Segmenting Speech into Broad Phonetic Categories using Neural Networks

Murali Gopalakrishnan
B.Sc., University of Madras, India, 1984
B.E., Indian Institute of Science, India, 1987

A thesis submitted to the faculty
of the Oregon Graduate Institute
of Science and Engineering
in partial fulfillment of the
requirements for the degree
Master of Science
in
Computer Science and Engineering

August, 1990

The thesis "Segmenting Speech into Broad Phonetic Categories using Neural Networks" by Murali Gopalakrishnan has been examined and approved by the following Examination Committee:

Dr. Ronald A. Cole
Professor
Thesis Research Advisor

Dr. Todd Leen
Assistant Professor

Dr. Mark A. Fanty
Post Doctoral Fellow

ACKNOWLEDGEMENTS

I thank Ron Cole for giving me this opportunity to work in the exciting field of speech recognition. Without his guidance, endless patience and most of all kindness and support, I would not have been able to complete this work. I thank Mark Fanty for supervising my research work. He has been a constant source of encouragement and innovative ideas which have greatly helped me in my research work. Both Ron and Mark have taught me a great deal about how to be a good researcher. I also thank Yeshwant Muthusamy, Rik Janssen and other members of the speech recognition research group for their help in testing out the system and giving me feedback as well as new ideas as to improving the system. I thank Vince Weatherill for recording the speakers which provided me with the data I needed for the research, helping me label the utterances and for all the little things that are so important in bringing a thesis to completion. I thank John Inouye and Srikanth Kambhatla for their valuable advice, support and friendship that helped me keep my spirits up all through the period of study and research here.

I thank the faculty of OGI for allowing me this wonderful opportunity to study in the United States. I thank the staff and students too for making my stay here a pleasant one.

Table of Contents

ABSTRACT

Feature-based approaches [6] to speech recognition tasks focus attention on regions of the signal we as humans recognize as containing the important information. Explicit segmentation allows us to select features that are most important to recognition.

In this thesis we investigate the problem of segmentation and classification of speech into one of the four broad phonetic categories Sonorants, Fricatives, Closures and Stop consonants using neural networks as classifiers.

We first limit this task to spoken letters of the English alphabet. We use a back propagation neural network with conjugent gradient optimization. We choose a set of speech parameters best suited for the task. We then experimentally determine the configuration which best segments a section of speech and classifies it. The parameters we vary are the neural net characteristics, the set of features extracted from the speech parameters, and the methods of normalization of these features. We compare this system to speech hand-labeled by two experts in the field. We then compare the performance with that of a well tuned rule-based system on a letter recognition task.

