STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO AIM-213

STAN-CS -73-385

RECOGNITION OF CONTINUOUS SPEECH: SEGMENTATION AND CLASSIFICATION USING SIGNATURE TABLE ADAPTATION

BY

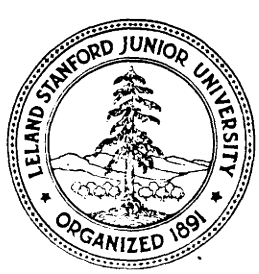
R. B. THOSAR

SUPPORTED BY

ADVANCED RESEARCH PROJECTS AGENCY
ARPA ORDER NO. 457

SEPTEMBER 1973

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY



STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO NO. AIM-213

SEPTEMBER 1973

COMPUTER SCIENCE DEPARTMENT
REPORT NO. CS-385

RECOGNITION OF CONTINUOUS SPEECH: SEGMENTATION AND CLASSIFICATION USING SIGNATURE TABLE ADAPTATION

by

R. B. Thosar

Abstract: This report explores the possibility of using a set of features for segmentation and recognition of continuous speech. The features are not necessarily "distinctive" or minimal, in the sense that they do not divide the phonemes into mutually exclusive subsets, and can have high redundancy. This concept of feature can thus avoid apriori biding between the phoneme categories to be recognized and the set of features defined in a particular system.

An adaptive technique is used to find the probability of the presence of a feature. Each feature is treated independently of other features. An unknown utterance is thus represented by a feature graph with associated probabilities. It is hoped that such a representation would be valuable for a hypthoesize-test paradigm as opposed to one which operates on a linear symbolic input.

This research was supported by the Advanced Research Projects Agency, Dept. of Defense under contract DAHC 15-73-C-0435.

Recognition Of Continuous Speech: Segmentation and Classification Using Signature Table Adaptation,

by

R.B.Thosar

1.0 Introductory

A very thorough study into the problems of speech-understanding systems has been reported by Newell at, al .[1]. They consider a variety of task and speaker dependent requirements for such a system. In the past two years, several models for speech-understanding systems and their implementaions have been reported[2,3,4]. These models view the speech-understanding process in its totality: the use of all the knowledge, roughly categorized as the acoustics, syntax and Semantics of the language (or speaker) and the specific task under consideration.

This report does not concern itself with a total system. It deals with the representation problem in the acoustic domain, independent of the task but withoutloosing sight of the fact that eventually it must feed into, and get feed-back from a general natural language understanding system.

We desire a representation of speech signal which is in some sense optimum for subsequent analysis: say syllabification using phonological constraints in first instance which in turn may be used as input (with errors) to an interactive natural language understanding system.

The representation sought in NOT in terms of a linear string of symbols(phonemic/phonetic or even sub-phonetic), but in terms of "features" which are non-mutually-exclusive. The representation is thus a feature "graph" in which the features Overlap and the overlaps are not constrained by a pre-defined relationship between the features.

This representation Is favored despite the obvious complexity because we feel that it is unrealistic to expect segments of the signal to fall into nice, clear-cut Phonetic or sub-phonetic slots when the high context sensitivity of not only the transitional sounds (liquids and glides [5] but also the relatively stationary vowels[6] and fricatives [7] is known to exist. This is true for a single speaker, so when multiplicity of speakers is considered the situation is much worse. Further, we are still far away from Using another source of knowledge: the inter-phonemic dynamics introduced by the articulatory constraints of the vocal mechanism, which are not "redundant", at least for human speech perception.

This approach is related to the hyperphoneme clustering suggested by Astrahan[8] but differs in one crucial aspect. The set of hyperphones are mutually exclusive snd hence result in a linear, non-overlappingsequence of segments. The model described by Reddy[2] uses a moregeneral two-levelsegmentation where voiced/unvoiced ond fricated/non-fricated dichotomy disambiguats firstlevel clustering similar to Astrahan's. However we use a generalized concept of feature where the feature set is used both for "segmentation" and labelling. Thus we do not consider segmentation and classification as todistinct steps in the recognition Process,

This concept of feature necessarily imposes another condition. It can not be a dichotomous decision whether a feature is present or absent, The presence of a feature is associated with a confidence figure: the probability of that featurebeingpresent when specific input signal is given. Thus ifthe value of probability is significant in a relative sense thenitis "present", Or if it is higher than a opposing feature (as requested by syntax, say) then we may say that the feature is present.

Use of an adaptive classifier is not essential to this approach per se, however it does free one from making many a ad hoc decision when buildingaclassifier.Also it has a clear advantage in an evolving system. We need not make aprjori decisions as to the speaker invariance of the parameters measured of the set of features adopted in an implementation. Clearly there is an upperlimittothe range speaker dependent variations a classifier would tolerate, An adaptive system can be modified to meetthese variations without major changes in the decision methodology.

In view of above discussion a set of general requirements for a recognition system may be summarized as follows.

1) It is "potentially" capable of extracting all the information contained in the signal, Stationary as well as non-stationary.Context independent as well as context dependent.(The co-articulation studies: Ohman[6], show that the two dichotomies are not necessarily identical),

2) The symbolic information that is extracted, either as single feature-property, or combinations thereoff, must have a confidence level associatedwith it, This seems necessary to cut down the combinetorics though not essential conceptually, because any complete system would eventually recover from an error,

3) It should be capable of generating an estimate as to a given piece of signal having certain property: a verification capability as opposed to 1) above, when the Context (at any level in the acoustic- syntactic-semantic soup) has high expectation of x being Y, but the acoustic input interprets itdifferently.

and 4) It should be adaptive.

The three requirements above are probably satisfyable by most acoustic systems, Put when we consider the fact that a speech communication system would be most useful when it accepts many speakers, some form of speaker adaptation, preferably not starting from scratch for each newcomer, seems desirable,

In the following sections we outline a system which satisfies above requirements. The present system exists more as a loosely organized collection of programs. The stress has been on showing the feasibility rather than working towards a computationally efficient and hence basically rigid code, The next section is devoted to the description of the general specifications, capabilities and the built in constraints Of the existing programs. The following section discusses the reasoning behind the choice of the sub-Phone& elements and their associated distinctive features, Experiments performed using a set of sentences provided by the ARPA data base which demonstrate the capabilities of the system are given in the fast section,

1.1 System Specifications

This Section summarizes the specifications and the constraints operative on the current system, However, at several places we point out the generality and extendebility of the system where ever it may not be obvious from the very broad view taken in the Introduction,

1) The acoustic input is sampled at 20 KHz. The parametric space is created by taking 256 sample FFTs which are overlapped 128 samples, 19 parameters (formants and such) are extracted from each FFT, scaled and quantized to 6-bits. Table 1 describes the set of parameters. For details see [9]. Thus the only acoustic information retained is the ordered set of 19-dimensional vectors, each taken every 6.4 msec.

These parameters are not the "best", nor are they sufficient. Pitch information and pitch synchronous analysis would provide better, a more consistent set of parameters, The inverse filtering technique [10] or an equivalent one would provide much better formant information as compared to the simple peak-picking used at present.

2) The system operates in a stationary world, No attempt has been made to extract information from the non-stationary parts of the signal directly or from the reduced parametric representation in 1) above. The knowledge of what to look for is fairly extensive- from the speech synthesis, perception and analysis experiments including the co-articulation analysis. The main Problem of where to look, i.e. the necessity of apriori segmentation is more or less obviated by feed-back from the first order segmentation: the primary results presented in this paper. The main reason for this lack is that the "classification" box used in this system is basically a probabilistic, trainable classifier and hence needs a large sample to produce reliable results. it is an open question whether a simpler classifier would suffice at this stage in the recognition process or the very high context-sensitivness would make a probabilistic decision more attractive.

3) The adaptation process is totally supervised. Each training sample is labeled as one of a set of phones(which includes NULL for non-stationary and undecidable sections). A phone however, is used as a convenient way to define a set of distinctive features and the linguistic-phonetic conotations of the symbol used to represent a PHONE are more of A convenience than having any conceptual relation. Our experience so for encourages us to believe that even without any extensive Syntactic support, it would be easy to bootstrap the system, at least for a single speaker for all the features and for higher level features (eg which decide between the set of FRONT VOWELs) for multiple speakers. The second factor may be more important in a larger, non-stationary environment where the speaker habits(such as nasalizing certain vowels in specific contexts) may be crucial even with all the syntactic, semantic support.

4) All the analysis shownin the results has been obtainedby processing the input instrictly left to right fashion and with a set of confidence limits on each feature being pre-defined. The multiple sets of results on the same sentence demonstrates the fact the system can also be used in the "verification" node where the probability (in the worst case the apriori) of a feature being present in a specific region of the input can be abstracted from the system

2.0 System Description

We would like to stress at the outset that the specific set of phones, features and tables described further on are not hard-wired into the programs nor in our minds. The programs used to create, learn and interprate the signature table set up described herein are developed for exploratory research and hence are very general.

Table 2 gives the list of phones and the features associated with them. The phones are defined with the stationarity of the classification process in mind. (we adapt to input over a long range learning session, but for interpretation of a specific input the classifier does not dynamically adapt).

Thus most of the continuents (vowels, fricatives etc.) which are inherently stationary appear with unaltered phonemic label. The nasals show a fairly consistent formant Structure over the closure interval though they tend to die out in amplitude in time. So the nasals are included though they are not stationary in the signal processing sense.

The glides (w,y) and their fricated counterparts (v,z) are included mostly for the segmentation purposes: we would like to locate the sections of tne signal that are weaker than vowels but stronger than voiced stops. The liquids (l,r) are somewhat more stationary, in some specific occurrences (they are undeniably context sensitive) and there is a fair chance that they might be segmented (lumped with glides and nasals) and also lend themselves to classification for those occurrences.

The stops, affricates and dipthongs are combinations of phones and hence do not appear at all. The voiced stop gaps are defined as VS. The bursts, particularly the strong/aspirated (p,t,k) cannot be spectrally distinguished (remember, we are in a 6.4 msec cross-section) fro-1 their fricative cognates (f,s,sh), but are included to make the implanted learning labeled information a little closer to a phonetic transcription. To repeat, a phone merely makes it easier to define a set of features or a bundle of features, so to say. However, the feature set is more important for both segmentation and classification.

AI though the features are clearly "bound" as a hierarchy by the definition of phones (there is no way to define a nasalized vowel in the present set except by addition of extra symbols), the feature (each) is treated independently of others during the interpretive phase, Thus even if the system was never shown a nasalized vowel, the feature NASAL may show up in parallel with a VOWEL (if nasalization is learned appropriately and is strong enough in the specific nasalized vowel) or can be "verified" using a lower confidence level interpretation for the segment if demanded by other consfderatfons.

