A SPEECH UNDERSTANDING SYSTEM WITH LEARNING CAPABILITY

R. De Mori - S. Rivoira - A. Serra
Istituto di Elettrotecnica - Centro per l'Elaborazione Numerale dei Segnali - Istituto Elettrotecnico Nazionale Galileo Ferraris - Politecnico - Corso Duca degli Abruzzi, 24 - 10129 TORINO (Italy)

Abstract

A speech understanding system with learning capabilities is presented.

Its relevant aspects are:

a) The spoken sentence is represented concisely by a description that can be used to reconstruct the sentence and to verify whether its meaning was not degraded by the coding.
b) Syllables or broader coarticulation segments are the smallest units.
c) The evaluation of an hypothesis is based on the probability that a syllable, a word or the sentence can generate the spectrogram of the spoken message and that a syntactic structure can generate its pitch contour.
d) Coarticulation effects are described in terms of pattern grammars, generating all the possible formant trajectories for a given utterance.
e) Spectral and prosodic features can be learned by inference of stochastic finite-state-automata.
f) More formant choices are allowed for a single syllabic segment and an algorithm is provided for assigning to each choice a probability.

Introduction

Large efforts have been devoted in the last years to the design of speech understanding systems (SUS) [1] ÷ [3]. These efforts have revealed many aspects of the problem that are still open and whose solution is a necessary condition for the success of such designs.

The purpose of this paper is to present a framework in which some of the above problems can be solved with successive refinements. Particular care is devoted to the problem of emitting and verifying hypothesis at the syllabic level, this takes *into account*, as much as possible, the knowledge about the structure of speech that is available to a phonetician performing a visual inspection of spectrograms.

On the line of a previous work [3], coarticulation effects are taken into account by a grammar of speech that is inferred by an automatic procedure.

Concepts from phonetics, syntactic pattern recognition theory, grammatical inference, stochastic automata, information theory, identification and modelling, and natural language processing theories are used in this project to characterize sources of knowledge and data bases at different levels of the processing, namely parametric, acoustic, phonological, lexical, syntactic and semantic.

The most relevant and original aspects of this project are the following:

a) The spoken sentence is represented concisely by a description that can be used to reconstruct the sentence and to verify whether its meaning was not degraded by the coding.
b) Syllables or broader coarticulation segments are the smallest units.
c) The evaluation of a hypothesis is based on the probability that a syllable, a word, or the sentence can generate the spectrogram of the spoken message and that a syntactic structure can generate its pitch contour.
d) Coarticulation effects are described in terms of pattern grammars generating all the possible formant trajectories for a given utterance.
e) Spectral and prosodic features can be learned by inference of stochastic finite-state-automata.
f) More formant choices are allowed for a single syllabic segment and an algorithm is provided for assigning to each choice a probability.

The organization of the system and the hypothesis evaluation algorithm, operating at the syllabic level, allow one to recover information lost due to errors in segmentation, spectral estimates, formant tracking and formant coding.

The system has been designed also to be used to investigate some problems related to fields other than SUS, like speaker characterization, investigation of phonetic rules in continuous speech and modelling of human speech perception. It has to be considered more an instrument for research than a commercial implementation of a SUS.

The acoustic level

At this level the information about the spoken sentence is the time waveform. It is sampled at 20 kHz and processed for extracting the amplitude and pitch, using an algorithm based on comb-filtering and pattern recognition. Furthermore some global features are extracted by a bank of digital filters, operating in real-time.

The time evolutions of such parameters are described by a very simple language that is processed by a set of finite-state automata (FSA) that locate silences (SL), unvoiced tracts (UT), vowels (V) and tracts of successive voiced consonants (VC). The details of the detection of such "elemental fragments" are reported in [3] and the description of the special purpose that performs digital filtering in real-time can be found in [4].

The sequence of elemental fragments is then processed by a simple FSA that locates the speech segments, called pseudo-syllable-segments (PSS), where coarticulation is expected to be more predominant. Notice that the PSSs can overlap.

The preliminary segmentation has two tasks.

1 - It delimits the voiced tracts of the speech waveform on which the spectra are computed, using pitch-synchronous linear prediction, while for the unvoiced tracts an asynchronous FFT is performed.
2 - The preliminary segmentation delimits the fields of spectrograms corresponding to PSSs, in which possible formant paths are specified and organized in a multilinked data structure. Furthermore this segmentation anchors on the steady-state portions of vowels, the starting points of the estimates of the best possible formants.

These formants are used to emit the first hypothesis about the possible syllables present in the spoken sentence, while the request for a hypothesis verification from higher levels results in a searching for paths in some specific allowed passages. This search is carried out in the multilinked data structure and if possible paths are found, their probabilities are also computed.

A node in the data structure contains information about a frequency interval of a given spectrum, where a concentration of energy and the maximum of this energy have been found.

The organization of the data structure and the algorithms for formant tracking are reported in [5]. The way in which probabilities are assigned will be described briefly at the end of this paragraph. The possibility of performing pitch-synchronous FFT on the voiced tracts is also allowed.

The difference between FFT and linear prediction (LP) in this application is that FFT generally introduces many spurious paths in the spectrogram, while LP requires a careful operation of gap-filling; both the aspects are treated by the formant tracking algorithm, but one method can be better or poorer than the other, depending on the waveform.

Formants, amplitude and pitch are presented at the higher levels as a list of descriptions; the form of these descriptions are introduced in the following account.

Amplitude description (ADES)

The amplitude description has the primitive forms represented by the following non-terminal alphabet:

$$V_1 : \{BLT,\ C1,\ C2,\ C3,\ C4\} ;$$

These forms are related to a terminal alphabet containing three symbols representing a linear aproximation of the amplitude curve; the terminal alphabet is:

V_2 = {h: horizontal tracts, a: ascendent tracts, d: descendent tracts} ;

and the composition rules are:

(BLT) = (C3)* (C1) (C4)*

C1 = ad + ahd

dC2a = dha

C3a = aha

dC4 = dhd

The compositions are performed by a simple finite-state automaton, operating in a time close to the real-time.

Each primitive symbol is followed by four attributes: the duration, the amplitude of the first point, the abscissa and the amplitude of the maximum. These parameters allow one to reconstruct the amplitude waveforms with a polynomial approximation.

Plich description (PDES)

Plich description is very similar to the amplitude description, except that a new symbol is provided for the silences or the unvoiced tracts.

Description of spectral features for an entirely voiced PSS

The description SDES of an entirely voiced PSS is characterized by the following rules:

$$SDES \rightarrow \beta\beta^* \quad ,$$

$$\beta = F1T + F2T + F3T + F4T + F5T \quad ,$$

where F1T is a tract where i formant lines have been detected.

The descriptions of such tracts are:

F1T = (F1D) (A1D) (b)

F2T = (F1F2D) (A1A2D)

F3T = (F1F2D) (A1A2D) (F3D) (A3D)

F4T = (F1F2D) (A1A2D) (F3F4D) (A3A4D)

F5T = (F1F2D) (A1A2D) (F3D) (A3D) (F4F5D) (A4A5D)

Where FhD is the description of the time evolutions of the h-th formant frequency,

hD is the description of the time evolutions of the h-th formant amplitude;

(b) is a set of symbols describing the 1 ÷ 10 kHz spectrum of voiced fricative and voiced stop sounds;

FiFjD is the description of the evolution of F_i and F_j in the $F_i - F_j$ plane,

AiAjD is the description of the evolutions of the amplitudes A_i and A_j in the $A_i - A_j$ plane.