CHAPTER 1

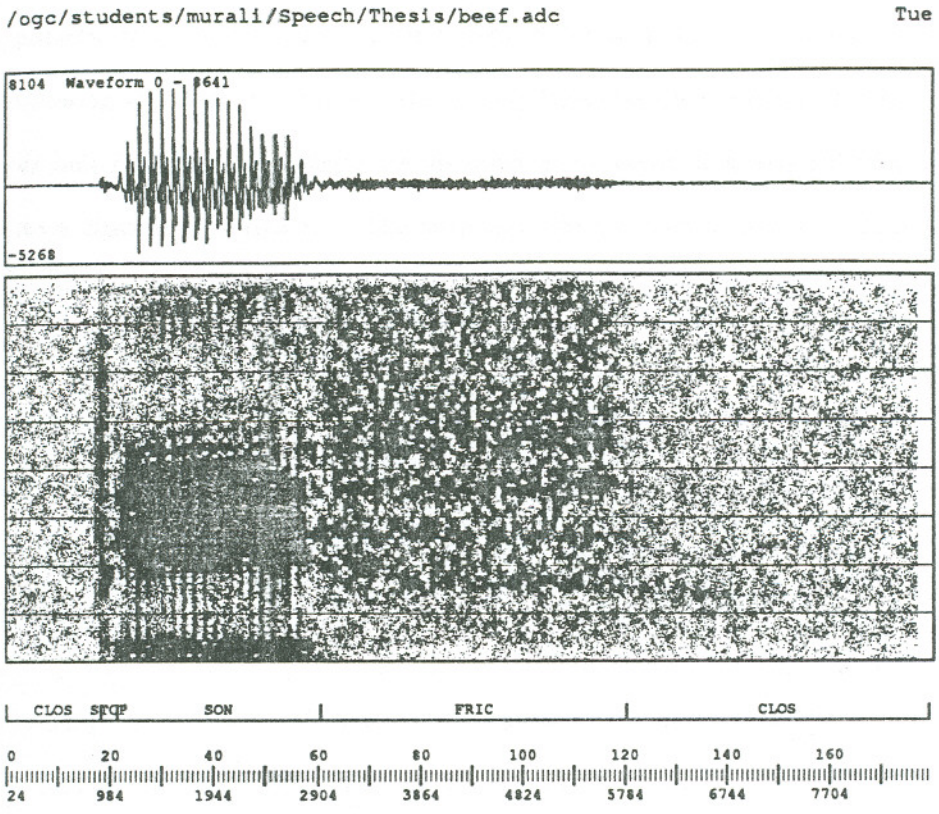
Introduction

1.1. The Problem

In this thesis, we attempt to segment speech into broad phonetic categories corresponding to the major articulatory behaviours that underly production. The research will apply a novel approach to the problem that combines application of knowledge-based features and neural networks. The algorithm will be evaluated in terms of human labeling performance and within working speech recognition systems.

Based on an analysis of the task described below, we have chosen the broad phonetic categories to be Sonorants, Fricatives, Stops and Closures. Figure 1.1 shows a waveform and a spectogram of the utterance "beef" which has been segmented and labeled by hand. The acoustic features describing the corresponding broad phonetic categories can be seen in this figure. **Closures** represent intervals of relative silence produced by a closure of the vocal tract preceding a stop consonant, a pause, or background noise. **Stops** are formed by a sudden release of pressure built up during the closure preceding the stop. If the airflow is not completely blocked, a high frequency hissing sound is produced. This is called a **Fricative**. If there is no restriction of the airflow, a fairly steady sound is produced. Such sounds are called **Sonorants**.

Stops are relatively short, often 9 to 12 msecs long, and therefore need to be detected with high accuracy. Moreover, they often occur before sonorants, the onset boundaries of which are also important to us. The sonorant, when preceded by a stop, contains a lot of information in the first few frames after its onset about the place of articulation



The utterance **beef** *showing the broad phonetic categories*

Figure 1.1

of the stop and it is crucial to detect these boundaries accurately. As an example, the rising pattern of the formants in the spectrum of the sonorant in the region immediately following the stop /*b*/ (figure 1.1) is a very important cue for discriminating /*b*/ from /*d*/ and /*g*/. Sonorant offsets, on the other hand, taper off gently till they disappear into a closure or a fricative. The drop off is sharper into a fricative. This boundary is often hard to place, even for experienced labelers. A fricative can be very long (/*s*/) or relatively short (/*jh*/). Fricatives are often distinguished by their length and their spectrum which gives clues as to the place of articulation. Often, voiceless (/*s*/) and voiced (/*z*/) are distinguished by their lengths. Though accuracy in determining the bounds of such segments is important, very high accuracy is not a necessity as the lengths differ by a fair number of frames.

In evaluating this system we distinguish the regions of the signal that are more important to recognition from those that are not as important, and we place more importance in the accuracy in recognition of the former.

1.2. Why segment speech?

The speech input is a digitized acoustic signal containing a very large number of data points. In order to find out what was said in that fragment of speech, we need to analyze the numbers that represent the signal. The problem is made more tractable when the signal is converted into a smaller number of basic units which preserves those characteristics of the signal from which meanings can be derived. This helps an automatic system to focus attention on regions of the signal we as humans recognize as containing the important information. It allows us to select features [1] [2] that are most important to recognition. For example, for stop consonants before vowels, we can select features at the burst release and just after the sonorant onset.

An example of a feature-based recognition system is EAR [3] , a neural network based English alphabet recognition system. Given the digitized representation of an utterance, the system classifies it as one of the 26 letters of the English alphabet. EAR uses a feature-based approach to classification. It selects important features from specific regions. It uses a segmenter to identify these specific regions. A comparison of EAR with other systems using Hidden Markov Models (HMM) to model speech is reproduced here from [3] in table 1.1. The success of the EAR system demonstrates the feasibility of using explicit segmentation, feature selection and neural network classification in computer speech recognition.

Recent letter classification results

Study	Conditions	Speakers	Approach	Letters	Results
Brown (1987)	20 kHz Sampling 16.4 dB SNR	100 speakers (multi-speaker)	HMM	E-set	92.0%
Euler et al. (1990)	6.67 kHz Sampling (telephone bandwidth)	100 speakers (multi-speaker)	HMM	26 letters + 10 digits + 3 control words	93.0%
Lang et al. (1990)	Brown's data	100 speakers (multi-speaker)	Neural networks	B, D, E, V	93.0%
Cole, Fanty (1990)	16 kHz Sampling 31 dB SNR	120 training 30 test (speaker-independent)	Knowledge-based features and neural networks	26 letters E-set B, D, E, V	96.0% 95.0% 94.2%

