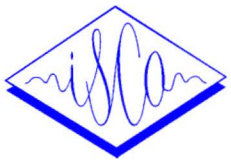

First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

MANNER-BASED LABELLING OF SPEECH SIGNAL USING TOTAL ENERGY PROFILE

A. K. DATTA

Indian Statistical Institute
203 Barrackpore Trunk Road
Calcutta 700 035, INDIA.

ABSTRACT

The paper presents a system for manner-based labelling of the speech signal for isolated words from total energy profile. The manner-based classes selected are sibilant, vowel, nasal murmur, semi vowel and lateral, trill, unvoiced unaspirated interrupt, unvoiced aspirated interrupt, voiced unaspirated interrupt , voiced aspirated interrupt and consonant clusters. The results of labelling for 100 words uttered by one male and one female informants are presented. The relavance of this approach in very large vocabulary phoneme-based spoken word recognition system is discussed.

1.0 INTRODUCTION

The paper deals with a subproblem of the acoustic expert subsystem (AEX) 'in an association of expert subsystem' approach to ASR [1]. This subproblem relates to the manner-based primary labelling of the speech signal from instantaneous energy profile. The speech signal is rectified, low-passed with an integrator circuit having a time constant of 8 ms and then sampled at a rate of 2 KHz using an 8 bit A/D converter chip connected to the port of 8085 based microprocessor system. The microprocessor is used to acquisition the data for an onward transmission to a personal computer.

Ten manner based categories, namely, 1) sibilant, 2) vowel, 3) nasal murmur, 4) semi vowel and lateral, 5) trill, 6) unvoiced unaspirated interrupt, 7) unvoiced aspirated interrupt, 8) voiced unaspirated interrupt, 9) voiced aspirated interrupt and 10) consonant cluster are selected on the premises that these may be efficiently detected and classified from a study of the energy profile alone. Such a primary classification will limit the word hypothesis to a very small number, on an average to.....[1], even with a very large vocabulary size. Further within this particular group of words directed by this 'manner word' phonemic disambiguation will be required for a very limited no of phoneme position in the word. Furthermore the disambiguation path can be controlled to a great extent by the reliability factor associated with classification of different phonemes. For example the first priority of disambiguation may be vowel classification with distant vowel seperation getting the higher priority.

The procedures and algorithms described here have been implemented on a personal computer using pascal, basic and turbo-prolog. The results of a pilot run of 100 multisyllabic words uttered by one male and one female native informant in standard colloquial Bengali (SCB) is reported. A discussion on automatic modification of rules using split-marge paradigm is also included.

2.0 PRE-PROCESSING

The energy profile is first normalised such that the peak amplitude in a word in equal to 1. The local gradient information from the digitised energy profile is used to determine the local maxima points. The gradient information used are only +ve, -ve and zero gradient. These local maxima information will be required for detecting voiced and unvoiced zones. The signal envelop is drawn by joining consecutive local maxima. Since the original signal is relatively busy locally this method gives a good envelope preserving speech related variation while smoothing out the random variations in glottal vibration or the fricative production mechanisms. A comparison of fig.1 and fig.2 brings out this point.

The primary segmentation consists of dividing the energy profile into significant hills. The only constraint is that the $depth_i$ and $depth_f$ should be less than 0.9 of the Ampl (fig.3) for a segment. The 12 parameters extracted from each segments are T_i, X_2, X_3, T_{max}, X_5, X_6, T_f, $depth_i$ $depth_f$, ampl (fig.3) slope i and slope f. the slope i is define as follows.

$Slope_i$ $=(.8\ ampl-depth_i)/X_3-X_2$ when X_3, X_2 exists
$= (.9\ ampl-depth_i)/X_3-T_i$ " X_2 not available
$= (.9\ ampl-depth_i)/T_{max}-X_2$ " X_3 not "
$= (ampl-depth_i)/T_{max}-T_i$ " X_3, X_2
both are not available

$Slope_f$ is defined in the same way replacing $depth_i$ X_2, X_3 and T_i by $depth_f$, X_6, X_5 and T_f respectively. One instance of segmented example is shown in fig.2.