The descriptions have two types of primitive forms, namely lines and stable zones.

These primitives are detected using an algorithm presented in [6] A stable zone corresponds to a quasi-stationary portion of the speech waveform and is represented by a symbol S with the following attributes: the duration and the gravity center coordinates of the points belonging to the stable zone.

The lines are represented in the SUS by one of eight slope symbols, k, l, m, n, o, p, q, r, corresponding to angular sectors of 45°, followed by two numerical attributes: the duration and the line length.

The choice of the maximum allowable error in linear approximation is a compromise between compression and fidelity. This value certainly need not fall below the limens of the perception of formants, that is 3% of the formant values.

The maximum error has been selected in order to obtain the detection of a single stable zone for all the non reduced vowels in a corpus of 50 sentences. Values of 70 Hz for the first formant, 150 for the second and 300 for the third, have been found.

Assuming a gaussian distribution of the errors over the stable zone, a total average value of the errors over the three formant regions of 5% is obtained. Similar results are obtained for the lines.

Silence and unvoiced tracts description

Silences are described by a symbol SL and its duration.

To represent the unvoiced tracts, the symbol UT is used, followed by a concise description of an average spectrum of the unvoiced segment. For separating two successive fricatives an analysis of the spectral derivate is performed, according to previous experiences in the segmentation of continuous speech.

Nasals and nasalized vowels often generate a segment of spectrogram with more than three formants. The possibility of nasalization depends on frequencies and amplitudes of the first formant. For these segments for which this possibility is detected, a more accurate research for clusters or other formant lines is started.

Two tracts having a different number of formants can be concatenated in different ways, these situations are described introducing predicates between formant descriptions [3].

The descriptions obtained in this way allow one to reconstruct and listen to the spoken sentence and to evaluate if the sentence still preserves its meaning.

Probability assignment

The influence of segmentation affects only the preliminary emission of hypotheses about possible phonetic transcriptions of PSSs. In this operation and in the complementary operation of hypothesis verification it is important to know the probability $P(d_{ij}/v_i)$ that the description d_{ij} truly represents the segment v_i of the spectrogram of the spoken sentence.

The subscripts of such parameters are to be interpreted as follows:

d_{ij} : is the j-th possible description of the i-th spectral segment,

$i = 1, 2, \ldots, L$,

$j = 1, 2, \ldots, I_i$,

where L is the number of segments evidenced on the spectrogram at a certain instant of the understanding,

I_i is the number of possible descriptions of the i - th segment.

The probability $P(d_{ij}/v_i)$ will be used in the next paragraph for the evaluation of a syllabic hypothesis.

Let v_i have G formants; let $F_1, F_2, \ldots, F_G$ be the formants of v_i ; thus:

$$P(d_{ij}/v_i) = P(F_1/v_i)\ P(F_2/v_i F_1) \ldots P(F_G/v_i F_1 \ldots F_{G-1}) \qquad (1)$$

Let F_1 be the concatenation of ν_1 paths $\pi_{11}, \pi_{12}, \ldots, \pi_{1\nu_1}$, where each path ends at a node where it can be followed by more than one way. Then one has:

$$P(F_1/v_i) = P(\pi_{11}/v_i)\ P(\pi_{12}/\pi_{11} v_i) \ldots P(\pi_{1\nu_1}/\pi_{11} \ldots \pi_{1\nu_1-1} v_i) \qquad (2)$$

The generic conditional probability in (2) is the probability that given a path $\pi_{11}\pi_{12}\ldots\pi_{1\nu}$ on v_i for the first formant, this path continues with $\pi_{1\nu+1}$.

This probability is evaluated considering the average amplitude, the number and duration of gaps and how far it is possible to go following $\pi_{1\nu+1}$ instead of the other possible directions.

With relations similar to (2) the other conditional probabilities in (1) can be computed, remembering that, given the first f formants, the (f + 1)-th should be in a proper frequency range and above the first f formants. This approach allowes one to use the information about formant evolutions that have been proven very significant by many perception experiments. Several possible formant patterns and their probabilities can be obtained avoiding that an error in formant tracking compromise irreparably the success in recognition.

The syllabic level

At the syllabic level the information about the spoken sentence is the description d_{ij} of the time evolutions of formants and amplitudes, together with the probabilities $P(d_{ij}/v_i)$ of the description d_{ij}.

The a-priori information source for the syllabic level is a grammar of speech generating a language recognized by a set of stochastic - finite-state automata (SFSA), with some auxiliary units that will be described later on. The formant patterns corresponding to a given VCV utterance pronounced in various contexts cannot be random patterns, they have to respect some rules underlying the relations between the articulatory commands and the spectra of the generated waveforms. On the other hand, the patterns are not equal because some distortions on the archetypes do not alter the interpretation given by a listener to the original utterance and it is important to know

the statistics of such distortions. Thus the SFSA are a proper tool for representing the possible patterns of a given utterance. Their implementations and their learning are also feasible because the segments considered are short in time and their possible patterns are described by a concatenation of few symbols belonging to a very small vocabulary.

Two operations are performed at the syllabic level, namely hypothesis emission and verification.

At the beginning of the understanding, the segmenter represents the spoken sentence as a continuous sequence of PSSs. This sequence is the first segmentation hypothesis. Then for the voiced portions of each PSS, the most likely formants are extracted and a description is generated. The descriptions are processed under the control of a grammar of speech and are translated into possible phonetic transcriptions, with associated the probability $P(S_j/u_i)$ that the syllable S_j corresponds to u_i, the i-th segment of the spectrogram of the spoken sentence.

The grammar of speech is a stochastic grammar representing the possible patterns for each coarticulation instance, corresponding to the concatenations of elemental fragments generated by the segmentation grammar.

The starting symbol of the grammar of speech is PSS; the terminal alphabet contains all the symbols with which the descriptions are made; the non-terminal alphabet contains all the possible concatenations of phonemes for which coarticulation affects the formant patterns even after the description approximations. The non-terminal alphabet contains also the symbols emitted by the auxiliary units preceding each SFSA, and acting as translators of the descriptions made of symbols, attributes and probabilities into symbols and probabilities, provided that some relations hold between the attributes. Finally, the productions of the grammar of speech are right-linear stochastic rewriting rules, derivable from the SFSA inferred by the procedure that will be presented in the next paragraph.

The operations of hypothesis emission and verification contain hypothesis evaluation as a component. thus this operation will be described first.