Table 1.1

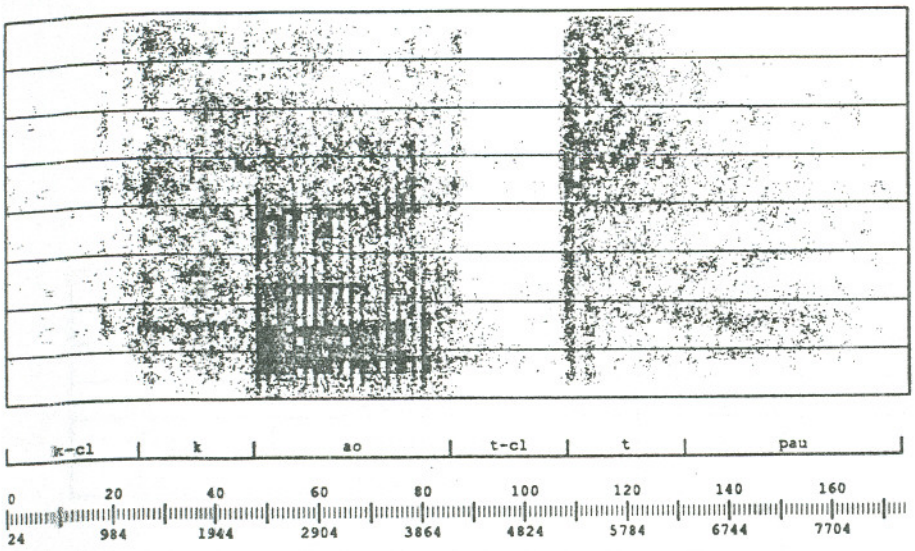
1.3. Approach

Our approach to segmentation is to use a neural-network classifier to classify each time frame of the signal into one of the four broad phonetic categories and then fix segment boundaries where the category changes from one frame to the next. Once segmentation and broad classification has been done, further processing of speech is much easier.

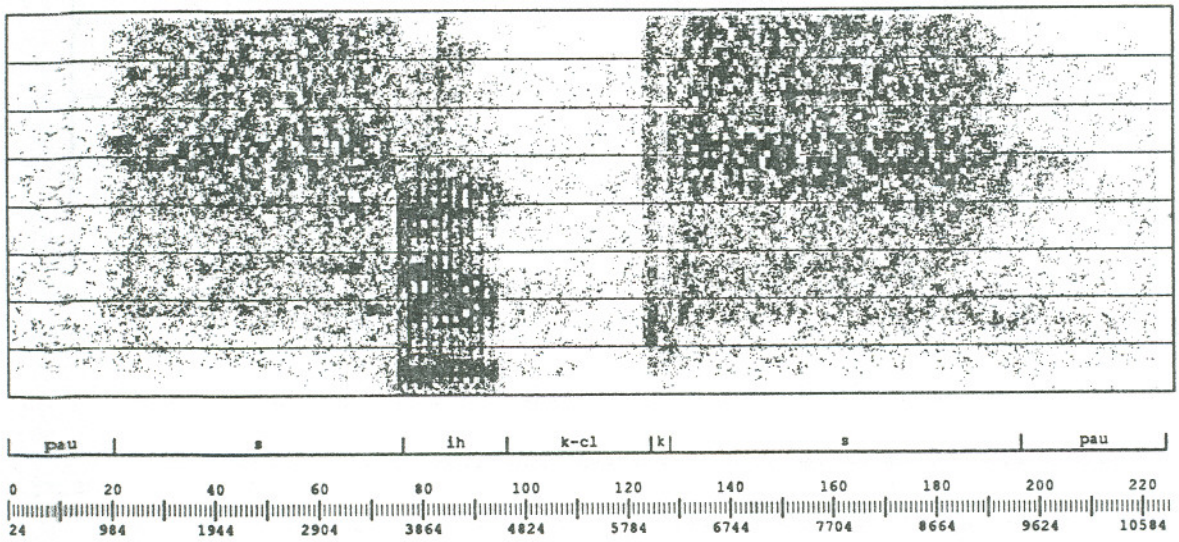
1.4. Choice of Task Domain

The goal of this thesis research is to help extend spoken letter recognition to continuous speech. A step in this direction is a system that can locate and classify letter strings spoken with pauses between the letters. In order to locate the letters it is necessary to segment speech, and it is sufficient if we do so into four broad phonetic categories: **stop, sonorant, fricative** and **closure**. In fact, with this approach we can perform context sensitive classification. To illustrate this let us take an example from continuous speech. The realizations of the Stop */k/* in *cot, scot, six, pick* and *risk* are all different as seen in table 1.2 and figure 1.2. With segmentation into broad categories, we can treat the realizations as different kinds of events and thereby make letter classification more reliable. We shall develop a neural network based segmentation and broad classification algorithm for this purpose. From later discussions it is seen that the rule-based segmenter used in the EAR system does not extend well to more complex domains. Hence, our choice of neural networks as the tool for segmentation and classification. The broad phonetic categories produced by the segmenter are intended to locate the phonetic events listed in table 1.3. (*Glot* is a glottal stop, the stop-like event that is typically observed before vowels). Preliminary results showed that voiceless stops were classified more easily as FRICs. From further research it was