Now consider the set of features in this table, The features VOICED, FRIC, STOP, and VOIFRI are intended for obtaining a preliminary segmentation: based on energy considerations in the spectrum, That they are not mutually exclusive is clear from the fact that "VS" the voiced stop is included with "sI", which indicates stop-gaps and silence, This avoides confusion between strong voicing

and weak voicing, A cleaner set might be obtained by measuring one more parameter, the pitch, if it can be reliably obtained for voiced stops, or by postponing the decision until after the voiced/un-voiced nature of the STOP found must be resolved, The same applies Lo the feature VOIFRI,

The voiced category of sounds is the most Prolific in speech. The features VOWEL, NASGLI (for nasal or glide) and NASAL are, used to subdivide the voiced streches of speech, again with no regard to the spectral detail, The feature NASAL is supposed to detect nasalization, But this feature is adapted to nasal stops where there is no oral output, For nasalized vowels and other sounds, the nasal pole-zero position and other nasalization effects (Fant[11]) depend on the oral shape and the degree of nasal coupling, As we shall indicate later, our results do show overlap between VOWEL and NASAL feature sets, but it is difficult to assert that the feature is detecting nasality and not merely responding to a weakened vowel,

Table 3 shows the particular parameters used to detect above features, It should be noted that only the amplitude parameters are used up to this point, The reason is that for these broad classes the variation and/or stationarity in terms of energies is more important than the frequency peaks, Thus if we were to use format position to detect a VOWEL the parameter would cover most of its range of values, At best it would produce a uniform distribution and contribute nothing or, more likely, it would add noise by creating clusters for the vowels with high frequency of occurrence and thereby weaken the VOWEL feature for others,

The three features FRONT, MID and BACK attempt to divide the frequency range in three regions, The choice of three is primarily due to the traditional classification and that in the frequency domain the three vowels /i/, /a/ and /u/ form three extrema, However, our results show more confusion between the MID and BACK, probably due to the vowels O U and the schwa AS, It may be advantageous to break-up the MID-BACK set into three, Thus having four features to delineate the voiced sounds, Again, as we are not looking for a mutually-exclusive set of-features, it does no harm if certain sounds are found as both MID and BACK, Only further interpretation becomes a little more difficult,

The features F/P, S/T awl SH/K are implied by thier names, It snows that in the present stationary classification state, we could not distinguish between th9 frication in the burst iron the corresponding fricative,

LONG and STEADY are features intended for classification of a Specific vowel, Also, these are the only non-redundant features as their absence carries the same information as their presence. in the results given in this report these do not figure prominently for two

main reasons: 1) they range over all the vowels and hence would have bad distribution in the frequency domain and 2) the total "learning" for these features is necessarily small in comparison with other features. These features have been included as an aid to further processing. Thus it would be more meaningful to process an "averaged" vowel vector after initial segmentation to give more reliable estimate of LONG and STEADY features.

2.1 System Implementation

The theoretical foundation and certain implementation details regarding the signature table adaptation are given in [12]. The hierarchy of tables (Table 3) generates a probability surface in the input parameter space which is conditional to the learnt feature. Thus, the table called by mnemonic label VOICED has as its output

$$p(AVE, HPE, LPE|VOICED) \qquad [1]$$

where AVE, HPE and LPE are the input parameters. The two inputs AVE and HPE are repeated at the earliest stage (table VOI1) in order to have explicit five variable space, the same as for other features (e.g. NASGLI). These repeated inputs are clearly redundant for this feature and are included to avoid scaling problems that arise as a result of having unequal number of inputs. This is true particularly when the amount of learning information is scanty in comparison with the size of the feature space, 2*(3*5), 32K in this case.

We wish to find the postiriori probability of a feature F given the specific input vector X,

$$P(F|X) = P(X|F)*P(F)/P(X) \qquad [2]$$

where P(F) is the apriori probability of the feature F and P(X) is the unconditional probability of X. In the present system P(F) is computed using only the information aquired during the learning phase. Thus it is given by the ratio of count this feature was specified to the total count over all the features. Clearly, these probabilities should also be modified to some extent by the known apriori distribution of the features for the language under consideration.

The unconditional probability P(X) is obtained indirectly. The phone-feature relationship (Table 2) indicates that the features VOICED, FRIC, VOIFRI and STOP are mutually exclusive and totally exhaustive. Therefore,

$$P(VOICED|X)+P(FRIC|X)+P(VOIFRI|X)+P(STOP|X) = 1. \qquad [3]$$

This constraint used in conjunction with eq.2 gives an estimate of P(X).

This method of obtaining the postirior probabilities allows us to treat all other features which do not figure in eq.3, independently of each other: one of the main advantages of this approach stressed in the introductory section.

2.2 Implementation of Counters

The feature probabilities obtained in the section above are for a input parameter vector X, which represents a short time slice of 6.4 msec., and are completely independent of the time context. Whereas, some features, particularly those related to vowels and fricatives are stationary over fairly extended time segments. This fact is used to improve the probability estimate at a time say T, by using the compound probability:

$$p(F\|X_T) = p(F|X_{(T-1)})*p(F|X_{(T-2)})*\ldots, \qquad [4]$$

where (T-N) represents a delay of N units.

In order to reduce the number of pre-assigned threshold values that must be specified to the program, the same value of this delay is used for all the features except the inherently non-stationary features (NASAL, NASGLI, VOIFRI) for which a delay value reduced by On9 unit is used.

"Counter" is the device used to detect "presence" of a feature using these compound probability estimates. A feature counter is triggered when the probability for that feature exceeds a pre-specified value, and remains high for several consecutive time units. Accidental dropouts or spikes are eliminated by using time hysterisis. A counter thus reports the onset time, the duration for which the feature was present and the average probability value over the duration.

Clearly, there is no "optimum" value of probability that can be set as the threshold for a counter. It would depend on the purpose for which the outputs are to be used and of course, the organization of the program whose task is to assign interpretation to it. Since this report deals exclusively with demonstration of the performance and capabilities of this approach, we shall present results using identical input and Varying the threshoid and the delay parameters.

It is not necessary to specify a separate threshold for each feature., The confidence witn which a feature may be detected is clearly related to the amount of learning for that feature: its apriori probability,

The actual thresholds used far each Counter are thus obtained bymultiplying its apriori Probability by a global confidence figure, whichis specified as a percentage, ranging from 1 to 100, In some "bad" cases where the probability distribution is biased because the feature is specified for unrelated phone sub-sets (e.g. LONG and STEADY), the threshold figure may be arbitrarily chosen,

3.0 Results of Experiments

The results given In this report are based on 26 utterances which were used as the data base at a speech Segmentat ion workshop held at tne Carnegie-Mellon University in July, 1973,

The utterances are listed in Table 4. The sentencesare divided in two sets of 13 sentences each; those with a "*" following the identification label are used for the training and the rest for a gross evaluation of the system,

Detailed results are given for utterance 19 to show the consistency of the feature set, The word "Tower" occures three times inthis sentence and has been used five times In the training phase(sentences 25 and 26),

Fig.1 gives the waveform of utterance 19, The vertical lines are spaced 6.4 msec, intervals. Table 5 shows the phone Information associated with this utterance and the starting and the ending segment number for each phone. This association is dune at present by visual inspection of the waveform and is rather conservativesince the classifier operates in a stationary domain,

Table 6 shows the counter outputs obtained when the confidence threshold is set to 80 and the delay parameter is 3, This combination biases the result towards higher probability and greater stationarity, Fig, 2 shows a graphic representation of these counter outputs and also the position of the phone string(Pony) associated with It, The necessary compression of thedata may cause a relative shift between the features, But the error is not more than 1 character on either side,

It may benotedthat thefeaturesVOICED and STOPare found with nigh reliability, Even when the friction in the first t-burst (seg, 82-94) is missed completely, itis not substituted by any other feature, Other t-bursts are also rather sketchy, On the other hand

the "s" at the end of the sentence(Tower C) is definitely picked up as a fricative and also identified as a S/T,

The feature VOWEL shows an interesting pattern, Most vowels indicate fair to high confidence, But note the vowel AA in the second occurrence of Tower(seg. 289-296) which is shorter in duration(anyway it is a part of a dipthong) and weaker than others, The graph in Fig.2 shows it as brokenin two sections and has low probability.

Going a step further we can see that for all the vowels in this utterance the sun-classes FRONT and MID are identified correctly, The intervening glide"w" in the word Tower is defined to be a BACK (Table 2), and this feature does not show up at all.

Fig.3 shows the graphic representation of the same utterance, analyzed with confidence level set to 60 and the time delay at 2. A comparison with Fig.2 indicates that more Information, a lot of which is redundant with respect to Fig.2, Is extracted, But the interesting aspect is tnat now the t-bursts show up as fricatives, though they are incorrectly labelled aseithar F/P or SH/K. The likely reason is that there are more instances of F and SH in the training set which are closer to these, rather weakly articulated bursts than the t-bursts and "s" sounds in the training set,

The second interesting aspectis that the "w" glides(feature NASGLI) are more definitively located and also declared as BACK, albeitwith lower confidence and morebreaks.

Figures 4 and 5 show the graphic representation for the utterance 2, with the confidence figures set to 80 and 40, and the delay Parameter set to 3 and 1respectively.

This example is intended to demonstrate another important characteristics of thisapproach: the feature set being mutually independent during the analysis phase.

Fig.4 is "clean" in the sense that none of the competing features (VOICED,FRIC,STOP), (VOWEL,NASGLI) and (VOWEL,NASAL) show any overlap, Whereas in Fig.5 the vowel EE (time 0.5 sac) has the features VOWEL and NASAL overlapping for most of its duration. On the other hand, vowels AE (time 0.65 sec), E (time 1.25 sec.) and AW (time 1,55 sec) show little or no overlap with feature NASAL,

From the phone-feature relationship (Table 2) used during the training, it is clear that the feature NASAL is associated only with nasal Phones M,N and NG. Theexampleaboveshows that nasalization in vowels can be detected without creating an explicit class for nasalized vowels,

To repeat, we do not assert that we have found the way to define and detect "nasalization" in the strict sense of the term. This is an example to substantiate our strategy for keeping the set of features independent Of each other, and for not using pre-defined relationships between features as A way to improve recognition and/or provide a simple algorithm for the phonetic labelling Of the utterance,

Tables 7 and 8 give a gross evaluation as to how well the signature tables perform when the learnt data itself is analyzed. The testing program compares the associated learning information (as in Table 5) with the output of the counters on a segment by segment oasis,

Table 7 gives the overall figures for the various features. The entry "Excess" is the sum of wrong classification, when the correct feature is not found at all and also of other features which overlap the correct feature. Thus excess is mainly a measure of seperability of the competing features. Table 8 represents the Same information with a phone-wise breakdown. This information is useful for 1) redefinition of the input parameters and feature relationship and more important, 2) redefinition of the set of features themselves. Thus result for vowel AE in the Table 8 indicates that it is classified as FRONT and MID about equal number of times. Similarly BACK vowels O and U switch between MID and BACK.