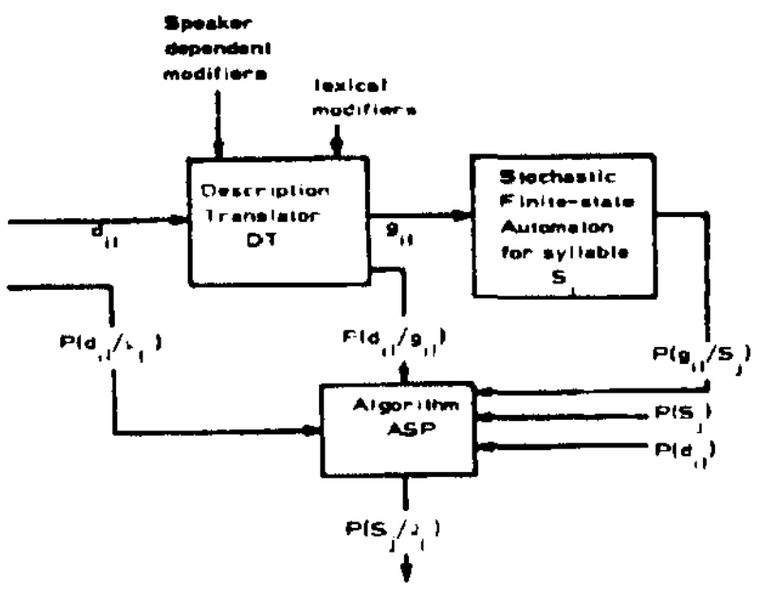


Fig. 1 - Scheme for hypothesis evaluation.

Hypothesis evaluation

Hypothesis evaluation la performed with a procedure whoae block diagram is shown in fig. 1 ,

When a syllable or a coarliculalion segment la hypothesized on a certain portion of lhe spectrogram, tha SFSA of that segment and the corresponding description translator are built up from the grammar of speech.

This operation is simple and fasi due to the straightforward ratalloni between the productions and ihe automata. The syllable description is processed by the description translator (OT). The DT translates a primitive form, with its attributes to describe correctly some torment evolutions, into a symbol and its associated probability. The symbol is emitted only if some conditions on the attributes are verified. These conditions are stored for a talker *mna* a typical lexical position of the segment and can be modified by soma epea war depeitdent modifiers, acting mainly on the form ant loci for the etable-xones constraints. In addition, some lextcal- dependent modifiers, acting mainly on lhe durations , change lhe constraints in accordance with the stress of the salient. The design of such modifiers that also involves the knowledge of probability di«ir ibutions Is actually limited to spactral loci end ttia influence of the stress on the durations, pitch and amplitudes. A lot of investigations are still to be done for characterizing the individual influence on formant evolution and loci for syllables embedded in continuous speech and the approach followed here should be considered a first approximation. The input to the DT is the description d_{il}. The output of the DT is a string g_{il} of symbols without attributes obtained from d_{il} which is a string of symbols with attributes. together with g_{il}, the probability $P(d_{il}/g_{il})$ is obtained. Finally, the g_{il} is processed by the SFSA and, if the g_{il} is recognized, the probability $P(g_{il}/S_j)$ is given. All the obtained probabilities are processed by the following algorithm ASP, that gives $P(S_j/u_i)$.

Algorithm ASP

Remembering that the syllable S_j is the union of all the descriptions recognized by the SFSA, i.e :

$$S_j = \bigcup_{k=1}^{k_j} g_k^j \qquad (3)$$

where $G^j = \{g_1^j, g_2^j, \ldots, g_k^j, \ldots, g_{kj}^j\}$ is the set of all the descriptions recognized by the SFSA of the syllable S_j; the following relation can be written :

$$P(u_i/S_j) = \sum_{k=1}^{k_j} P(u_i/g_k^j) \frac{P(g_k^j)}{P(S_j)} \qquad (4)$$

Combining (3) and the following relation:

$$\frac{P(g_k^j)}{P(S_j)} = \frac{P(g_k^j \cap S_j)}{P(S_j)} = P(g_k^j / S_j) \qquad (5)$$

one gets :

$$P(u_i/S_j) = \sum_{k=1}^{k_j} P(u_i/g_k^j)\ P(g_k^j / S_j) \qquad (6)$$

where $P(g_k^j / S_j)$ can be obtained at the output of the SFSA when g_k^j are recognized.

Let now :

$$D_i = \{d_{1i}, d_{2i}, \ldots, d_{t_i i}\}$$

be the set of the possible descriptions of the spectral segment u_i and

$$D_{ki}^j = \{d_{k1_1}^j, d_{k1_2}^j, \ldots, d_{k1_{M_{kj}}}^j\} \subseteq D_i$$

the subset of descriptions of u_i translated by the DT of the syllable S_j into g_k^j. Let M_{kj} be their number.

Let D_k^j be the set of all the possible descriptions that can be translated into g_k^j; let N_{kj} be their number. As the descriptions are disjoint elements one can write:

$$P(u_i/g_k^j) = \sum_{n=1}^{N_{kj}} P(u_i/d_{kn}^j)\ P(d_{kn}^j / g_k^j) \qquad (7)$$

but:

$$P(u_i/d_{kn}^j) = \frac{P(d_{kn}^j/u_i)\ P(u_i)}{P(d_{kn}^j)} \qquad (8)$$

The descriptions obtained by u_i must belong to the set D_i and those that are mapped into g_k^j and are obtained by u_i must belong to D_{ki}^j ; thus (7) reduces to:

$$P(u_i/g_k^j) = \sum_{m=1}^{M_k} \frac{P(d_{k1m}^j/u_i)\ P(u_i)}{P(d_{k1m}^j)} \cdot P(d_{k1m}^j/g_k^j) \qquad (9)$$

where the $P(d_{k1m}^j/g_k^j)$ are learned with the SFSA of the syllable S_j and the $P(d_{k1m}^j/u_i)$ are obtained from the descriptor.

Using the relations obtained so far, the evaluation of the hypothesis that the syllable S_j is the trascription of the portion u_i of the spectrogram, can be computed as follows :

$$P(S_j/u_i) = \frac{P(u_i/S_j)\ P(S_j)}{P(u_i)} =$$

$$= P(S_j)\left\{\sum_{k=1}^{k_j} P(g_k^j/S_j)\left[\sum_{m=1}^{M_k} \frac{P(d_{k1m}^j/u_i)}{P(d_{k1m}^j)} P(d_{k1m}^j/u_k^j)\right]\right\} \qquad (10)$$

In practice, it is expected that often only one description d_{i1} is recognized and translated by the DT into a g_{i1} that is recognized by the SFSA; in this case the hypothesis is simply evaluated as follows :

$$P(S_j/g_{ij}) = P(S_j)\,\frac{P(d_{ij}/g_{ij})}{P(d_{ij})}\,P(d_{ij}/g_{ij})\,P(g_{ij}/S_j) \qquad (11)$$

where the three conditional probabilities are given respectively at the outputs of the descriptor, the DT and the SFSA.

Let the description d_{ij} be recognized as PSSs $S_1, S_2, \ldots S_j, \ldots, S_w$. The probability $P(d_{ij})$ can be computed as follows :

$$P(d_{ij}) = P(d_{ij}/S_1)P(S_1) + \ldots + P(d_{ij}/S_j)P(S_j) + \ldots + P(d_{ij}/S_w)P(S_w) \qquad (12)$$

because $S_1, S_2, \ldots, S_j, \ldots, S_w$ are disjoint sets.