/ogc/students/murali/Speech/Thesis/cot.adc Sat Aug 25 11:39



/ogc/students/murali/Speech/Thesis/six.adc Sat Aug 25 11:37:58 1990 (Page 1 of 1



Spectograms of cot and six
different realizations of the phoneme /k/

Figure 1.2

Allophones of /k/

Word	Characterisation	Broad Category Representation *
*c*ot	aspirated	C-S-V-C-S
s*c*ot	unaspirated	F-C-S-V-C-S
si*x*	affricated	F-V-S-F
pi*c**k*	unreleased	C-S-V-C-S
ris*k*	released	V-F-C-S-C

* C - Closure, V - (vowel like) Sonorant, F - Fricative, S - Stop.

Table 1.2

Phonetic events represented by Broad Phonetic Categories

Broad Category	Phonetic event it represents
STOP	b, d, and Glot
FRIC	p, t, k, jh, ch, s, z, f, v
CLOS	closures before stops, background noise, pauses, silence
SON	vowels, l, r, w, y, m, n

Table 1.3

found that, in spoken letters, such stops are always seen in word initial position and therefore have high frication, almost equivalent to that in fricatives. Further, the main difference betwen stops and fricatives is that the former are characterized as having a clear closure before them. In spoken letters, due to the pauses between letters, there is a clear closure before fricatives in word initial position too. Hence, in this task domain, unvoiced stops and fricatives look alike and are treated alike. The category of sonorants is left as it is.

1.5. Previous Work

1.5.1. Review of Early Work in Speech Recognition Systems

There have been a number of early attempts at word and continuous speech recognition using rules and statistical pattern classifiers with different representations of the signal. These were largely exploratory methods aimed at solving such problems as finding good representations of the signal, looking for basic elements of the spoken language, and finding ways to detect them in the signal using the constraints of the language.

The first ARPA project (1971 - 76) produced systems mainly based on statistical classification methods or AI techniques with rules built from spectral data or other features. Some examples of these systems are the Harpy system [4] , and HEARSAY II [5] both from CMU, and HWIM [6] from BBN. These systems were all medium vocabulary (about 1000 words), low perplexity (branching factor 33 except for HWIM where it was 196) and multi speaker (1 to 5 speakers) systems working on very specific tasks domains. Performances were in the range of 90% for the CMU systems and 44% for HWIM. The phoneme recognition accuracies were much lower at around 40% - 50%. Such a high utterance accuracy from the mediocre phoneme accuracies leads us to believe that we could build bigger and more versatile systems which can still have a high recognition rate by greatly improving the segmentation accuracies.

In the period following the first ARPA project, HMMs emerged as the leading technology. A notable exception in this regard was the Feature System [7] from CMU. HMM based systems use no explicit segmentation or speech knowledge to guide their search.

1.5.2. Review of Early work in Explicit Segmentation

There were a number of early systems for explicit segmentation. Some of these studies are shown in table 1.3. Most of these systems were designed to segment and classify speech into voiced and unvoiced [8] regions, and sometimes silence [9] too. Some of the efforts were hardware based [10] [9] , while others were based on rules or pattern classification. Some systems used hierarchial decision making strategies [11] [12] to aid classification. Some [13] performed broad classification as a step before phonemic classification. They used features mostly based on LPC and sometimes on DFT. Zero crossing count was also used as a feature by some. Some [9] came up with ingenuous features that work fairly well with hardware. However, most of the systems used very few speakers with very few utterances except for one system by Wilcox and Lowerre in 1986 [12] , which dealt only with continuous digits.

In 1978, Victor Zue [14] showed that

- phonetic structure is recoverable from the spectogram,
- explicit segmentation is a viable alternative in speech recognition, and,
- knowledge can be used effectively in speech recognition.

These results gave a big boost to research in feature based approaches [7] [15] to speech recognition problems, and explicit segmentation. With the development of neural networks and other better strategies for classification, the performance of segmentation and classification sub-systems at more complex tasks and in truly speaker independent environments began improving.