This apparent confusion is indicative of the variability of the formant structure of the vowels with context. One may get around this problem by increasing the number of features: one for overlap between FRONT and MID and another for the overlac between MID and BACK. A better solution might be to disambiguate these confusions by 1) averaging input over the vowel duration and re-interpreting the resulting vector (amounts to a local feed-back) and/or 2) postponing the decision until it becomes essential, and using acoustic and Other context information.

Table 9 gives the averaged results on the learnt data when the confidance threshold is lowered to 60 and the delay to 2. The increase in classification rates for the feature NASAL, NASGLI and BACK, among others, show that the phones which define these features are not particularly stationary.

Tables 10 and 11 give similar results for the unseen data. Similarity of the phone breakdown in Table 11, with the one in Table 8 is indicative of the consistency of classification as well as the confusion.

4.0 Conclusion

The examplesin above section give a fair idea of the capabilities and the potentiality of the present approach towards a speech recognitionsystem. The system satisfies all the requirements outlinedin the introductory section.

At this point we may make a projection as to how the present system might fit into a full-fledged recognition system. A possible strategy is outlined in a very general way by the following steps.

1) FIND sections of the unknown utterance setting the confidence parameters "high".

2) GO through a hypothesize-test procedure to identify and label these sections. Verification can be done on certain sections, with loweredconfidence levels If demanded by the context.

3) MASK those sections marked in steps 1) and 2). The recognition algorithm cannot do a better job than this!

4) If parts of the utterance are left un-interpreted then lower the confidence parameters, and go to step 1).

Clearly, the most crucial step 2 above is dependent on the constraintsand the goal of the recognition schema. One may merely include language-specific phonological rules at this stage. The system would then accept a wider class of utterances and produce a phonetic transcription, whereas a task oriented, limited vocabulary system might get away with fewer phonological rules and still provide acceptable results.

Parameter	Lower Limit	Upper Limit [Hz]
F1: first formant	200	800
F2: second formant	700	2050
F3: third formant	2,743	3200
A1: F1 amplitude		
A2: F2 amplitude		
A3: F3 amplitude		
FP1: fricative pole 1	1800	3200
FP2: fricative pole 2	3200	5000
FP1A: FP1 amplitude		
FP2A: FP2 amplitude		
FZ: fricative zero	FP1	FP2
FZA: FZ amplitude		
NP: nasal pole	800	1500
NZ: nasal zero	NP	NP+500
NPA: NP amplitude		
NZA: NZ amplitude		
LPE: low region energy	0	450
HPE: high region energy	2500	10000
AVE: average energy		10000

Table 1. Input Parameters and Their Ranges.

PH NU	IPA	Significant features				
EE	i	VOICED	VOWEL	FRONT	LONG	STEADY
AE		VOICED	VOWEL	FRONT	LONG	
E	e	VOICED	VOWEL	FRONT	STEADY	
I	I	VOICED	VOWEL	FRONT		
AS		VOICED	VOWEL	MID		
AA	a	VOICED	VOWEL	MID	LONG	STEADY
AR		VOICED	VOWEL	MID	LONG	
A	ʌ	VOICED	VOWEL	MID	STEADY	
OO	u	VOICED	VOWEL	BACK	LONG	STEADY
U	U	VOICED	VOWEL	BACK	STEADY	
AW	ɔ	VOICED	VOWEL	BACK		
O	o	VOICED	VOWEL	BACK	LONG	
Y	y	VOICED	NASGLI	FRONT		
R	r	VOICED	NASGLI	MID		
L	l	VOICED	NASGLI	MID		
W	w	VOICED	NASGLI	BACK		
M	m	VOICED	NASAL	BACK	NASGLI	
N	n	VOICED	NASAL	MID	NASGLI	
NG	ŋ	VOICED	NASAL	FRONT	NASGLI	
F	f	FRIC	F/P			
S	s	FRIC	S/T			
SH		FRIC	SH/K			
H	h	FRIC				
V	v	VOIFRI				
Ż	z	VOIFRI				
ŻH		VOIFRI				
PB		FRIC	F/P			
TB		FRIC	S/T			
KB		FRIC	SH/K			
SI		STOP				
VS		STOP				

Table 2. The Phone-Feature relationship used.

(second column gives the nearest IPA equivalent, where possible)

Name	TYPE	Learn	IN1	IN2	IN3	IN4	IN5	IN6	IN7
VOI1	P2	VOICED	AVE	HPE	AVE	HPE	LPE		
FRI1	P2	FRIC	AVE	HPE	AVE	HPE	LPE		
VOFR1	P2	VOIFRI	A1	A2	AVE	HPE	LPE		
STO1	P2	STOP	AVE	HPE	AVE	HPE	LPE		
VOW1	P2	VOWEL	A3	A2	A1	AVE	LPE		
GLI1	P2	NASGLI	A3	A2	A1	AVE	LPE		
NAS1	P2	NASAL	A1	NZA	NPA	AVE	LPE		
FRN1	P2	FRONT	A1	A2	F3	F1	F2		
MID1	P2	MID	A1	A2	F3	F1	F2		
BCK1	P2	BACK	A1	A2	F3	F1	F2		
XFP1	P2	F/P	FP1A	FP2A	FZ	FP2	FP1		
ST1	P2	S/T	FP1A	FP2A	FZ	FP2	FP1		
SHK1	P2	SH/K	FP1A	FP2A	FZ	FP2	FP1		
LNG1	P2	LONG	A1	A2	F3	F1	F2		
STD1	P2	STEADY	A1	A2	F3	F1	F2		
GAP1	P2								
GAP11	P2								
VOI2	P2	VOICED	VOI1	AVE	HPE	LPE			
FRI2	P2	FRIC	FRI1	AVE	HPE	LPE			
VOFR2	P2	VOIFRI	VOFR1	AVE	HPE	LPE			
STO2	P2	STOP	STO1	AVE	HPE	LPE			
VOW2	P2	VOWEL	VOW1	Al	AVE	LPE			
GLI2	P2	NASGLI	GLI1	Al	AVE	LPE			
NAS2	P2	NASAL	NAS1	NPA	AVE	LPE			
FRN2	P2	FRONT	FRN1	F3	F1	F2			
MID2	P2	MID	MID1	F3	F1	F2			
BCK2	P2	BACK	BCK1	F3	F1	F2			
XFP2	P2	F/P	XFP1	FZ	FP2	FP1			
ST2	P2	S/T	ST1	FZ	FP2	FP1			
SHK2	P2	SH/K	SHK1	FZ	FP2	FP1			
LNG2	P2	LONG	LNG1	F3	F1	F2			
STD2	P2	STEADY	STD1	f-3	F1	F2			
GAP2	P2								
GAP22	P2								
VOI3	P2	VOICED	VOI2	HPE	LPE				
FRI3	P2	FRIC	FRI2	HPE	LPE				
VOFR3	P2	VOIFRI	VOFR2	HPE	LPE				
STO3	P2	STOP	STO2	HPE	LPE				
VOW3	P2	VOWEL	VOW2	AVE	LPE				
GLI3	P2	NASGLI	GLI2	AVE	LPE				
NAS3	P2	NASAL	NAS2	AVE	LPE				
FRN3	P2	FRONT	FRN2	F1	F2				
MID3	P2	MID	MID2	F1	F2				
BCK3	P2	BACK	BCK2	F1	F2				
XFP3	P2	F/P	XFP2	FP2	FP1				
ST3	P2	S/T	ST2	FP2	FP1				
SHK3	P2	SH/K	SHK2	FP2	FP1				
LNG3	P2	LONG	LNG2	F1	F2				
STD3	P2	STEADY	STD2	F1	F2				

Table 3. (cont.)

GAP3	P2			
GAP33	P2			
VOICED	P2	VOICED	VOI3	LPE
FRIC	P2	FRIC	FRI3	LPE
VOIFRI	P2	VOIFRI	VOFR3	LPE
STOP	P2	STOP	ST03	LPE
VOWEL	P2	VOWEL	VOW3	LPE
NASGLI	P2	NASGLI	GLI3	LPE
NASAL	P2	NASAL	NAS3	LPE
FRONT	P2	FRONT	FRN3	F2
MID	P2	MID	MID3	F2
BACK	P2	BACK	BCK3	F2
F/P	P2	F/P	XFP3	FPa
S/T	P2	S/T	ST3	FP1
SH/K	P2	SH/K	SHK3	FP1
LONG	P2	LONG	LNG3	F2
STEADY	P2	STEADY	STD3	F2

Table 3, The Signature Table Hierarchy,

NO	IDENT	UTTERANCE
1	B10*	What is the average uranium lead ratio for the lunar samples?
2	B27	Do any samples contain troilite?
3	B34	Do you have any references on payalitic olivine?
4	B35	Do any samples contain tridymite?
5	B36*	Has whitlockite been measured in any lunar sample?
6	B40*	What are the pyroxene concentrations in each type A rock.
7	B51*	Give me the cristobalite concentrations for each type B rock.
8	D7*	Count where type equals linear equations and runtime less than fivesix.
9	D10	Repeat where key word equals Gauss elimination or key word equals eigenvalue.
10	CV1300*	Alpha becomes alpha minus beta.
11	CV3200	Alpha gets alpha minus beta.
12	LS1	I wantto do phonemic labelling on sentence six.
13	LS21	Who's the owner of utterance eight?
14	LM3	Who is the owner Of utterance eight?
15	LM13	Display the phonemic labels above the spectrograms.
16	LM14*	Put the left boundary on first "s" segment on the tenth frame.
17	LM18*	Move the right boundary of the first "ah" one position to the left.
18	LM24*	Display the root mean squared function and the silence threshold above the spectrogram
19	RB2	They are Tower A, Tower B, and Tower C.
20	RB6	Do you have any right squared boxes left?
21	RB7	Do you have any rectangular cylinders left?
22	RB11	The white block in the picture is called a box.
23	RB12*	The orange block in the picture is not a box.
24	RB16*	Put the other red block on the red block.
25	RB19*	From left to right, they are Tower A, Tower B, Tower C, and Tower D.
26	RB26*	Is there a red block in front of Towers C and D?