Each addend in (12) is obtained from the a-priori knowledge of the system, as follows :

$$P(d_{ij}/S_j)P(S_j) = P(S_j/d_{ij})P(d_{ij}) = P(d_{ij}/g_{ij})P(g_{ij}/S_j)P(S_j) \qquad (13)$$

The probabilities in (13) are the a-priori informations that need to be learned. There are several possible approximations for $P(d_{ij}/g_{ij})$ that correspond to different reductions in the computation of (11) and (12).

The simplest approximation consists in considering $P(d_{ij}/g_{ij})$ independent from j. A better approximation, that seems to be very realistic, consists in considering the probability of an attribute of d_{ij} uniform over the range allowed for this attribute, in order to translate d_{ij} into g_{ij}. In this case, one need learn only the intervals allowed for the attributes of d_{ij}. Finally, the exact distributions of the attributes of d_{ij} over the allowed intervals of the DT could be learned and used to compute $P(d_{ij}/g_{ij})$. This last approach would require a very large number of experiments, probably without greatly improving the system performances.

Emission of hypotheses

When a PSS has been described, its description is processed in order to decide whether or not it belongs to the language generated by the grammar of speech. The grammar of speech is ambiguous because a description can be generated by many syllables; thus it may be recognized by different SFSAs. This is one of the reasons why the linguistic approach previously followed [3] has been extended now to the field of stochastic grammars. In practice, it is time-consuming to try to find if the description of a PSS is recognized by any among all the SFSAs of the grammar of speech. For this reason, the PSS description is preprocessed by the FSAs shown in fig. 2. The outputs of these automata define the sub-set of SFSAs that may accept the unknown description.

Only this subset is used to process the input description, the syllable corresponding to those SFSAs that accept the description with an associated probability is emitted as hypothesis of the syllabic level. The set of probabilities emitted by the SFSAs, allows one to compute the total probability that the description belongs to the language generated by the grammar of speech.

From the PSS description, a first formant description (FFD), a second formant description (SFD) and the amplitude description (A1A2DES), are extracted and processed separately.

The automata of fig. 2 are deterministic.

The first automaton (FBA: front-back automaton) processes FFD, recognizes if there is one or two vowels in the PSS and if they are front (F) or back (B).

The second automaton (HMLA: high, middle, low automaton) recognizes if the vowels are high (H), middle (M) or low (L).

The third automaton recognizes if the consonant between the vowels is nasal (N), voiced stop or fricative (VSF), liquid or glide (LG), or other (O). The automaton may recognize more than one of such classes because of the ambiguity of the consonant segment or of the presence of two consonants.

Such automata are learned without probabilities, using the same procedure presented in the successive paragraph.

Verification of hypotheses

When an hypothesis about a syllable S_j has to be verified on a certain portion of the description, the description is processed by SFSA of S_j ; if a portion of the description is recognized, the recognition probability

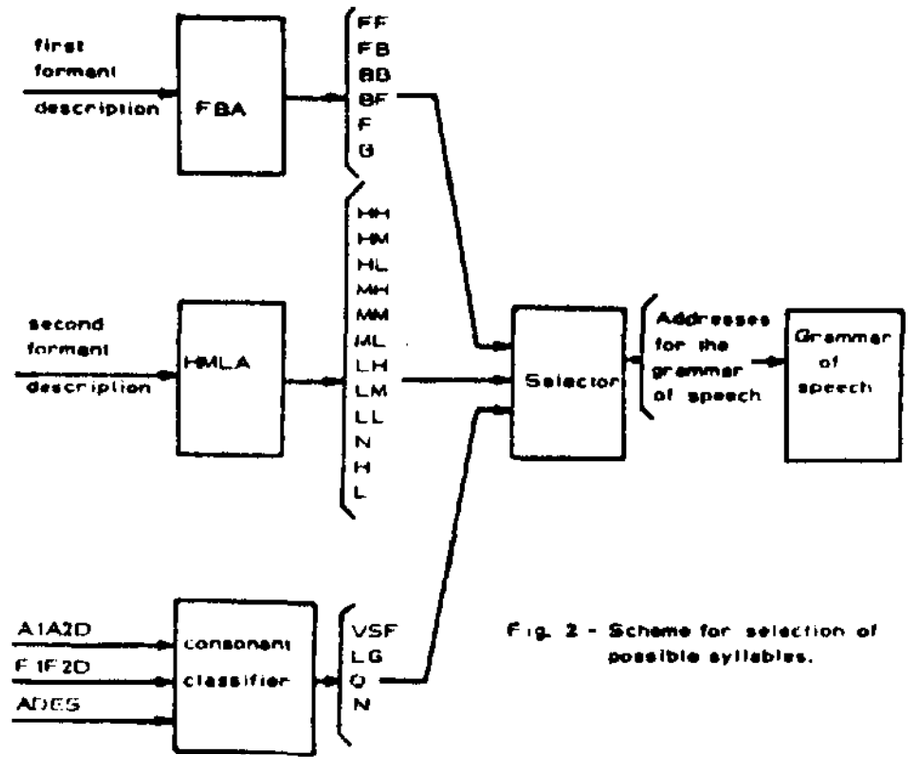


Fig. 2 - Scheme for selection of possible syllables.

is emitted as a measure of the verification.

Otherwise, possible insertions, omissions or substitutions on the description are considered in order to recognize the syllable.

The description modifications are then verified in the spectral data base and if they are found possible, their probabilities are computed and the set of modifications which make the highest probability of recognition of the syllable is accepted. The corresponding probability is assumed to be the measure of the verification. If these modifications alter the previous segmentation, new hypotheses are generated at the syllabic level.

If some of the description modifications are found to be inconsistent with the spectral data base, a loss function is computed that increases with the units of information of the events that are not received at the proper time by the SFSA. If many solutions are possible, the one corresponding to the minimal loss is chosen.

Learning of syllabic features

Learning is accomplished through the construction of a stochastic finite-state automaton for a finite set of stochastic strings of symbols describing the same syllable [8].

Every string with probability of being different from zero is generated by the grammar associated with the automaton, while none of the other strings can be generated.

The samples of information are presented sequentially to the inference machine, which determines a stochastic finite-state automaton after the t-th observation on the basis of the t-th sample and of the automaton generated after the (t-1)-th observation.

The probability of generation associated with each string is obtained by the relative frequency of occurrence and is approached as t approaches infinite.

The learning system is based on an interactive procedure with a human trainer. The teacher checks for the correspondence of the string to an acceptable phonetic description of a given syllable and decides if it must or must not be learned by the inference machine.

He can be helped to decide by listening to the synthesis of the syllable obtained with the phonetic description parameters.

The stochastic finite-state automaton satisfies two conditions: it is unambiguous and it has the minimal number of states. A straightforward procedure generates from it a right-linear grammar with a minimal number of non terminal symbols.

The grammar inferred can be continuously updated during recognition by adding new states and transitions to the SFSA if a configuration is not recognized, or by refining the probability assignments if the configuration is recognized. The SUS can in this way update continuously its knowledge. A measurement of the stability of this knowledge is obtained considering a syllable as an information source; the possible spectral configurations of the syllable are the messages emitted by the source. As the probability of each message can be obtained by the SFSA, the entropy of the source can be measured in

different moments of the learning. When the entropy remains quasi-stationary after a large enough number of experiments and no new configurations are learned, a good degree of knowledge has been reached.