Author/Year	Task	Data	Features	Decision Method	Performance
Kasuya, H., Wakita, H. (1979)	Voiced and Unvoiced	10 utterances of continuous speech; 2 male, 2 female spkrs	LPC based features	rule-based decision	93.3% overall
Knorr, S.G. (1979)	Voiced & Unvoiced	12 sentences of continuous speech; 5 male, 3 female, 4 child spkrs	filtered spectrum features	hardware switching based decision	97%
Un, C.K., Lee, H.H. (1980)	Voiced, Unvoiced & Silence	25 seconds of continuous speech	bit-stream from linear delta modulation and zero crossing as features	hardware switching based decision	94%
Siegel, L.J., Bessey, A.C. (1982)	Voiced, Unvoiced & Mixed excitation	continuous speech; 8 sentences each from 4 spkrs train, 2 spkrs test	LPC and DFT based features with zero crossing	linear perceptron based decision at every node of a decision tree	94%
Regel, P. (1982)	1) classify into Silence, Unvoiced, Voiced Fricative, Unvoiced Fricative; 2) classify phonemes	continuous speech with 4 male & 2 female spkrs for 40 sentences (60% train, 40% test)	LPC based features	Bayesian classification	92% broad category, 60% phonetic
Wilcox, L., Lowerre, B. (1986)	Silence, Vowels, Nasals, Strong Fricatives, Weak Fricatives, Others	continuous digit database of TI	LPC based features and zero crossing	Gaussian classification at every node of a decision tree	87.4%

Early work in Explicit Segmentation

Table 1.4

1.5.3. Rule-based Segmentation

A fairly robust segmentation and broad classification system based on rules formed around different representations was built by Cole and Hou [15] at CMU in 1988. This system showed the advantages of a knowledge-based approach. Further, the broad classes are more or less distinct and fairly easily recognizable from the signal. They help constrain the search for phonemes and to some extent for word boundaries. The work on Cole and Hou's rule-based segmenter was extended to spoken letters in 1989 at OGI and a system to recognize letters of the English alphabet spoken in isolation [3] was developed. However, the disadvantage of rules was quickly apparent. As the number of speakers in the system increased, more variations of the signal were found, and though minor, they affected the rules. The rules had to be continuously tuned, and more rules had to be added as the domain grew. As is seen from results later in this thesis, rule-based segmentation performs well in smaller and more limited domains like isolated letter utterances. But, when we tried to extend it to connected letters – sequences of letters of the English alphabet spoken with distinct pauses in between the letters, the accuracy fell. The rules could be tuned to improve performance, but the effort required is far too much as the complexity of the domain increases.

1.5.4. Dendrograms

Another approach to segmentation is the use of a dendrogram [16] to represent multilevel hierarchies of segments. The basis of segmentation in this approach is the use of difference measures applied to some spectral representation of the signal. The difference in the spectrum across adjacent frames is computed using this difference measure, and similar frames are clustered together. This clustering is done repeatedly at higher and higher levels until the entire utterance forms one segment. An advantage here is that clustering is based on spectral differences in the actual signal. This results in segments smaller than phonemes. Therefore, few segments are lost due to

deletion. However, finding a good difference measure for speaker independent speech is difficult. Any general difference measure at this low level is prone to weighing unimportant differences equally with the important ones. The characteristic feature distinguishing one segment from another could get swamped by other insignificant but strong differences. This can happen when discriminating among some vowels. Another, more important, problem arises when trying to match the regions for likenesses to phonemes and finding a path through the network. The classification into phonemes is done by matching the spectrum within the segment with each of a set of prototype spectra for the phonemes. As in HMM based systems, here too it could fail because the characteristic similarities could be offset by insignificant differences like across speaker differences.