The quasi-periodic and quasi-random zones are determined using the following rules :

(1) $t_n - t_{n-1} < 1.5$ ms : For a nominal fundamental frequency of 200 Hz this indicates variational range from 125Hz to 800Hz which is more than enough in any real situation

(2) $t_n < 7.5$ ms : This corresponds to a fundamental frequency of 80Hz.

(3) If the local amplitude in a quasiperiodic zone is less than 15% of the amplitude in a word this is a quari random zone

(4) A quasi-random zone of length less than or equal to 7.5 ms between two quasi-periodic zone is a quasi-periodic zone.

This algorithm works quite well. Fig.5 examplifies one such instance.

4.0 PROCEDURE

The complete labelling procedure after primary segmentation consists of two distinct phases [Fig.4]. Phase 1 is vowel labelling and phase 2 is consonant labelling.

4.1 VOWEL LABELLING

The vowel segments are characterstically high and wide with nearly smooth top. However they are often split into two or more segments due to interaction with trills, bursts and due to other personal artifacts. Moreover their intensities depend upon the openness of the vowel, the distance from the begining of the word, stress pattern and conjunction with laterals, nasals and semivowels. This has made the vowel detection procedure some what complex. This procedure is split into three seperate algorithms :

a) Combine likely consecutive segments and check-if-vowel.

b) Split and recombine likely cluster and relabel,

c) Search for weak vowels in leftout likely space

The basic constraints selected for a segment to be a vewel are :

1) The segment should be mostly quasi-periodic,

2) The length of a vowel segment is at least 50ms.

3) The amplitude of the 1st vowel segment is at least 0.6 and that for any other vowel is 0.6 times the amplitude of the preceeding vowel.

4) For weak vowels the amplitude threshold is half of normal in corresponding cases.

The details of rules are given in table 1.

4.2 CONSONANT LABELLING

The gap between, before and after the vowel is located first. The segments in these are are combined using different rules. Here the relative depth is not the criteria but the distance between two peaks and their relation to the group amplitute is exploited. The rule-base is divided into three groups for word-initial, word-medial and word final positions. The parameters used are gap, periodicity, amplitude of the small peaks in the gap and $slope_i$, $slope_f$, $depth_i$ $depth_f$ of the adjoining vowel segments. Rules are arranged in a tree-like structure with node-names such that the name itself denotes the traverse path. Table 2 gives an example of a part of the rules.

4. RESULTS AND DISCUSSIONS

Table 3 summarises the result of vowel labelling. The recognition score is encouraging. The false inclusions are mostly due to conjunction of laterals, semivowels and nasals with vowel segment. The recombination and splitting rules need to be carefully augmented. Table 4 summarises the results of consonant labelling. Free vowel indicates that the corresponding terminal segment is a vowel. The detection of trill through consonant labelling program shows very large inclusion errors. The solution may lie in the fact that a trill introduces a colouring of the vowel itself. Therefore these phonemes should be detected at the vowel level. The aspirated unvoiced plosives also show a large error. Further improvement is required in the rule base. The large no. of rejections for "YYL" is trivial simply because there is no rejection elsewhere. The rules which triggered these rejection label "XXX" needs a change of the label to "YYL". The total error comes out at 10% to which the contribution of trill and aspirated voiced sounds are unexpectedly large. It is likely that corrective measures may bring down the error rate of these two sounds to the average error level. In that case one can reasonably expect an error of about 3%.

Because of the restriction in the Turbo-Prolog environment modification of rule base can not be incorporated directly. A seperate programme for modification of rule-base using split for each error and merge for batch is being tried. This produces explosion of rules due to atomisation of parametric ranges. They have to be manually supervised for achieving a reasonable compression of the number of the rules.

The significant as pect of the study is that a preliminary manner-based classification from the low-cost parameter of total energy is indicated. This may go a long way in improving the status of word recognition using phoneme as the sub-word criterion.