The learning of syllable configurations is performed in accordance with the theory of mechanical inference of SFSAs /7/. The following aspects clarify why the learning of syllabic features is practically feasible and do not require a large amount of memory and computation.

1) The strings to be learned have a limited length. Neither recursion nor concatenations of the same symbol can be present.
2) The terminal alphabet has few symbols.
3) The syllable patterns should respect the phonetic structure of the syllable. Thus the variety of descriptions accepted by a syllable automaton cannot be very large, even if the number of patterns can be very large. In fact each line can be generated by concatenation of elemental arcs connecting two points and the number of such arcs can be largely variable.

A detailed description of the inference procedure is described in another paper [8] and is only summarized in the following.

Let $\Gamma^j_S(n)$ be the SFSA of the syllable S_j after n samples have been learned. Let $\Gamma^{j'}(n)$ be the corresponding non-stochastic automaton. Because of the absence of recursions in $\Gamma^{j'}(n)$ it is easy to construct a corresponding regular expression $R^j(n)$. Let $I(n+1)$ be the information to be learned because it is not recognized by $\Gamma^{j'}(n)$. A new regular expression:

$$R^j(n+1) = R^j(n) + I(n+1)$$

is considered and the new automaton $\Gamma^{j'}(n+1)$ is constructed, considering the derivatives of $R^j(n+1)$ with respect to all the first symbols, the derivatives of the derivatives with respect to their first symbols and so on. Then the probability distribution on the automaton is computed. This operation is performed easily, because for each transition in the stored automaton the total number of times this transition has been executed is stored with the transition itself. The occurrencies of the transitions in the new automaton are computed from the knowledge of the old ones.

The state minimization problem is simplified by the fact that no recursions are allowed. Thus all the states from which the same final state is reached, with the same strings of length l and with the same probability, are equivalent and can be merged. This consideration can be extended to merge the states from which the same state is reached with same strings with the same probability.

With the above considerations, the research of equivalent states can be performed sequentially, going from the final state back to the initial one considering tails of strings of increasing length. Fig. 4 shows the automaton of the pseudo-syllable "uBA"; the operation of the corresponding translator is summarized in table A1.

When the automaton corresponding to a given PSS has been learned an algorithm described in [8] is applied to check if the considered PSS can be subdivided into acoustically independent units. Furthermore, many automata, corresponding to syllables whose phonemes differ for few distinctive features, have considerable portions in common (e.g. the formant amplitude pattern). These automata can be reduced to a single one where some translations or transitions are valid only for some syllables. The structure of the source knowledge at the phonetic level is characterized by stochastic rules that will be refined during learning when more insight in the properties of the auditory patterns is gained. Finally, the stochastic rules partially refer to different possibilities of articulation and partially to the noise introduced by the algorithms that transforms the speech waveform into auditory patterns. The noise effects are probably independent from the particular syllable compositions and could be characterized by a single grammar induced by error trasformations using an algorithm proposed by Fung and Fu [9].

The lexical, semantic and syntactic levels

These levels have been conceived in a simple way and will be described briefly. The protocol used so far was introduced only for experimental purpose and refers to an automated reservation and information system for travels. The vocabulary is actually of about 200 words and the syntax and semantics are described by simple networks. The semantic network contains a small set of key words for each node and issues the main requests for hypothesis verification.

The lexicon contains a stochastic finite state automaton for the root of each word, corresponding to the different ways the word can be pronounced. The probability assignment is strongly dependent on the speaker and his dialect; the possibilities for each word are few because the Italian Language does not allow a large variety of realizations.

The different pronunciations taken into account refer to utterances that are perceived as clearly different during the learning of such rules. Problems of alterations in word concatenation are also very limited in the Italian Language.

The syntactic network refines the hypotheses in order to create or recognize a sequence of words that is grammatically correct. Furthermore, a stochastic automaton, that is learned by experiments, gives the probability of having a certain description of the pitch contour, given a syntactic form.

The hypotesis about a sentence is evaluated by the probability of that sentence, given the spectrogram and the pitch contour. If the sentence does not match the spectrogram or the pitch contour, a general loss function is evaluated, proportional to the information carried by the components of the sentence that are not matched. Generalizing such approach requires an evaluation of the information carried by the various elements of a spoken sentence; the probability of such elements (syllables, words, phrases) depends on the context.

The likelihood function and the loss functions are the main factors on which the control strategy is based. Given a certain knowledge about a spoken sentence and a set of syllables detected with a proper probability, the hypotheses to be verified first are those corresponding to the maximum likelihood and/or the minimum loss among the possibilities that are semantically and syntactically correct. A successful verification of an hypothesis can alter the segmentation and the set of detected syllables on which the choice of the successive verification is based.

Conclusion

A project for SUS has been presented. Some methods have been introduced for evaluating rigorously an hypothesis made on a spectrogram or on a pitch contour.

It became evident that a lot of experimental work has to be done in learning spectral and prosodic features and in correlating them with the structure of the language. A tool for this purpose has been provided.

A sistematic application of the proposed algorithm for mechanical inference will help in investigating how the syllabic features can vary for a single speaker, how they are affected by the context and in which cases they can be subdivided into shorter segments indipendent each other or when segments larger then syllables have to be considered as units for speech understanding. In fact the elements of a description, that correspond to primitive spectral forms, have a probability that depends on the previous elements like in a Markov process, but it is not known how many should be these previous elements.

Certainly a special purpose hardware able to allow a parallelism of many simple operations will be suitable for carrying enough experiments and reaching an acceptable degree of learning. It is hoped that the development of speech understanding systems will help in conceiving and check models or theories on human speech perception.

Acknowledgements

This work was performed at the Centro Elaborazione Numerale dei Segnali, Turin, Italy and was supported by the Consiglio Nazionale delle Ricerche of Italy.

References

[1] Lesser V.R., Fennel R.D., Erman L.D. and Reddy D.R., "Organization of Hearsay II Speech Understanding System", Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University (April 1974), pp. 11 ÷ 21.

[2] Woods W. A., "Motivation and Overview of BBN SPEECHLIS: An Experimental Prototype for Speech Understanding Research", ibid., pp. 1 ÷ 10.

[3] De Mori R., "Design for a Syntax-Controlled Acoustic Classifier", Proc. IFIP Congress 1974, vol. 4, pp. 753 ÷ 757.

[4] De Mori R., Rivoira S. and Serra A., " A Special Purpose Computer for Digital Signal Processing", IEEE Trans. on Computers,(in press).

[5] De Mori R., Laface P. and Piccolo E., "Automatic Detection and Description of Syllabic Features in Continuous Speech", T.R.-CENS, Turin, Italy, October 1974.

[6] De Mori R., "A Descriptive Technique for Automatic Speech Recognition", IEEE Trans. on Audio and Electroacoustics, vol. AU-21, April 1973, pp. 89 ÷ 100 .

[7] Fu K.S., "Syntactic Methods in Pattern Recognition", Academic Press, 1974.

[8] De Mori R., Rivoira S. and Serra A., "Automatic Learning of Spectral features in Continuous Speech", Proc. Third International Congress of Cybernetics and Systems, August 1975.

[9] Fung L.W. and Fu K.S., "Syntactic decoding for computer communication and pattern recognition", TR-EE74-47, School of Electrical Engineering, Purdue Univ., Lafayette, Ind., December 1974.