Table 4. List of the 26 sentences used in the experiments.

(those with a "*" following the identification label were used for adaptation)

Header Hint Information from

FILE= SEG19,T0C77,TH0]

The message in this file is :::::

(19:RB2) THEY ARE TOWER A TOWER B AND TOWER C

	P1	BEG	END
22	vs	10	12
23	E	22	23
24	I	27	36
25	AA	49	54
26	R	62	65
27	SI	71	81
28	TB	82	90
29	AA	99	105
30	W	115	120
31	AR	129	133
32	SI	148	154
33	E	162	183
34	I	204	209
35	SI	266	273
36	TB	275	280
37	AA	289	296
38	W	305	310
39	AR	314	320
40	VS	327	338
41	EL	346	382
42	H	454	455
43	N	463	470
44	SI	477	481
45	TB	483	491
46	AA	499	505
47	W	515	522
48	AH	524	529
49	S	539	563
50	EE	571	584

Table 5. Phone Information Associated with Utterance 19.

SEG#infile name refers to the Utterance Number,

Data file SEG19,TØ[77,THO] 18-JUL-1973 1413:22

(19:RB2) THEY ARE TOWER A TOWER B AND TOWER C

Trained on: LRNMIX,TMP Threshold=8Ø & Delay=3

Begin	End	Label	Level	St.Seg	End	SegCnt
First level [voiced, fric, voiced-fric, voiced & unvoiced stop]						
128	704	STOP	4	2	11	10
896	4160	VOICED	6	14	65	52
4544	5056	STOP	5	71	79	9
6116	8512	VOICED	7	94	133	40
9344	9856	STOP	5	146	154	9
10148	13632	VOICED	7	157	213	57
14016	14080	STOP	1	219	220	2
14336	17408	STOP	5	224	272	49
17600	17664	FRIC	6	275	276	2
18176	19072	VOICED	7	284	298	15
19328	19520	VOICED	5	302	305	4
19776	20416	VOICED	6	309	319	11
20736	21696	STOP	3	324	339	16
21888	24708	VOICED	7	342	387	46
25650	29056	STOP	6	400	454	55
29376	30720	STOP	4	459	480	22
31168	3x94	FRIC	3	487	489	3
31488	33536	VOICED	7	492	524	33
33920	34240	VOICED	5	530	535	6
34432	36096	FRIC	5	538	564	27
36352	37440	VOICED	7	568	585	18
37824	37885	STOP	3	591	592	2
38144	38272	STOP	1	596	598	3
38592	38656	STOP	4	603	604	2
38912	39168	STOP	3	608	612	5
Voiced [vowel, nasal, nasal/glide]						
768	832	NASAL	6	12	13	2
1344	1664	VOWEL	3	21	26	6
1920	2112	VOWEL	3	30	33	4
2368	3392	VOWEL	4	37	53	17
3648	3904	NASGLI	1	57	61	5
4032	4160	NASGLI	2	63	65	3
6144	7360	VOWEL	7	94	115	20
7360	7552	NASGLI	Ø	115	118	4
7488	7616	VOWEL	4	117	119	3
7744	8000	VOWEL	4	121	125	5

Table 6. (cont.) (19)

8?64	8512	NASGLI	1	126	133	8
10?48	12480	VOWEL	7	157	195	39
12608	12800	VOWEL	1	197	200	4
12928	13184	VOWEL	2	202	206	5
13504	13568	NASAL	5	211	212	2
18048	18112	NASAL	5	282	283	2
18368	18624	VOWEL	4	287	291	5
18880	18944	VOWEL	0	295	296	2
19392	19520	NASGLI	0	303	305	3
19712	20416	NASGLI	2	308	319	12
22016	23232	VOWEL	3	344	363	20
23488	23680	VOWEL	3	367	370	4
23872	23936	NASAL	6	373	374	2
23936	24000	VOWEL	2	374	375	2
24128	24192	NASAL	6	377	378	2
24192	24320	VOWEL	6	378	380	3
24192	24384	NASGL I	4	378	381	4
24448	24512	NASAL	6	382	383	2
24640	24704	NASGLI	3	385	386	2
24640	24704	NASAL	1	385	386	2
31424	31552	NASAL	0	491	493	3
31744	32768	VOWEL	3	496	512	17
32896	33536	NASGL I	2	514	524	11
33856	33920	NASGLI	2	529	530	2
36416	37056	VOWEL	4	569	579	11
37184	37248	VOWEL	0	581	582	2

Fricatives [F/P, S/T, SH/K] and [LONG, STEADY]

6528	6656	LONG	7	102	104	3
6528	6656	STEADY	7	102	104	3
6848	7040	STEADY	6	107	110	4
19?08	19264	STEADY	3	297	301	5
24128	24256	STEADY	,	377	379	3
31168	31232	SH/K	4	407	488	2
34368	34432	SH/K	4	537	538	2
34816	36096	S/T	5	544	564	21

Table 6. (cont.)

[front, mid, back]

1536	2496	FRONT	2	24	39	16
3584	3648	MID	5	56	57	2
6144	6976	MID	7	96	109	14
7488	7616	MID	5	117	119	3
7552	7616	FRONT	1	118	119	2
8192	8512	MID	2	128	133	6
10048	12288	FRONT	7	157	192	36
12544	13568	FRONT	5	196	212	17
18368	19264	MID	5	287	301	15
19904	20160	MID	2	311	315	5
21888	24832	FRONT	7	342	388	47
31744	32192	MID	4	496	503	8
33344	33536	MID	1	521	524	4
36352	37440	FRONT	4	568	585	18

Table 6. Counter Outputs for the Utterance 19 with Confidence Threshold set to 80 and time delay to 3, (Begin and End times are in 10 microsec. units).

Feature	Given	Found	Excess	%Found	%Excess
VOICED	1816	1365	18	75	1
FRIC	741	477	44	64	6
VOWEL	1241	630	11	56	1
NASAL	322	69	7i	2i	22
STOP	698	605	116	87	17
VOIFRI	113	8	3	7	3
F/P	153	24	5	16	3
S/T	346	257	35	74	10
SH/K	222	111	19	50	9
FRONT	583	353	59	61	10
MID	991	332	11	34	11
BACK	242	48	24	20	10
LONG	642	11	6	2	1
STEADY	752	29	3	4	0
NASGLI	575	175	101	30	18

Table 7. Averaged Feature Performance of Learnt Data. (Threshold=80 and Delay=3)

	EE	AE	E	I	AS	AA	AR	A	OO	U	AW	O	Y	R	L	W	M	N	NG	F	S	SH	H	V	Z	ZH	PB	TB	KB	SI	VS
VOICED	108	63	221	145	90	278	71	50	19	30	19	22	4	68	64	43	43	56						5	5		1				7
FRIC		8		3																33	284	102		1	28	4		24	51			
VOWEL	48	39	168	83	44	208	33	22	19	25	11	5		3	1	4									3							
NASAL	6	8	12	6	9		4					4		8	6	1	28	41							2		1	1			3
STOP																2	5	49	1	39			6	7					85	30	98
VOIFRI																				2	1				8							
F/P																				24									1	4		
S/T																					261	4			31				6	1		
SH/K		8		1																1	1	84			3	4			1	28		
FRONT	93	211	47	98	13	7	12		8			1	4	7	4			4						1	1						1
MID		25	38	15	19	23	30	19	28	8	12	4	2		18	13		3	12					2							3
BACK		2	7		3	4	2		7	8	3	5		5	1	12	13														
LONG			4		1	13		1	1																						
STEADY			11		3	14		3	1																						
NASGLI	1	7		8	11	12	24	8	3	1	1	11	1	43	46	31	22	34		2				6							6
TOTAL	11?	652	19?	174	104	314	98	49	21	30	23	25	41	76	73	70	73	247		213	727	910	7	20	36	68	9	16	67	115	56	113

Table 8. Phone break-down for the result in Table 7.

Feature	Given	Found	Excess	%Found	%Excess
VOICED	1816	1432	31	79	2
FRIC	741	519	66	70	9
VOWEL	1241	876	34	71	3
NASAL	322	182	264	57	82
STOP	698	605	176	87	25
VOIFRI	113	30	38	27	34
F/P	153	41	25	27	16
S/T	346	281	70	81	20
SH/K	222	145	46	65	21
FRONT	583	411	127	70	22
MID	991	515	191	52	19
BACK	242	117	139	48	57
LONG	642	33	17	5	3
STEADY	752	67	14	9	2
NASGLI	575	360	457	63	79

