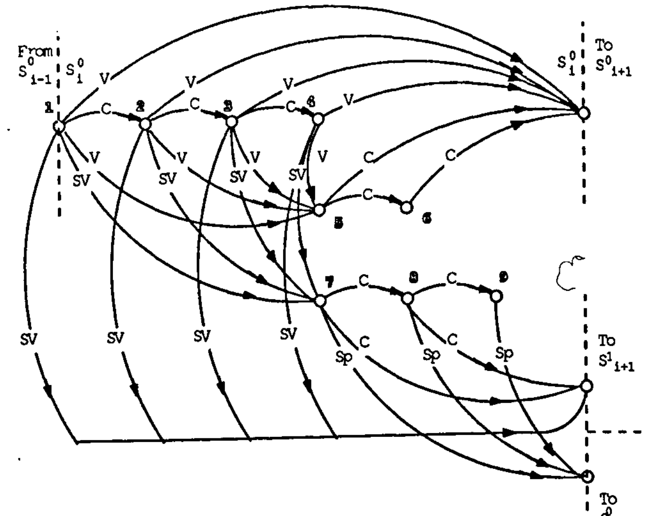

Fig. 2 - Syllabic Source Model transition diagram for S^0_i

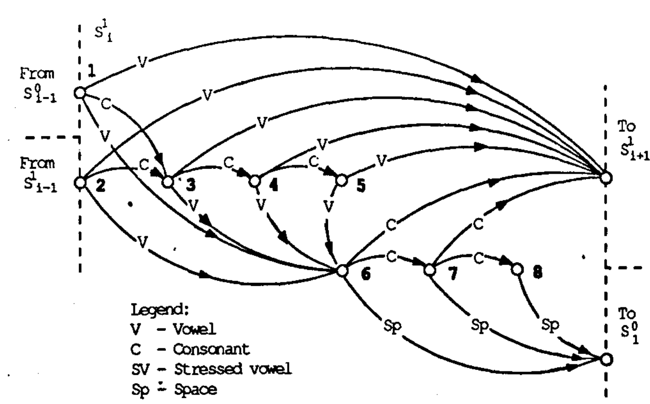

Fig. 3 - Syllabic Source Model transition diagram for S^1_i

(3) $H_0(a,ns,ss) = \Sigma_i Q(p(t_i/a,ns,ss))$

(4) $H_1(a,ns,ss) = \Sigma_j p(pp_j/a,ns,ss)\Sigma_i Q(p(t_i/pp_j,a,ns,ss))$

where $Q(a) = aLg_2(1/a)$.

The non stationary functions (1)-(4) can be statistically averaged over their arguments, in order to point out the dependencies from syllable number or syllable state or stress relative position only, or even a global value. Here below follows the formulas for obtain such values for a generic non stationary function R(a,ns,ss):

$R(a,ns) = \Sigma_{ss}p(ss/a,ns)R(a,ns,ss)$

$R(a,ss) = \Sigma_{ns}p(ns/a,ss)R(a,ns,ss)$

$R(a) = \Sigma_{ss}\Sigma_{ns}p(ss,ns,/a)R(a,ns,ss)$

$R = \Sigma_a\Sigma_{ns}\Sigma_{ss}p(a,ns,ss)R(a,ns,ss)$

IV. MODEL ESTIMATION

This section illustrates the estimation method for the probabilities values that appear in the above formulas by using a knowledge of the world.

A set of 14 text files were selected from newspapers, novels, textbooks, giving a total of 4981 words occurencies and 2073 different lexical items. These files were phonetically transcribed and then syllabized in agreement to the rules described in [4] by a computer program, thus obtaining a set of phonetic word sequences (referred as DB1) whose composition reflects the function and content words frequency of occurence which is proper of the language.

The WSM is then utilized as a parser, individuating the paths inside the model that generate the phonetic sequences belonging to the training corpus. This is a straightforward operation since every word in the lexicon corresponds to an unique path in the model; a simple count of the number of times that each transition has been crossed during the training corpus analysis allows us a maximum likelihood estimation of the probabilities values. In fact the ML estimate can be expressed by the ratio of the counts of the joint event and the conditioning one; the number of times count that a state (or a set of states) has been visited is easily derived by summation over the counts of the outgoing transitions.

As a first result, Fig.4 reports the stationary source entropies H_0 and H_1 variations while the source is being trained with the phonetic sequences. As knowledge is added, the new created transitions makes the WSM a more informative source. But larger is the data base already analysed and lesser will be the probability of observing new events; this consideration motivates the saturation effect visible in Fig.4. This entropies behaviour validates the results that will be given below as significative for our investigation purposes, even if obtained from a relatively modest data base size.