ACKNOWLEDGMENT

The author wishes to acknowledge with thanks the technical assistance of Miss Lopamudra Roy Chowdhury, Mr. Tarun Dan and Mr. S.C. Kundu for their assistance in programming and Mrs. Sandhya De Bhoumick for typing the manuscript.

REFERENCE

1. A.K. Datta, Anuradha Roy and N.R. Ganguli, "An expert system for key syllable based isolated word recognition". PRL 6, pp.145-150, 1987.

TABLE 1. Rules for Vowel Labelling

Rule No						Remarks
1	Ampl (n+1) > 0.6*Ampl(n); Ampl(1) > 0.6					Amplitude
	Parameter Status				Relation	
	X_6	X_5	X_3	X_2		
2		√	√	/	$(X_5 - X_3) > 10$	Duration
3		√	!	√	$(X_5 - X_2) > 10$	
4		√	!	!	$(X_5 - T_i) > 10$	
5	√	!	√		$(X_6 - X_3) > 10$	
6	√	!	!	√	$(X_6 - X_2) > 10$	
7	√	!	!	!	$(X_6 - T_i) > 10$	
8	!	!	√		$(T_f - X_3) > 10$	
9	!	!	!	√	$(T_f - X_2) > 10$	
10	!	!	!	!	$(T_f - T_i) > 10$	
11	√	-	-	√	$(X_6 - X_2) > 30$	Splitting
12	√	-	-	!	$(X_6 - T_i) > 30$	rules
13	!	-	-	√	$(T_f - X_2) > 30$	
14	!	-	-	!	$(T_f - T_i) > 30$	
15	Check 1st segment vowel Else combine till Vowel					Recombination Relabelling
16	Check last segment Vowel Else combine from last till Vowel					
17	Other segment non-vowel					
18	Gap before 1st Vowel > 150 ms					Location for weak-vowel-search
19	Gap between two Vowels >150ms					
20	Gap after last Vowel > 150 ms					

LEGEND √ = Exists ! = non-existent - = immaterial

TABLE 2. Consonant Labelling Rules for Word-initial Single-peak Case.

Rule No.	Rule
1.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K),K=1,Gap>=30,Ampl1>=0,1,Ampl2>=0.25,Ampl2<0.5,Label="AVP",Rule="$a_1b_3c_3$".
2.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K),K=1,Gap>=30,Ampl1>=0.05,Ampl1<0.1,Ampl2<0.05,Label="YYL",Rule="$a_1b_1c_2$"!.
3.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K),K=1,Gap>=30,Ampl1>=0.05,Ampl1<0.1,Ampl2>=0.25,Ampl2<0.5,Label="NVP",Rule="$a_1b_3c_2$",!.
4.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K),K=1,Gap<30,Ampl1<0.06,Ampl2<0.25,Label="ANP",Rule="$a_2b_1c_2$",!.
5.	do-label-initial-single-peak (K,-,Ampl2,-,Gap,Label,Rule):- free(K),K=1.Gap>=30,Ampl2>=0.5,Label="NVP",Rule="$a_1b_3c_2$",!.
6.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K) ,K=1,Gap>=30,Ampl<0.05,Ampl2>=0.25,Ampl2<0.25,Label="XXX",Rule="$a_1b_3c_1$",!.
7.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K) ,K=1,Gap>=30,Ampl1>=0.05,Ampl1<0.1,Ampl>=0.05,Ampl<0.25,Label="NVP",Rule="$a_1b_2c_3$"!.
8.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K) ,K=1,Gap>=30,Ampl1>=0.1,Ampl2<0.25,Lbel="XXX",Rule="$a_1b_1c_1$",!.
9.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free(K) ,K=1,Gap<30,Ampl1>=0.06,Ampl>=0.25,Label="YYL",Rule="$a_2b_2c_2$",!.
10.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free (K),K=1,Gap<30,Ampl>=0.06,Ampl2<0.25,Label=" AVP",Rule="$a_2b_2c_3$",!.
11.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free (K),K=1,Gap<30,Ampl1<0.06,Ampl2>=0.25,Labe"XXX",Rule="$a_2b_2c_1$",!.
12.	do-label-initial-single-peak (K,Ampl1,Ampl2,-,Gap,Label,Rule):- free (K),K=1,Gap>=30,Ampl<0.05,Ampl2<0.25,Label="XXX",Rule="$a_1b_2c_1$",!.
13.	do-label-initial-single-peak (K,-,-,-,-,Label,Rule):- free (K),K=1,Label="ZZZ",Rule="$a_0b_0c_0$",!.