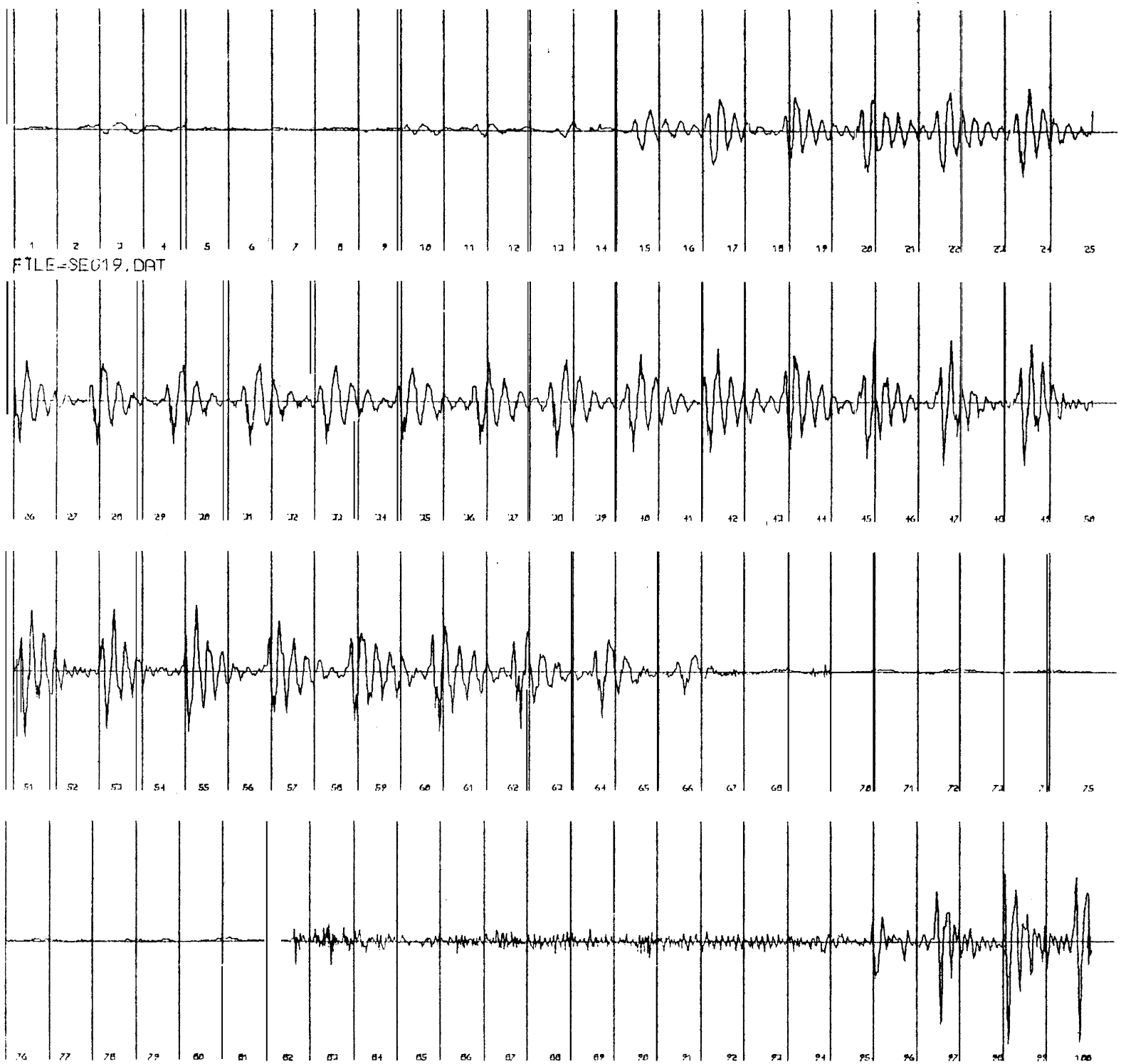
Table 9. Averaged Feature Performance for the Learnt Data.
(Confidence Threshold=60 and Delay=2)

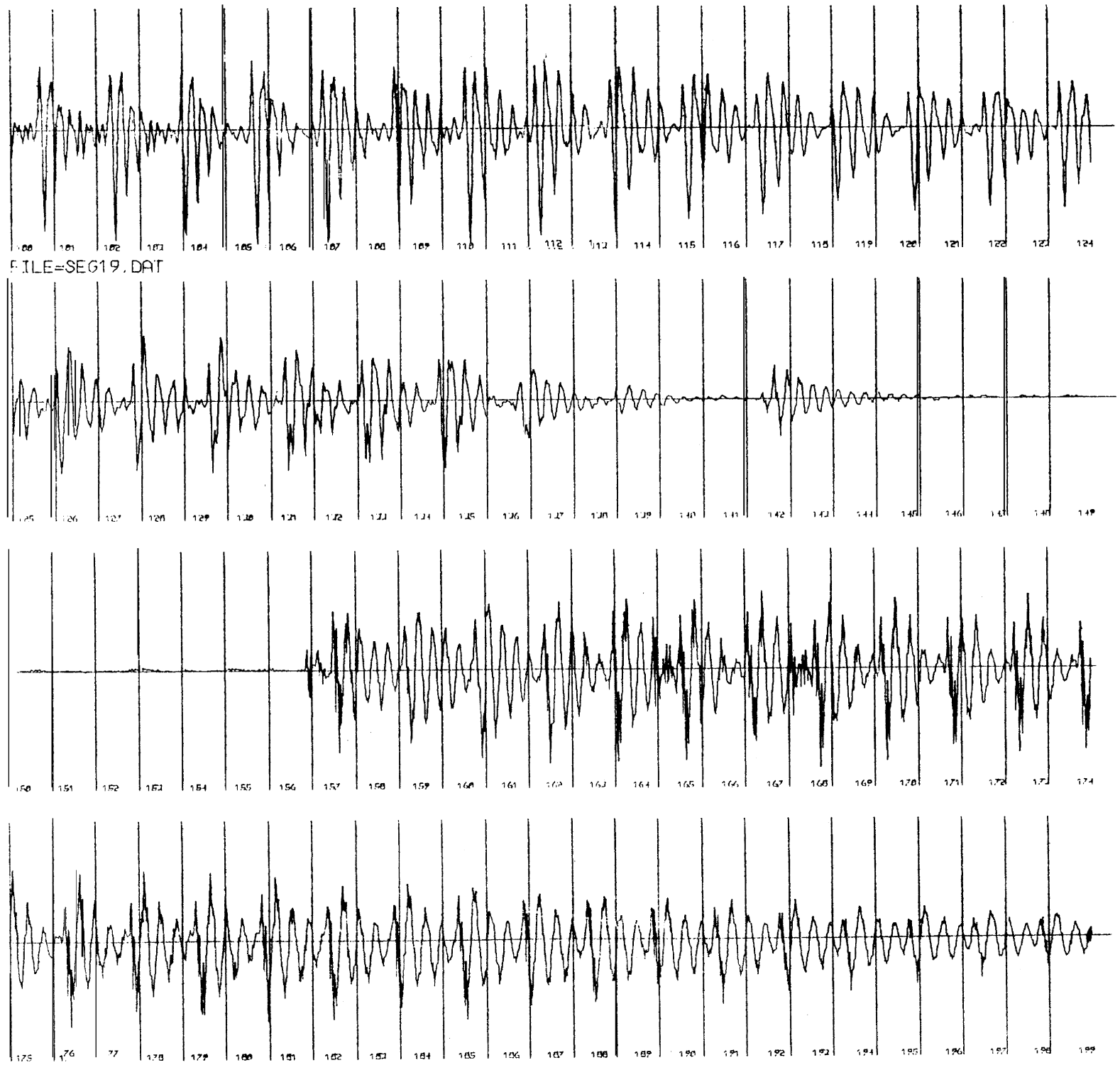
Feature	Given	Found	Excess	%Found	%Excess
VOICED	1553	1186	37	76	2
FRIC	523	317	39	61	7
VOWEL	1020	509	26	50	3
NASAL	220	45	81	20	37
STOP	410	327	111	80	27
VOIFRI	91	1	3	1	3
F/P	a2	9	2	11	2
S/T	329	211	38	64	12
SH/K	92	19	7	21	8
FRONT	630	342	64	54	10
MID	573	146	126	25	22
BACK	350	12	22	3	6
LONG	570	4	2	1	0
STEADY	597	8	6	1	1
NASGLI	533	139	146	26	27

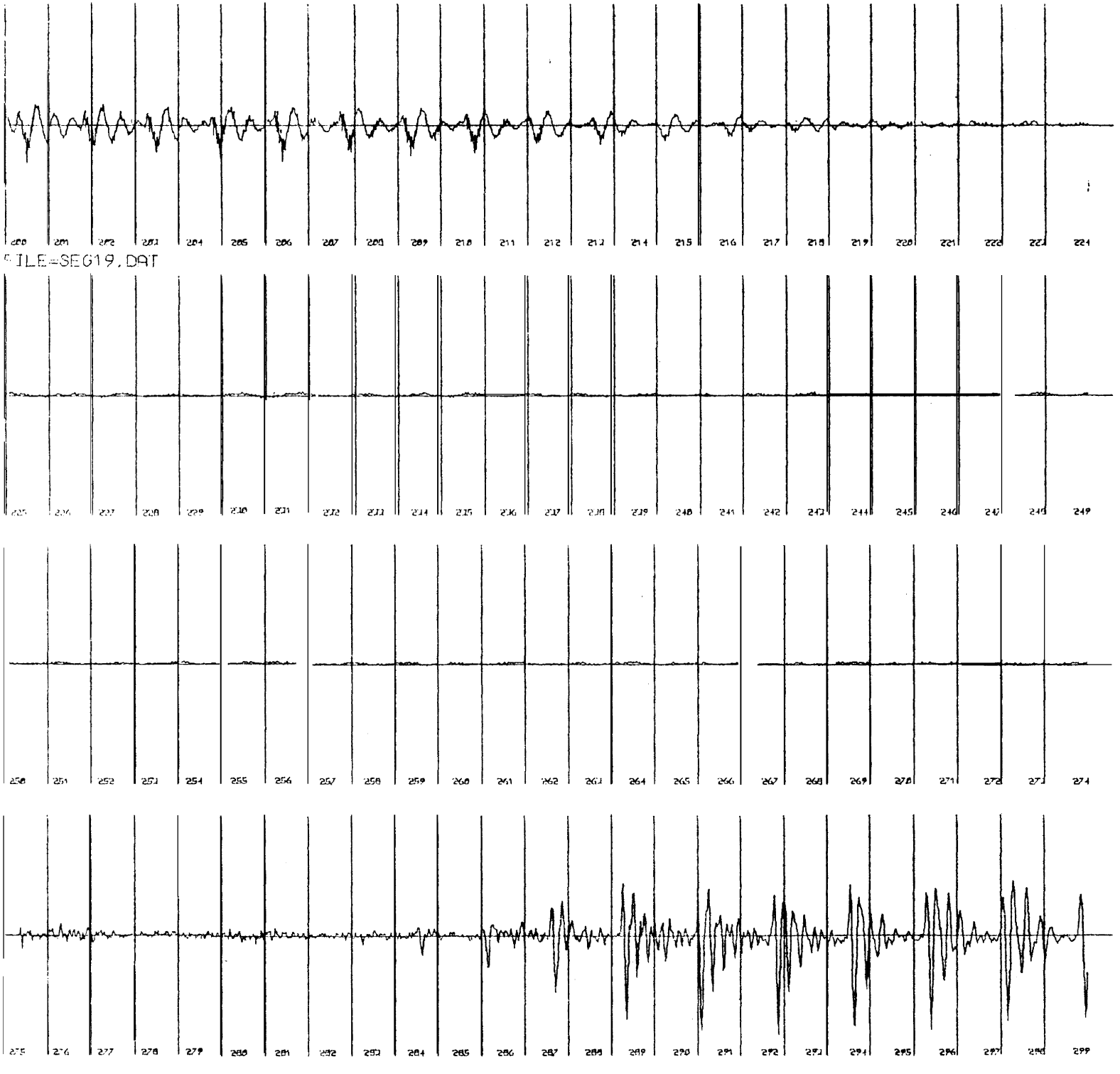
Table 10. Averaged Feature Performance for the Unseen Data.
(Confidence Threshold=80 and Delay=3)