Appendix I

This appendix presents an example of the evaluation of a syllabic hypothesis. Fig. 3 contains the frequency and amplitude patterns of the PSS "unA" belonging to the spoken sentence: vorrei prenotare una vettura (I would like to rent a car). The phonetic transcriptions are made by hand on the pictures and have been deduced by the patterns of the time evolution of the parameters. In fact, time flows according with the arrows in the pictures, but the durations cannot be read from the drawings. The amplitudes are in base-two logarithmic scale and are normalized with respect to a reference A_0. The picture shows also the slope-code. Lines have two parameters: duration in hundreds of microseconds and lengths in hertz for frequencies and conventional logarithmic units for amplitudes.

The stable zones are represented with a symbol S and three parameters: the duration, and the coordinates of gravity centers with the same units used for lengths.

The description d_1 of the syllable is :

F1F2D : S(6, 500, 700) o (4, 430) t (5, 420)
S(6, 700, 1300) p(4, 410)

A1A2D : S(7, 1.5, 1.4) r (5,0.5) n (3, 0.9) t (2, 0.4)
S(8, 1, 8, 2 , 0)

F3F4D : S(8, 2250, 0) • S(6, 1800, 3200) • S(11, 2400, 0)

A3A4D : S(8, 0 . 8, 0) • S(6, 0 , 05, 0 . 05) •
S(11, 1. 2, 0)

The formants have been extracted with linear prediction and have probability 1 (non alternative paths were possible).

The description translator (DT) for the syllable S_j = unA operates according with table A1, where le is for length, d for duration, b_1 and b_2 are respectively the first and second gravity center coordinates for the stable zone, the probabilities are assumed to be 1 for sake of simplicity.

Table A1

Symbol of the output description	Symbol of the input description	Constraints
A	S	($300 \leq b_1 \leq 600$; $b_2 \leq 800$)
B	S	($500 \leq b_1 \leq 1000$; $1200 \leq b_2 \leq 1800$)
C	S	($2000 \leq b_1 \leq 2300$; $b_2 = 0$)
D	S	($b_1 \geq 2300$; $b_2 = 0$)
E	S	($1500 \leq b_1 \leq 2200$; $2800 \leq b_2 \leq 3500$)
F	S	($b_1 \leq 400$; $1200 \leq b_2 \leq 1500$)
K	S	($b_1 \geq 1,4$; $b_2 \geq 1,2$)
L	S	($b_1 \geq 0,5$, $b_2 = 0$)
M	S	($b_1 \leq 0 . 1$, $b_2 \leq$ 0.1)
N	S	($b_1 \geq 1$; $b_2 \leq 0 . 7$)
P	p	(le ≥ 200)
Q	l	(le ≥ 200)
W	m	(le ≥ 200)
O	c	(le ≥ 200)
R	r	($0 . 3 \leq le \leq 1$)
S	n	($0 . 3 \leq le \leq 1$)
T	l	($0 . 3 \leq le \leq 1$)
U	q	($0 . 3 \leq le \leq 1$)

The description coming out from the translator is :

g_1^j = AOQBPKRSTKC•E•DL•M•L

The SFSA for this syllable is represented by the diagram of fig. 4. On the basis of that grammar, the input syllable G is recognized with a probability :

$$P(g_1^j / S_j) = 0.02 .$$

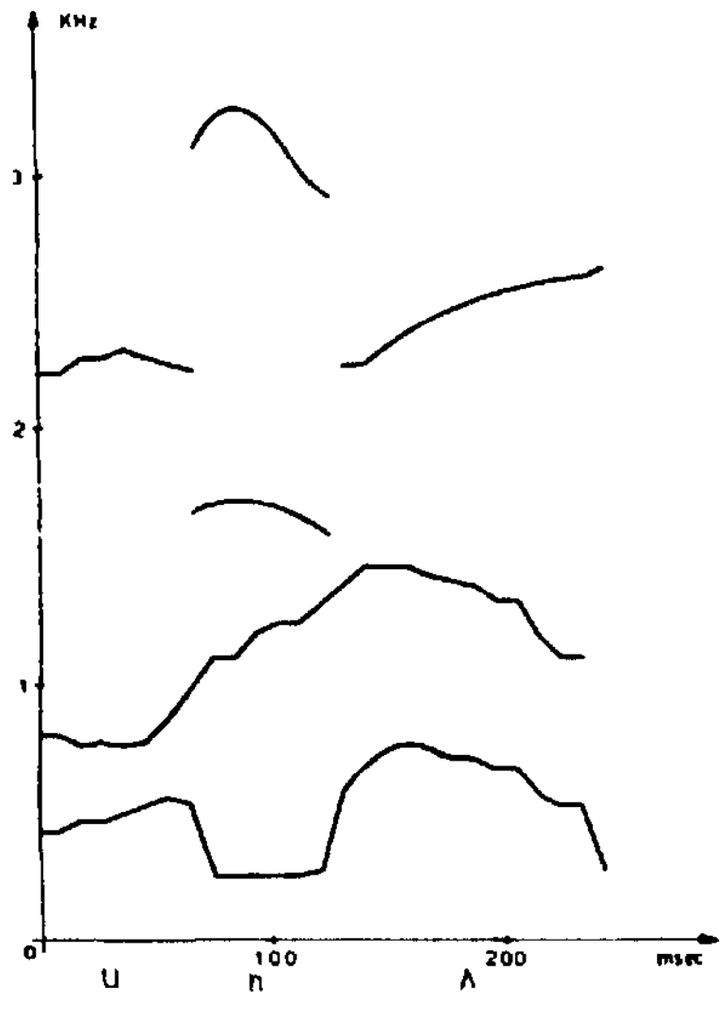


Fig. 3a - Formants of the syllable "unA"