1.5.5. Neural network based systems

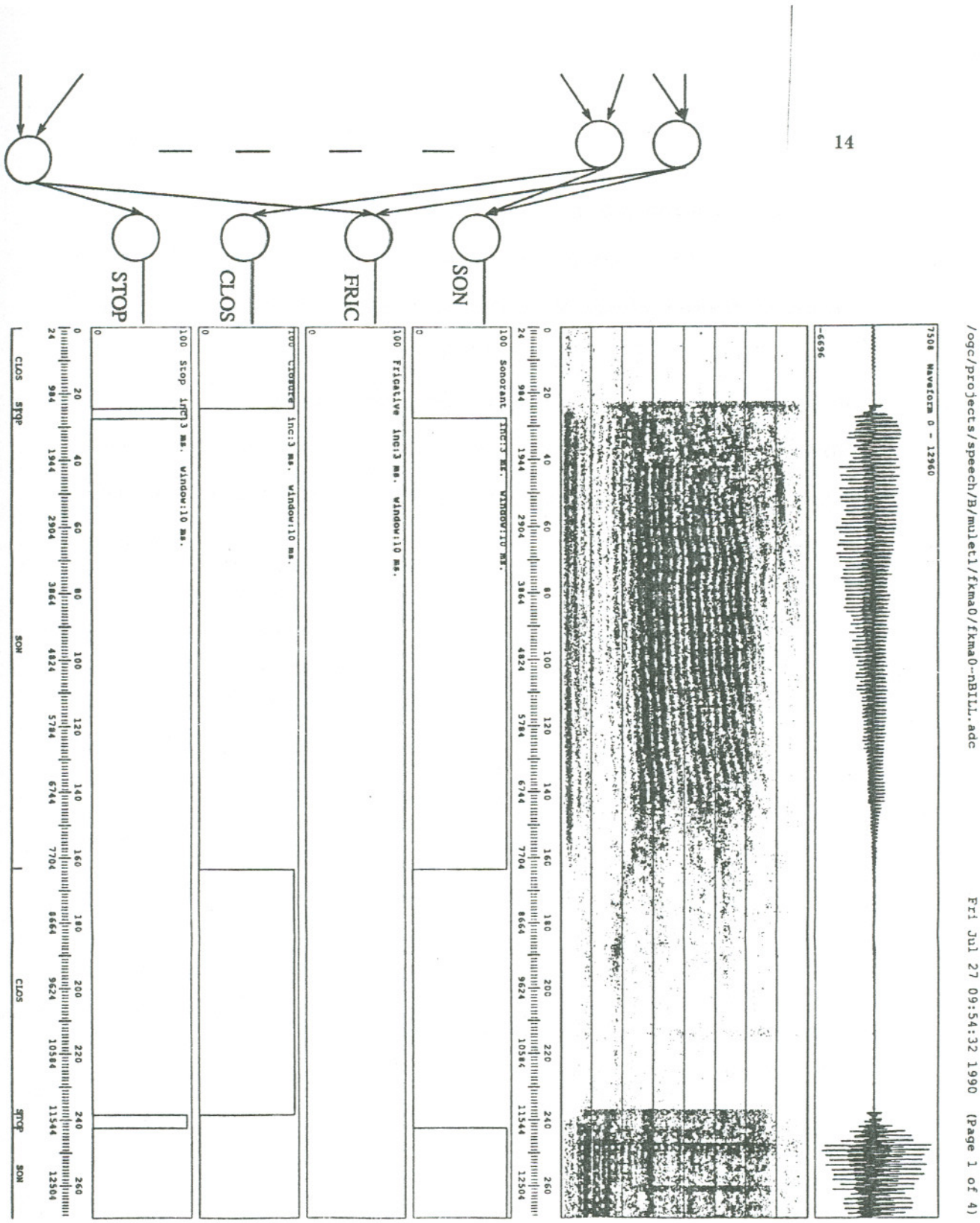
Neural networks, being powerful classifiers, have been recently used [17] [18] [19] to build segmentation and classification systems. The idea is to compute features from the signal at various time frames and input to a trained neural network. The net classifies the time frame as one of the categories specified. Identical adjacent frames are then grouped into segments of the category to which the net assigned them. One problem with neural network based systems is that they classify frames independent of each other. They do not provide a natural model of time and duration. Some methods have been developed by which time can be represented in the spatial organization of the network. The architectures either have the delay units and context represented in space [18] , or have recurrent links through delay units [20] into the input layer. A system [19] using the latter approach to segment and classify speech into seven broad categories: silence, unvoiced plosive, unvoiced fricative, voiced plosive, nasal like, sonorant like, and vocalic. Cepstral representation of the speech signal was used. The

classification was performed frame by frame with the cepstra of left and right context used along with the cepstrum of the frame in question. At the segment level, the errors were 7.9% insertions, 5.4% substitutions, and 11% deletions. The overall frame classification accuracy was 85.2%. Our approach is to provide a wide window of information at different resolutions and let the neural net capture the variations over time as represented in space.

1.6. Our approach and overview

We have seen on the one hand systems which make use of speech knowledge but which have no general procedure to learn the important characteristics that differentiate phonemes. On the other hand, there are systems which use procedures to learn characteristics automatically but do not use knowledge to highlight those characteristics and restrict the search. As we have mentioned earlier, neural networks are powerful classifiers that make no assumptions about the underlying probability distribution of the pattern space. Further, they can learn similarities and differences by correlating different kinds of features. They hence have the potential for using knowledge of speech if given in some acceptable format to discriminate different kinds of speech segments.

Our approach is to design good features for the neural network that capture the knowledge required to perform the discriminations, and to use those features to build (and train) a network to perform the classification. An idealized output of the segmenter is shown in figure 1.3. The ultimate goal is to build a segmenter that extends to continuous speech. In this thesis we start with spoken isolated letters and show how our approach easily extends to connected letters. The results have been sufficiently encouraging that work is now underway to extend this approach to continuous speech.



Idealized output of the segmenter

Figure 1.1

The organization of the thesis follows. In chapter 2, a detailed description of the algorithm that resulted from the research is presented. Chapter 3 discusses the design of the features and the research using neural networks to segment letters of the English alphabet spoken in isolation. In chapter 4, the system is extended to names and words spelled with pauses between letters. We also describe a method which helped us immensely — *training on errors*. The performance of the system is measured and the results are discussed in chapter 5. Chapter 6 discusses the approach and indicates future work in this area.