	EE	AE	E	I	AS	AA	AR	A	OO	U	AW	O	Y	R	L	W	M	N	NG	F	S	SH	H	V	Z	PB	TB	KB	SI	VS
VOICED	171	109	173	100	86	101	47	36	77	13	19	26	16	30	58	59	29	65	4				8	16	6		1	2		3
FRIC	3															2				23	228	7			36	1	28	39		
VOWEL	169	87	138	49	30	43	5	22	18	7	5	11	4		2	18		1					1							
NASAL	14	5	11	7	2				7			1		6	8	5	10	33	2				3	5	2		1			4
STOP														1		68		22		19				2			2		289	50
VOIFRI																					2				1			1		
V/P																				8						1		2		
V/T																					204				29		12	9		
H/K	3															1					1	13					2	7		
FRONT	136	43	99	63	14		1	8	16	5		2	12	8		1		6					2					1		
MID	3	45	15	6	20	64	26	16	10	3	8	20			16	10		5						6	1					
BACK		3	7	1	2	2			5				2	4	1		7													
LONG					1	2	2						1																	
STEADY	2	4	1		2	2		3																						
ALL	4	6	3	3	15	25	28	1	32	4	2	7	8	24	20	28	20	39						9	1			2		4
TAL	136	114	175	120	90	105	56	34	84	13	18	25	20	53	72	168	42	163	15	76	251	19	20	34	57	6	78	73	331	79

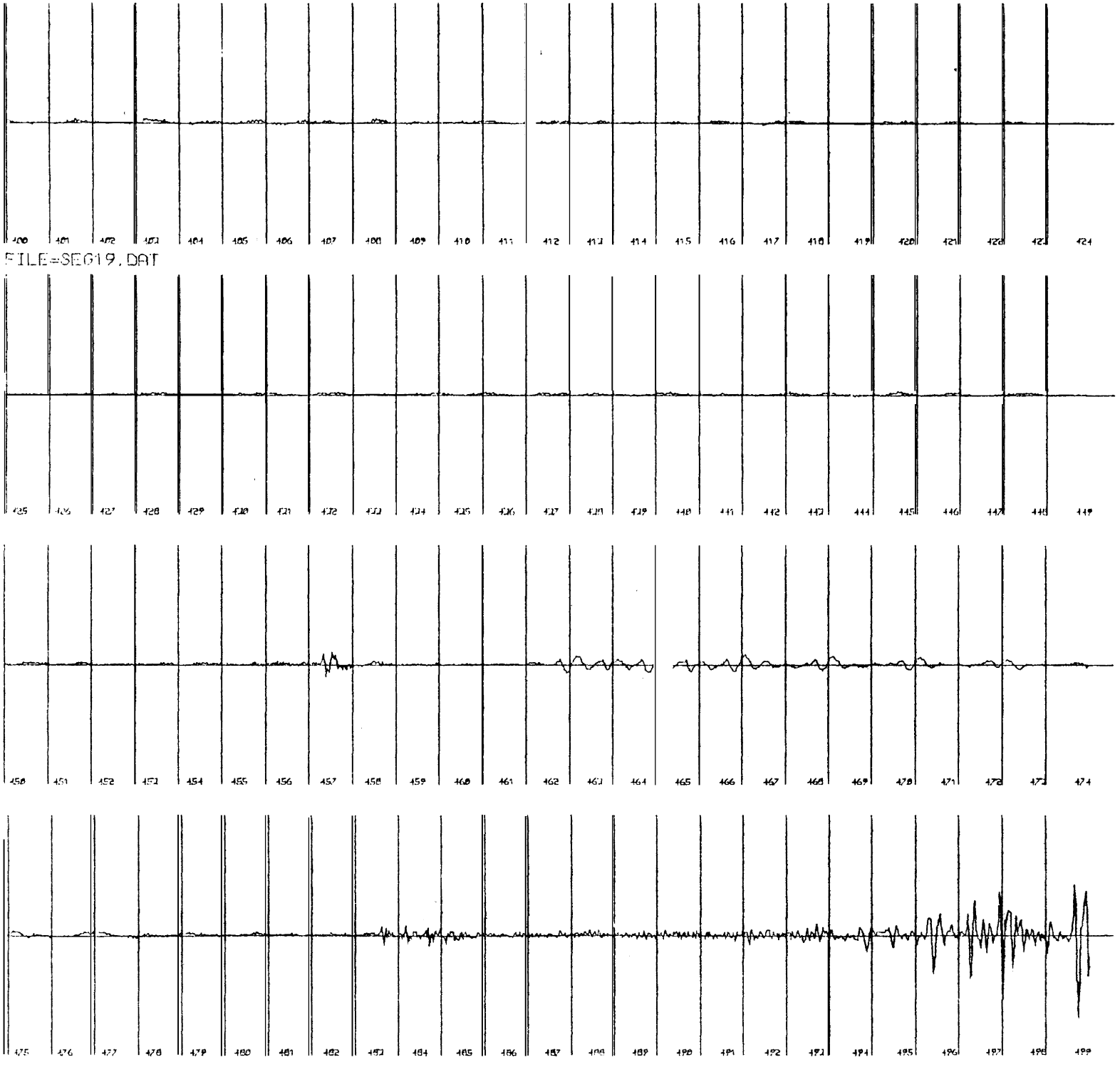
Table 11. Phone break-down for the result in Table 10.