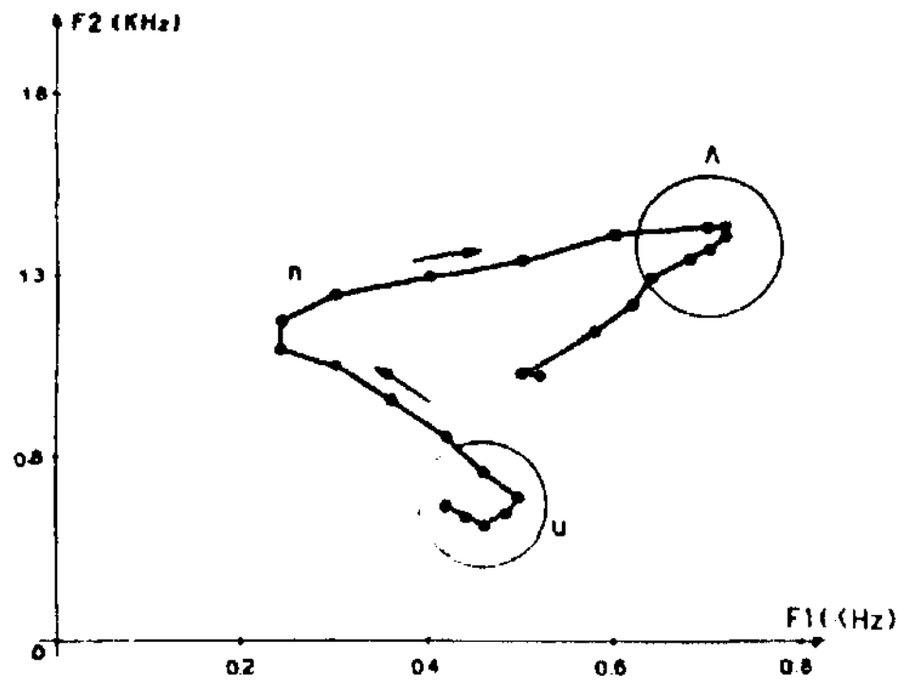


Fig. 3b - Parametric representation of F_1 and F_2

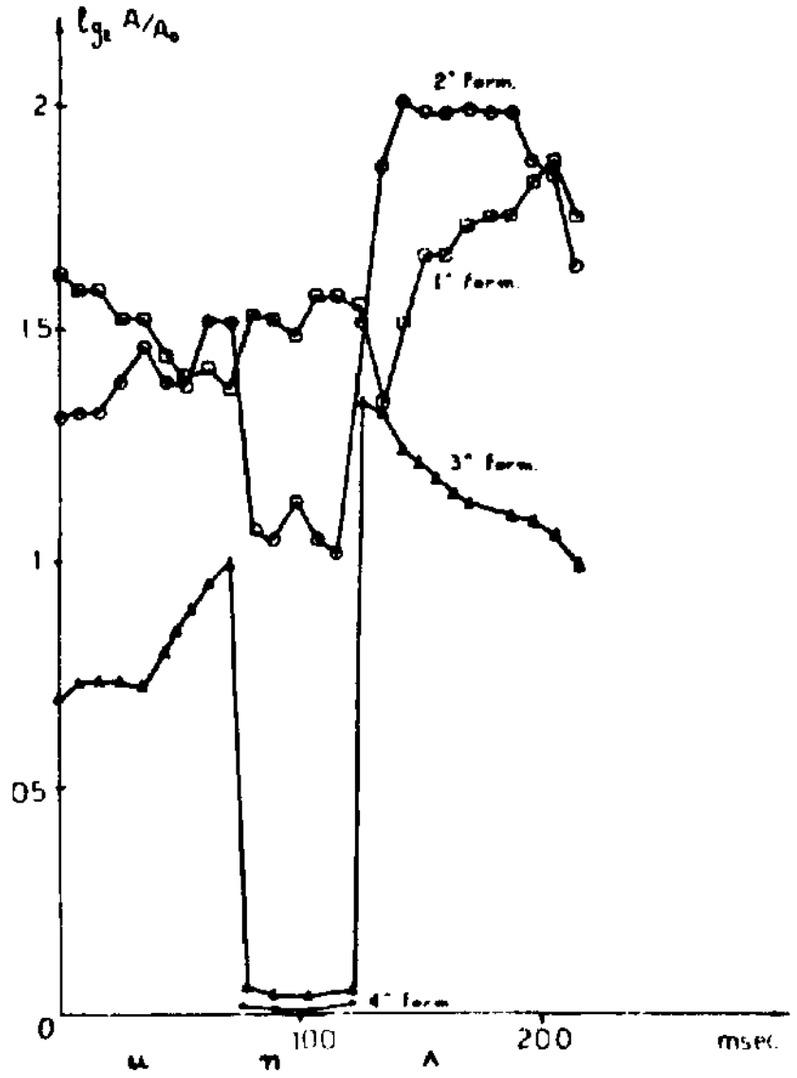


Fig. 3c - Formant amplitudes of the syllable "una"