CHAPTER 2

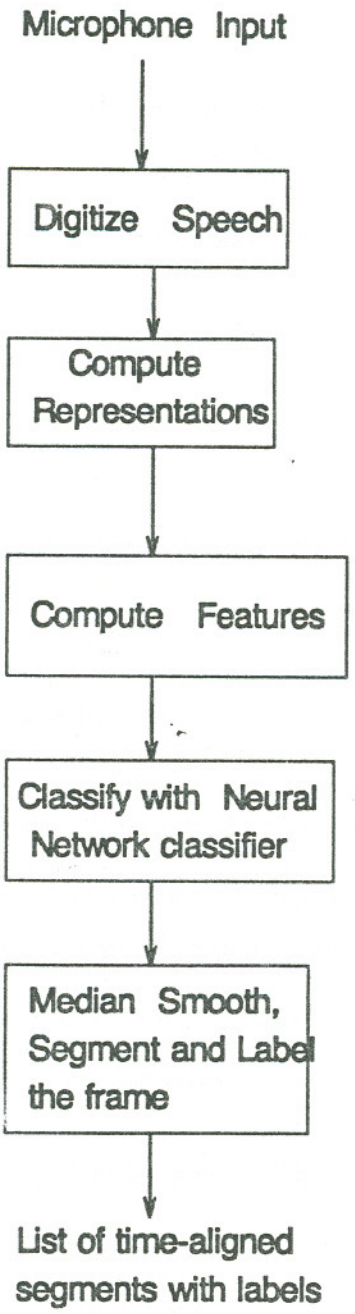
Description of the Segmentation and Broad Classification System

2.1. Overview

System modules that transform an input utterance into a contiguous sequence of broad phonetic category labels are shown in Figure 2.1. The sound input from the microphone is first digitized and stored as a waveform representation. Some useful representations of the signal are then computed. A set of features is computed on these representations within a 10 msec time frame every 3 msecs in the utterance, in a region around that frame. The feature vectors are input to a neural network that classifies the frame as one of four segment labels. The output activations from the network are then processed to determine the segment boundaries.

2.2. Data capture

Speech is recorded using a Sennheiser HMD 224 noise-canceling microphone. It is lowpass filtered at 7.6 KHz and sampled at 16 KHz per second with 16 bits used to represent each sample. Data capture is performed using the AT&T DSP32 board installed in a Sun4/110. The digitized utterance is stored in a buffer (2 sec long for isolated letters and 6 sec for connected letters) using the WAVES+ software distributed by Entropic systems. In order to speed processing time, the utterance is located within the buffer based on values observed in two waveform parameters, the zero crossing rate and peak-to-peak amplitude. The remaining representations, such as the DFT, are then computed in the region of the utterance only. In [3] a more detailed



Functional block diagram of the segmenter

Figure 2.1

description of the recording environment and the data capture process for isolated letters can be found. The data capture process for connected letters differs from that for isolated letters only in the size of the storage buffer.