As a side experiment, a secondary data base DB2 was derived from DB1 by elimination of duplicate words, thus obtainig a list of equiprobable lexical items. This list is then used in a separete model estimation for comparison purposes.

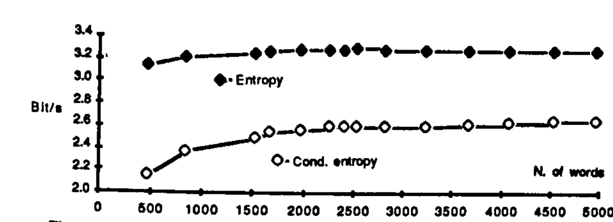

Fig. 4 - WSM entropies as function of the training corpus size

	DB1			DB2		
	a = 0	a = 1	all	a = 0	a = 1	all
H_0	3.86	2.14	3.29	4.05	2.17	3.36
H_1	3.26	1.43	2.66	3.35	1.49	2.67
I_m	0.59	0.7	0.63	0.7	0.68	0.69
I_N	0.15	0.33	0.19	0.17	0.31	0.20
P(a)	0.67	0.33		0.63	0.37	

Tab. I - Stress position dependent and global information values

V. RESULTS

Let us begin to examinate results derived from DB1 analysis. Formulas (1)-(4) were evaluated and the information matricies statistically averaged over their dimensions in order to deal with more readable series.

The first two columms of table I give the entropy and mutual information values as function of the syllable type only; by using the syllable type probability in the last row of table 1 as wheigths in a further averaging the stationary values of column three are obtained. The comparison of the entropy values reveal the syllables which follows the stress are less informative with respect to the type 0 ones (the difference is more than one bit/symbol) and exhibit a phonotactical constraints strenght nearly doubled. This fact is easily explained remembering that in italian the stress is often placed near the end of the lexical root, so type 1 syllables mainly belongs to a closed set of suffixes. Columns four to six of Table I report the values obtained by using DB2 as training corpus. The lexical items equiprobability produces only a slight entropy increase, mainly due the 0 syllable type.

A more accurate feeling on the symbols predictability changes is given in Fig. 5, where the non-stationary functions (1)-(4) expected values with respect to the syllable state number are reported in form of istograms, allowing an easier visual evaluation of the results. Black bars refer to type 0 SSM and grey bars to type 1 ones; Table I values are computed from Fig. 5 ones by means of the SSM probability wheighting reported in Fig 5.e. As an example consider the a) plot, which refers to the WSM entropy $H_0(a,ns)$. Also if the first three type 0 SSMs exhibit nearly equal entropy values, their contribution to the value reported in Tab.I are unequal because of their different frequency of occurence.

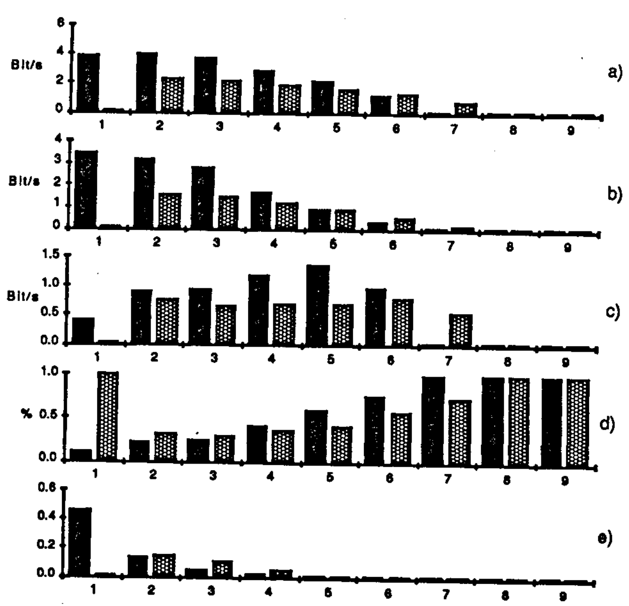

Fig. 5 - Information values as function of the syllable number and type:
a) - Entropy H_0 (a,ns)
b) - Conditional Entropy H_1 (a,ns)
c) - Average Mutual Information I_m(a,ns)
d) - Normalized Mutual Information I_N(a,ns)
e) - Syllable Probability P(a,ns)

By looking to Fig 5.a, it can be noted that the type 1 SSMs entropy is quite constant, while type 0 SSMs entropy decrease more quickly after syllable number three; about the same effect can be noted also for H_1 (Fig 5.b). This result can be explained by thinking that the statistical composition of syllables which follow the stress is nearly the same for every syllable number because it originates from the same suffixes joined with word roots of different lengths. For what regars type 0 SSMs entropy decrease, a motivation can be the word roots progressive fading in SSM statitical composition with the syllable number increase, making room for the influence of a closed set of inflectional morphemes.