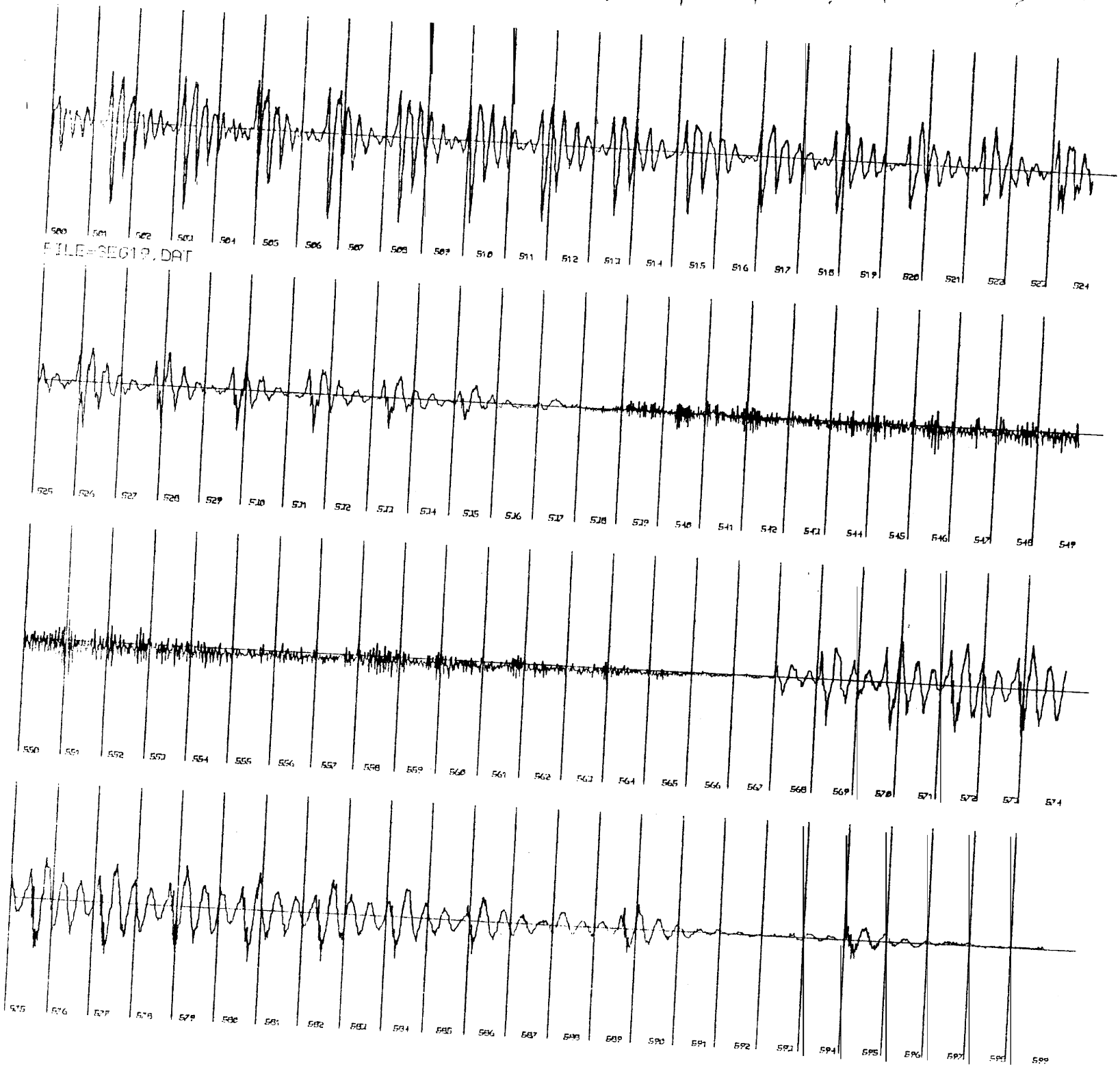

FILE=SE019.DAT


FILE=SEG19.DAT


FILE=SEG19.DAT

FILE=SEG19.DAT


FILE=SEG19.DAT


FILE=SEG12.DAT

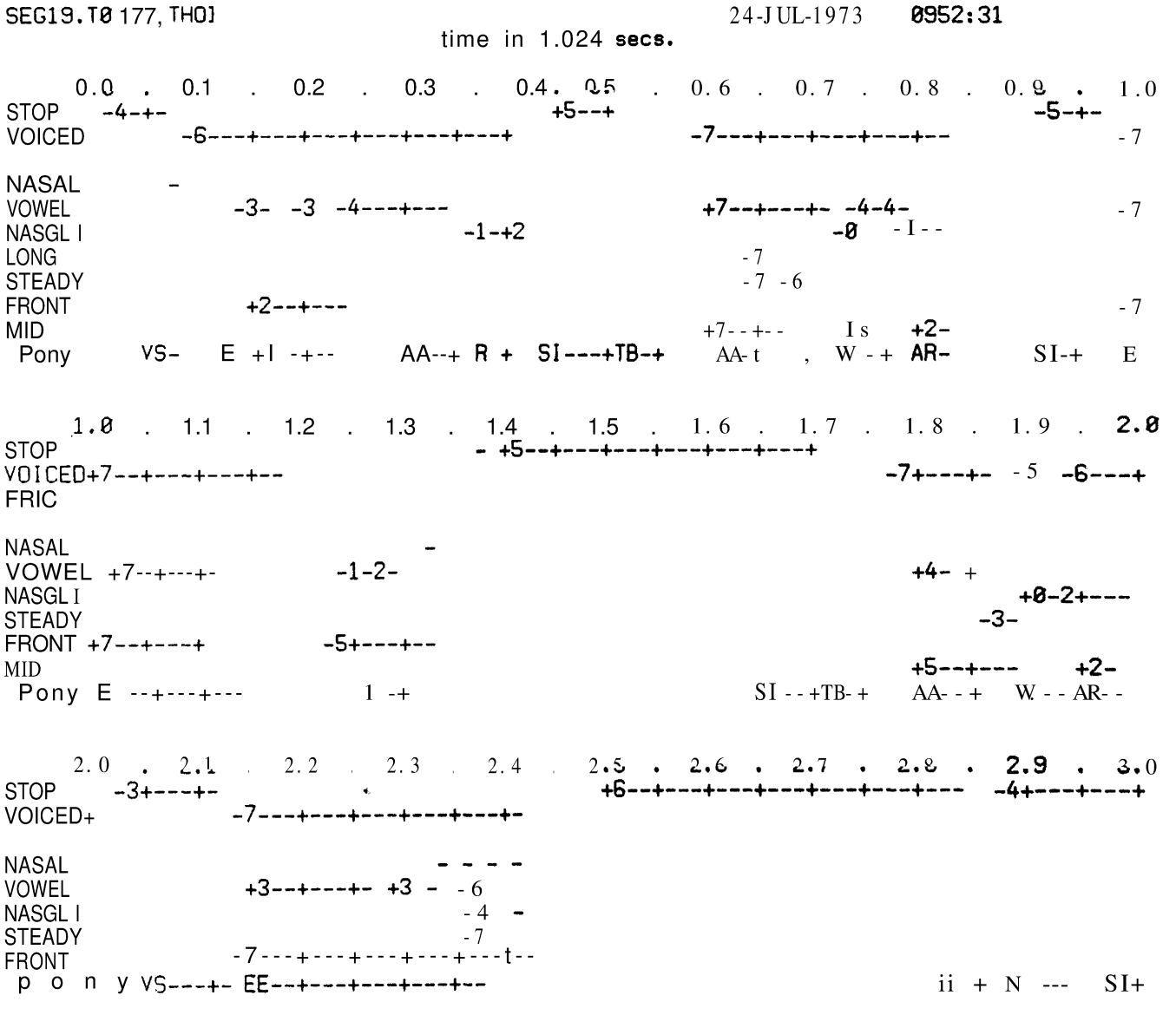
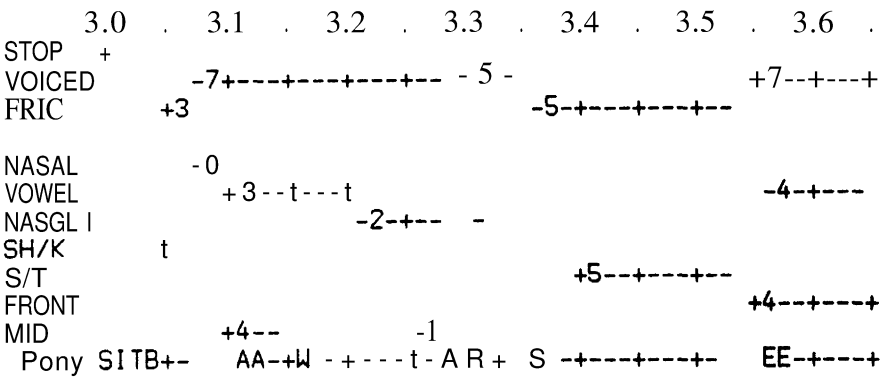


Figure 2. (cont.1

SEG19.T0[77,THO] 24-JUL-1973 0952: 32

time in 1.024 secs.



Notes: The t symbols denote scale divisions only.
The numbers on lines are confidence figures (unsigned, 0 to 8).
The position of pony data may not be exact because of scale compression.

Figure 2. Graphic Representation of Counter Outputs in Table 6.

time in 1,824 secs.

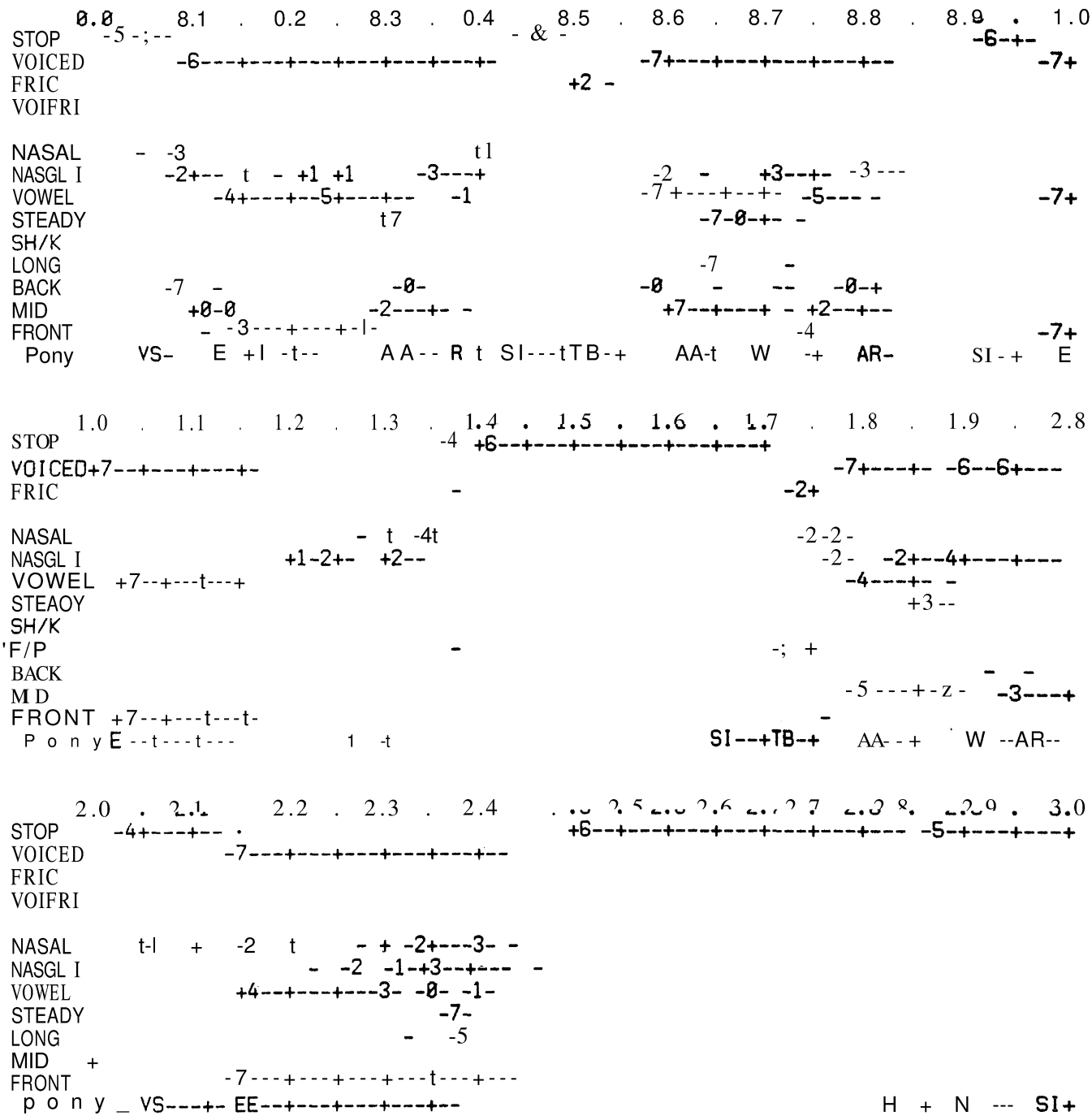
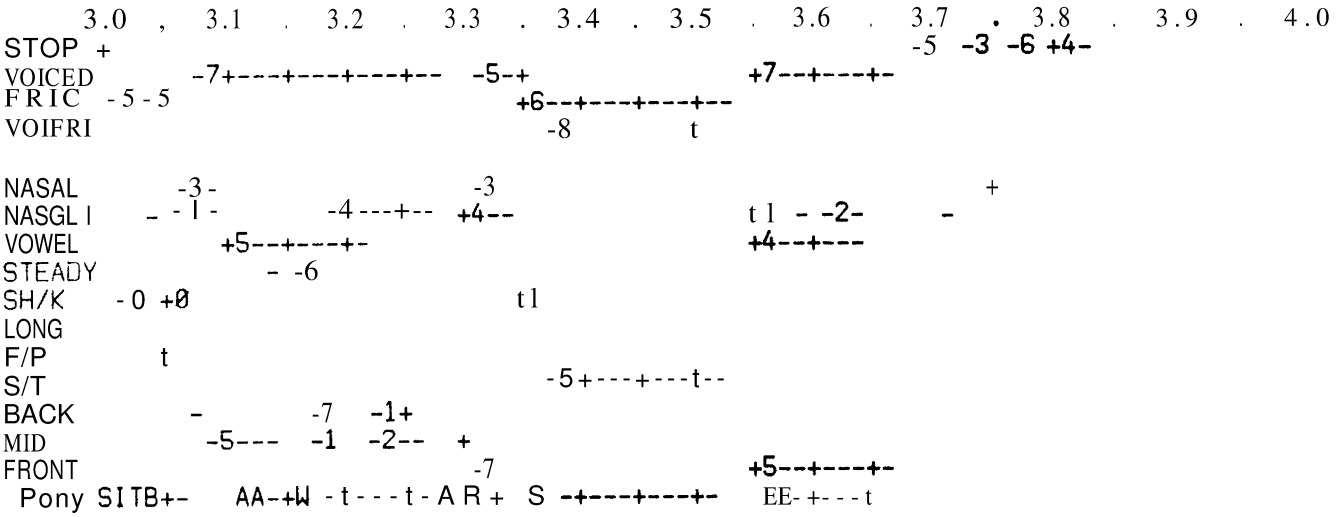