TABLE 3. Results of Vowel Labelling.

No. of words	No.of vowels	Correct Labelling Primary	Weak	Total	Rejection	False Inclusion
50x3	510	429 (84%)	81 (16%)	479 (94%)	30 (6%)	31 (6%)

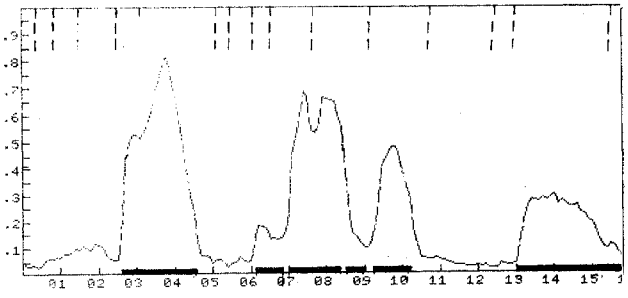

Fig. 1 Compressed Smooth Total Energy Profile of the word /bhuribhojɔn/

TABLE 4. Results of Consonant Labelling

Sl. No.	Class	Correct Labelling	False Labelling	Rejection
1.	Free Vowel	186	2(1%)	0
2.	Lateral & Semivowel	102	12(1%)	5
3.	Nasal	30	0(0%)	0
4.	Trill	18	6(33%)	0
5.	Sibilant	30	0(0%)	0
6.	Consonant Cluster INTERRUPTS	48	0(0%)	0
7.	Voiced-Aspirated	54	1(0.5%)	0
8.	Voiced-Unaspirated	114	30(2.8%)	0
9.	Unvoiced-Aspirated	48	6(12.5%)	0
10.	Unvoiced-Unaspirated	162	24(1.5%)	0
11.	Total	792	81(10%)	5