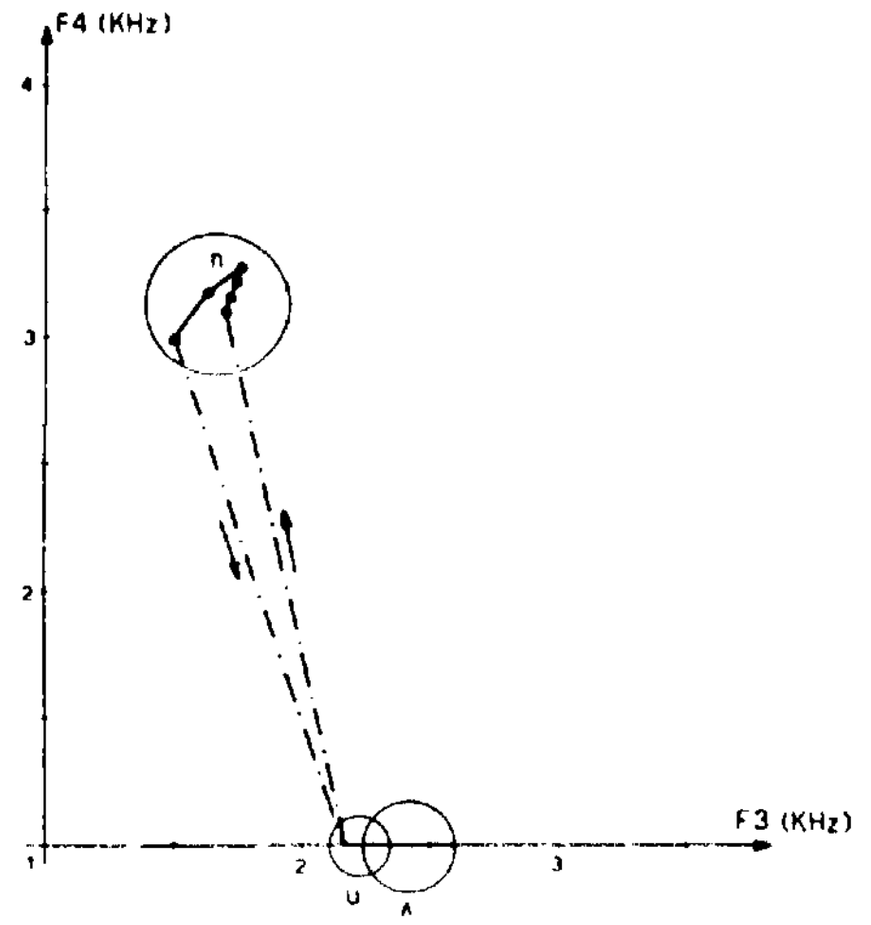


Fig. 3e - Parametric representation of F_3 F_4

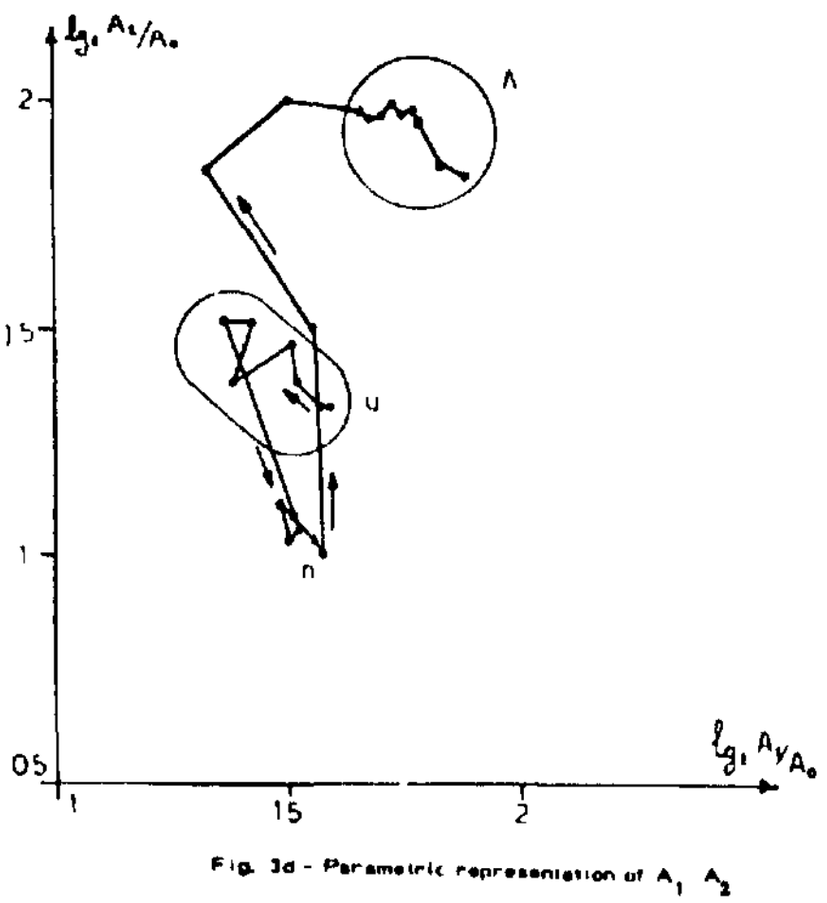


Fig. 3d - Parametric representation of A_1 A_2

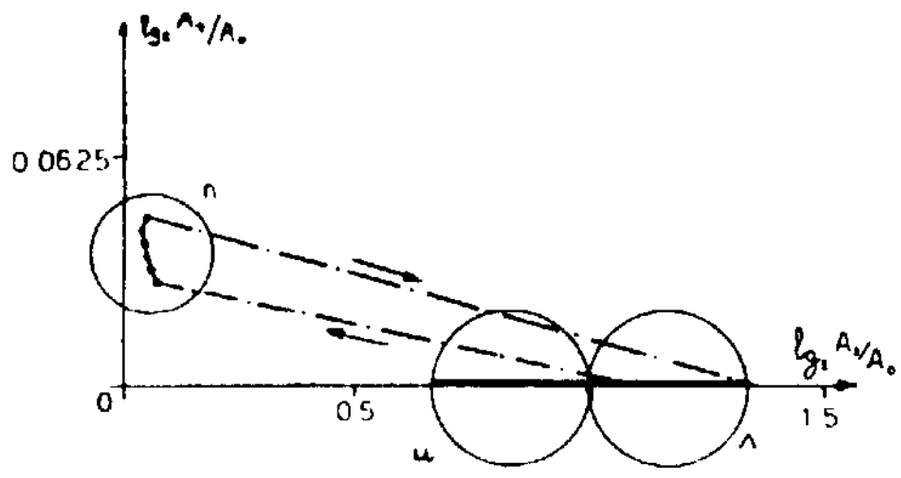
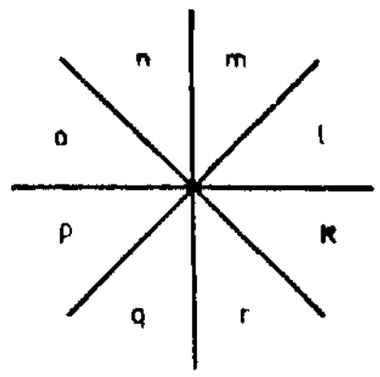


Fig. 3f - Parametric representation of A_3 A_4



slope code

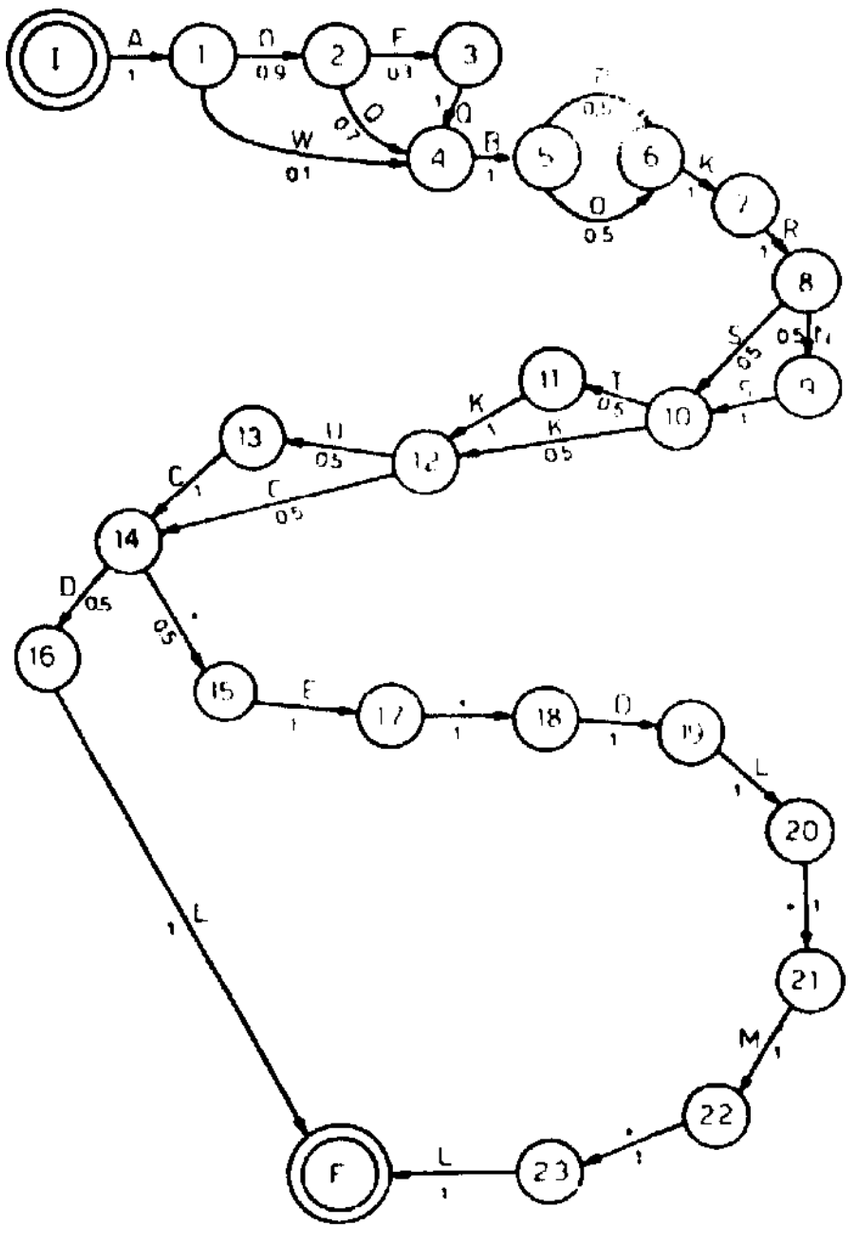


Fig. 4 - SFSA of the syllable " unW".