Figure 3. (cont.1

SEG19.TO[77,THO] 24-JUL-1973 0356: 18

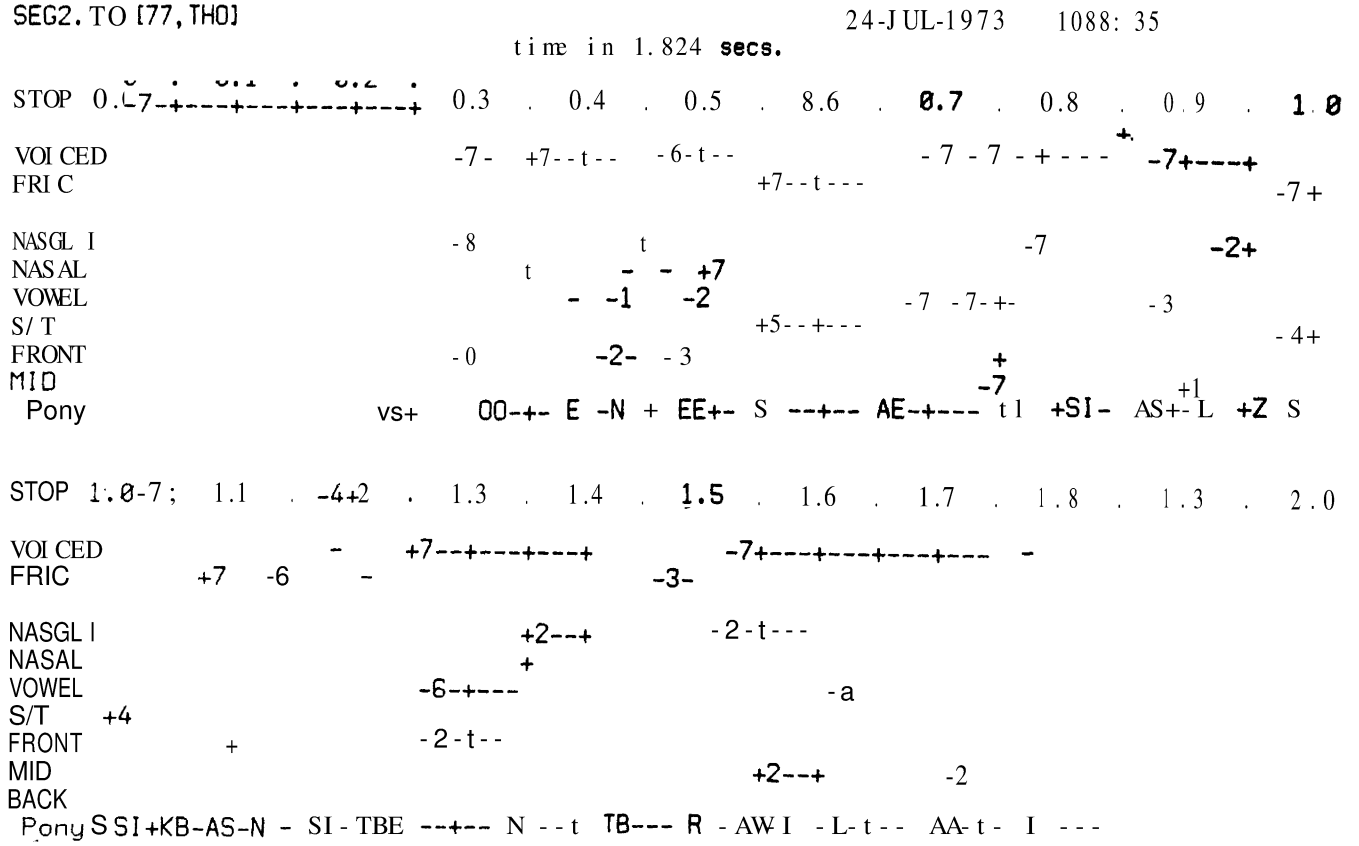
time in 1,024 secs.



Notes: The t symbols denote scale divisions only.
The numbers on lines are confidence figures (unsigned, 0 to 8).
The position of pony data may not be exact because of scale compression.

Figure 3. Graphic Representation of Counter Outputs for Utterance 19

(Confidence Threshold=60 and Delay-21



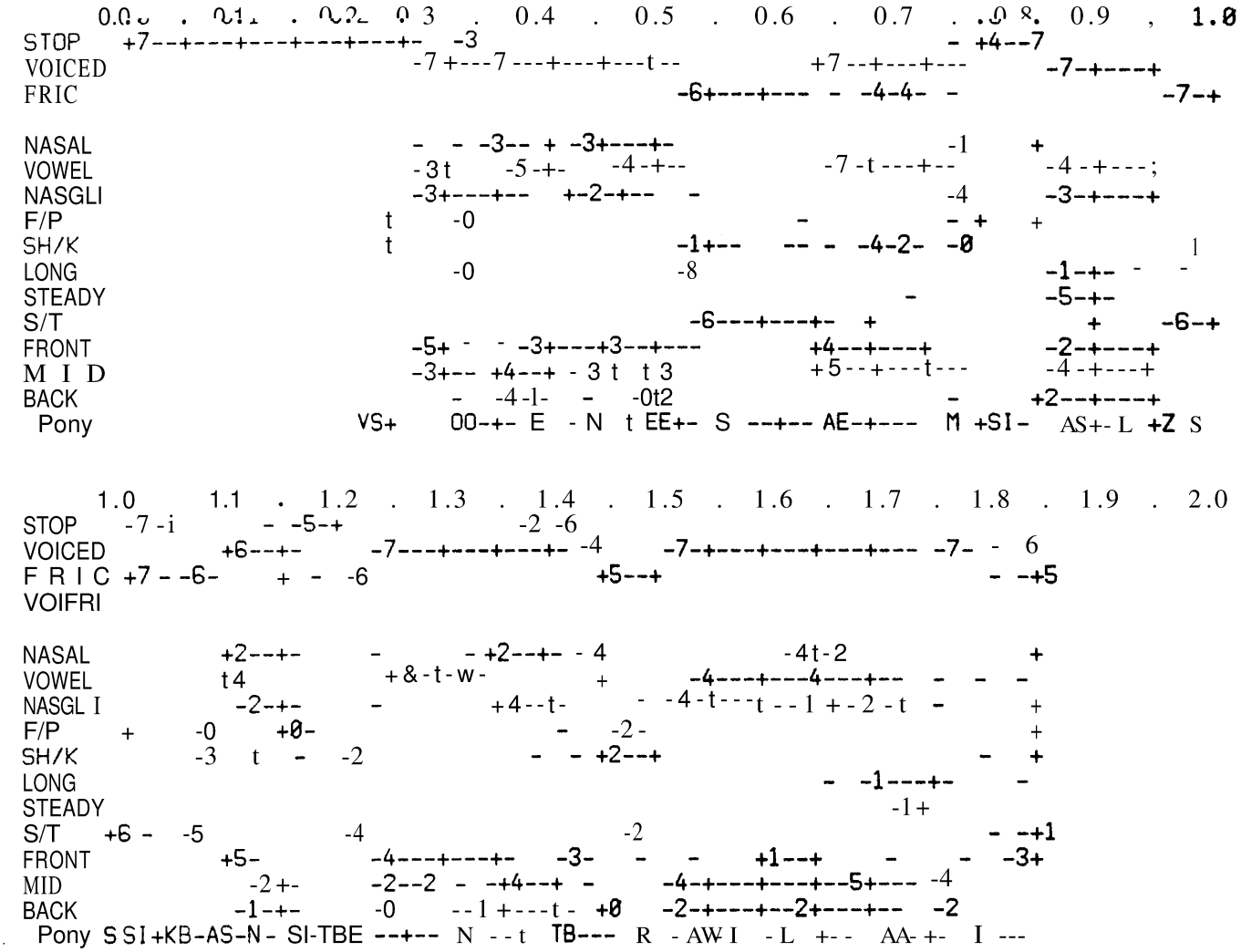
Notes: The t symbols denote scale divisions only.
The numbers on lines are confidence figures (unsigned, 8 to 8).
The position of pony data may not be exact because of scale compression.

Figure 4. Graphic Representation for Utterance 2.

(Confidence Threshold=80 and Delay=3).

SEG2.T0[77,THO] 24-JUL-1973 1005: 55

time in 1.024 secs.



Notes: The + symbols denote scale divisions only.
The numbers on lines are confidence figures (unsigned, 8 to 8).
The position of pony data may not be exact because of scale compression.

Figure 5. Graphic Representation for Utterance 2.

(Confidence Threshold=40 and Delay=1).

References

1] Newell,A.(Chairman),et. al,, "Speech-Understanding Systems! Final report of a Study Group",
Comp. sci, Dept., Carnegie-Mellon University, Pittsburgh,May,1971.

2] Reddy,D.R.,Erman,L.D.,Neely,R.B.,"A Model and a System for Machine Recognition of Speech",
To appear in IEEE Trans. on Audio and Electroacoustics.

33 Walker,D,"Speech Understanding Through Syntactic and Semantic Analysis?
Proc. of the 3rd IJCAI, Stanford, California, Aug. 1973,

43 Woods,W. and Makhoul,J.,"Mechanical Inference Problems in Continuous Speech Understanding",
Proc. of the 3rdIJCAI, Stanford, California, Aug. 1973.

5] Lehiste,I.,"Acoustical Characteristics of Selected English Consonants",
Indiana University,1964.

6] Ohman,S.E.G.,"Coarticulation In VCV Utterances, Spectrographic Measurements",
Jour. of Acoust. Soc. of Am, 39, 151,1966.

7] Menon,K.M.N.,Jenson,P.J.,Dew,D.,"Acoustic Properties of Certain VCC Utterrances",
Jour. of Acoust. Soc. of Am, 46, 44% 1969.

8] Astrahan,M.,"Speech Analysis by Clustering of the Hyperphoneme Method",
AI Memo 124, Comp. Sci. Dept., Stanford University,California.

93 Thosar,R.B.,Samuel,A.L.,"Some Preliminary Experiments in Speech Recognition Using Signature Table Learning",
SUR Note 43,ARPA Network Inform. Center,NIC 11621, SRI,California.

10] Markel,J.D."Digital Inverse Filtering, A New Tool for Formant Trajectory Estlmati an",
IEEE Trans. on Audio and Electroacou., AU-20,1972.

11] Fant, C.G.M.,"Acoustic Theory of Speech Production",
Mouton,The Hague, 1970.

12] Thosar,R.B.,"Estimation of Probability Density Using Signature Tables for Application to Pattern Recognition",
AI Memo 198, Comp.Sci. Dept., Stanford University, California,