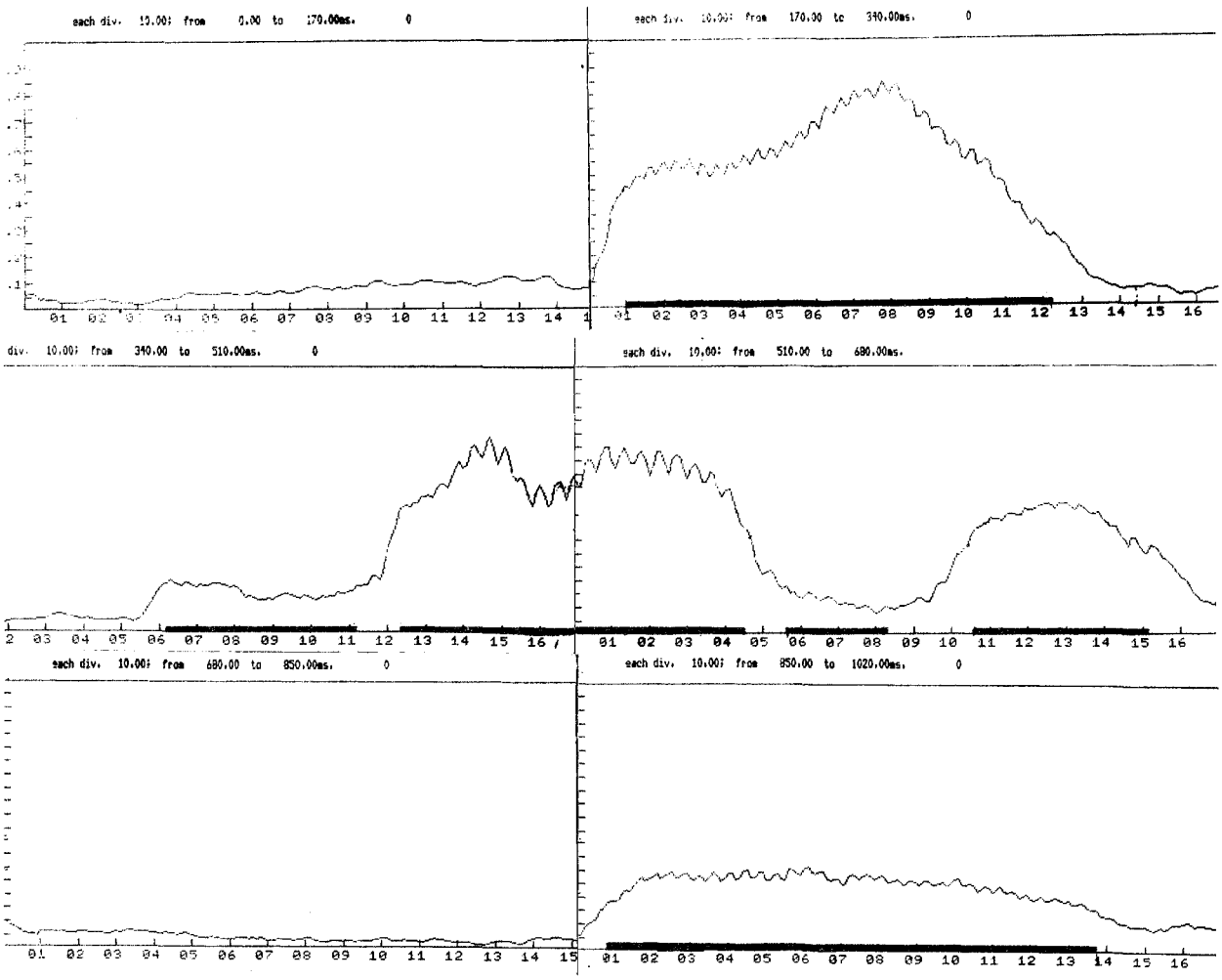

Fig.2. Raw Total Energy Profile for the Same Word.

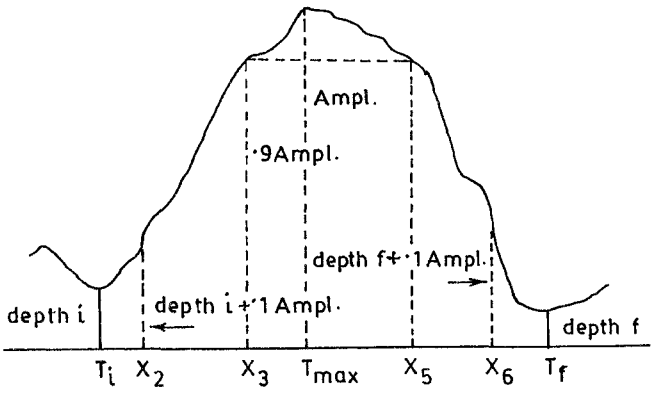

Fig.3. Parameters of a Segment

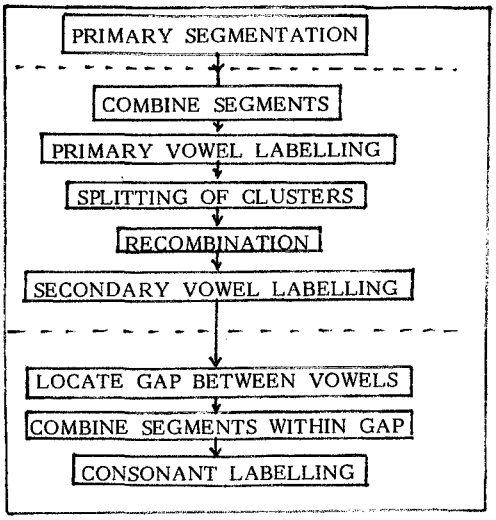

Fig.4. Flow Chart for the Complete Procrdure