Figg. 5.c and 5.d shows the average mutual information values, and it is possible to appreciate how the normalization of I_m whith respect to H_0 is important in giving the exact measure I_N of the constraints strenght variations. The neighbouring effects become strongher in a linear fashion as function of the syllable number for both the syllable types. The unitary value means an absolute predictability, and this is obvious for syllable numbers to which no word in the training corpus has given contribution!

The use of DB2 as training corpus do not give much additional information; just an entropy increase can be noted, mainly for the initial syllables whose statistic is heavily influenced by the high frequency short words.

As a final analysis it is interesting investigate on the constraining power of the SSM states as they are evidenciated in Figg. 2-3; for this purpose the expected values with respect to the syllable number of the functions (1)-(4) were calculated, obtaining the plots of Figg. 6-7 for syllable types 0 and 1 respectively.

Fig 6.a shows the values of the conditional entropy: it is possible to note an information decrease among the state numbers which are destinations of the consonants belonging to the initial consonant cluster (states number 2-4), while their constraining power shown in Fig 6.b by means of I_N(a=1,ss) is nearly equal. The state following an unstressed vowel (#5) is sligthly more informative than the one following a stressed vowel (#7), and exhibits a predictability just a little bit weaker than the latter. For what regards the after-vowel consonant cluster, its information contribute is very low. Fig 6.c reports about the probability of occupancy of the SSM states, which is in agreement with the high frequency of occurence of short syllables.

The states numbering of Fig. 7 reflects the type 1 SSM morphology of Fig. 3. The entropy and predictability distributions are given in Figg. 7.a and 7.b; the states probability values of Fig. 7.c evidenciate the great prevalence of CV syllable structure for type 1 SSMs.

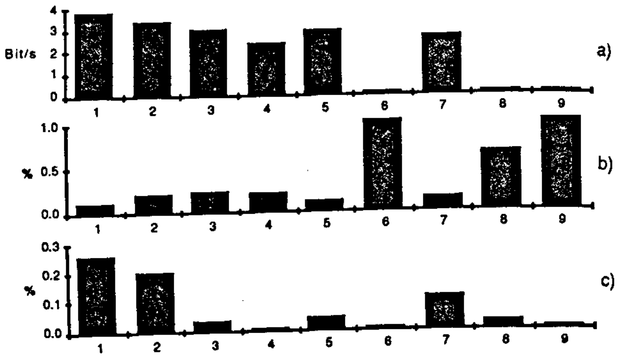

Fig 6 - Information values as function of the state number for syllable type 0

a) - Conditional entropy H_1 (0,ss)

b) - Normalized mutual information I_N (0,ss)

c) - State probability P(0,ss)

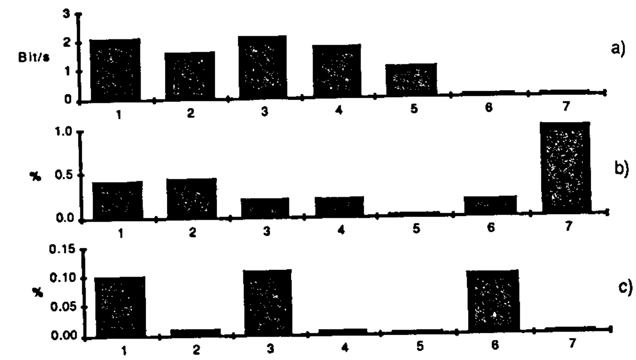

Fig 7 - Information values as function of the state number for syllable type 1

a) - Conditional entropy H_1 (1,ss)

b) - Normalized mutual information I_N (1,ss)

c) - State probability P(1,ss)

VI. CONCLUSIONS

The word and syllable structures driven non stationary statistical analysis of phonetic chains has highlighted a heavy disuniformity in the phonotactical constraints strenght as a function of the inside word syllable position. Although this effect could be foreseen from morphology, its quantification can result very useful in the area of automatic speech recognition for very large lexicon systems. Future work will be addressed towards the use of the WSM as a language model for an automatic phonetic recognizer.

REFERENCES

[1] - L.R. Bahl, F. Jelinek, R.L .Mercer, "A Maximum Likelihood Approach to Continuos Speech Recognition", IEEE Trans. PAMI-5, No.5, March 1983

[2] - S.E.Levinson, "Structural Methods in Automatic Speech Recognition", Proc. IEEE, Nov. 1985

[3] - J.P.Tubach, L.J.Boe, "Quantitative Knowkedge on Word Structure, from a Phonetic Corpus, with Applications to Large Vocabularies Recognition Systems", Proc. ICASSP 86, April 1986, Tokyo

[4] - R.A.Hall Jr., "La Struttura dell'Italiano", Armando Armando Ed., 1971 Roma