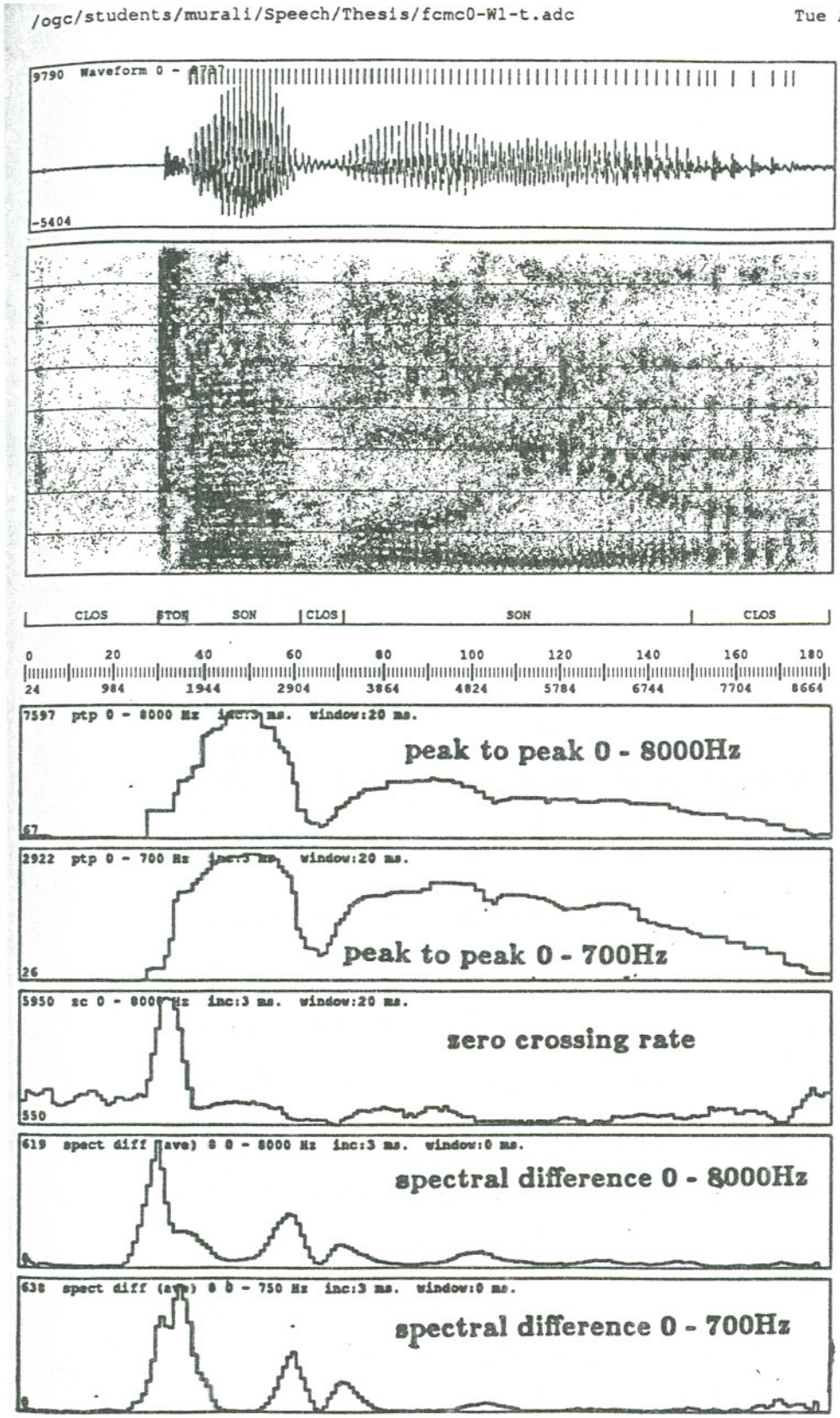
2.3. Signal Representations

The representations are shown in Figure 2.2. We can broadly divide the parameters into waveform parameters computed from the digitized waveform and spectral parameters computed from a 128 point discrete Fourier transform computed every 3 msecs with a 10 msec Hanning window.

2.3.1. Waveform Parameters

- **peak-to-peak amplitude 0 - 8 KHz** (ptp 0 - 8000) — The peak-to-peak amplitude is the difference between the maximum positive and maximum negative peaks of the original waveform in a 10 msec (160 adc points) wide window. As seen in figure 2.2 this parameter gives a measure of the waveform envelope and is a very good indicator of silence in clean (high S/N ratio) speech.

- **peak-to-peak amplitude 0 - 700 Hz** (ptp 0 - 700) — This parameter is computed from the waveform low-pass filtered below 700 Hz, which is the range in which the first formant (the lowest resonant frequency of the vocal tract), is located. Since formants are more salient during periodic signals, this parameter gives a fairly good estimate of sonorant intervals; most disruptive noise and other aperiodic signals have energy at higher frequencies. Periodic signals are comprised largely of sonorants but also include voicing in fricatives (/*v*/ and /*z*/) and prevoicing - which we classify as closure.



Parameters used for segmentation

Figure 2.2

- **zero crossing count** (zc) – The zero crossing count is the number of times the waveform crosses the zero line in a 10 msec window. A high zero crossing implies high frequency and therefore frication. Zero crossing is largely independent of the power or amplitude of the signal.

- **pitch** – Pitch is computed using a neural network pitch classifier [21] to locate peaks in the filtered waveform that begin pitch periods. It is a good indicator of periodic signals.

2.3.2. Spectral Parameters

- **spectrum** – The spectrum is obtained from a 64 point DFT computed every 3 msecs with a 10 msec Hanning window.

- **spectral difference 0 - 8 KHz** (SD 0 - 8000) – The spectral difference is computed as the mean squared difference of the spectrum averaged N frames before and N frames after the frame under consideration. In this project N is set to 8. This number was chosen after performing some experiments to determine the efficiency with which various widths accentuated stops and sharp spectral changes while smoothing out the effects of smaller changes. The spectral difference indicates changes in spectral energy from region to region. Sharp spectral changes indicate the presence of a stop burst in that region. Significant, but not very sharp, spectral changes also indicate changes in the signal patterns and are therefore useful in locating boundaries in speech.

- **spectral difference 0 - 700 Hz** (SD 0 - 700) – The spectral difference computed from the DFT coefficients below 700 Hz is used to help determine the onsets of voicing, especially Sonorant onsets. The Sonorant onset has been found to be most

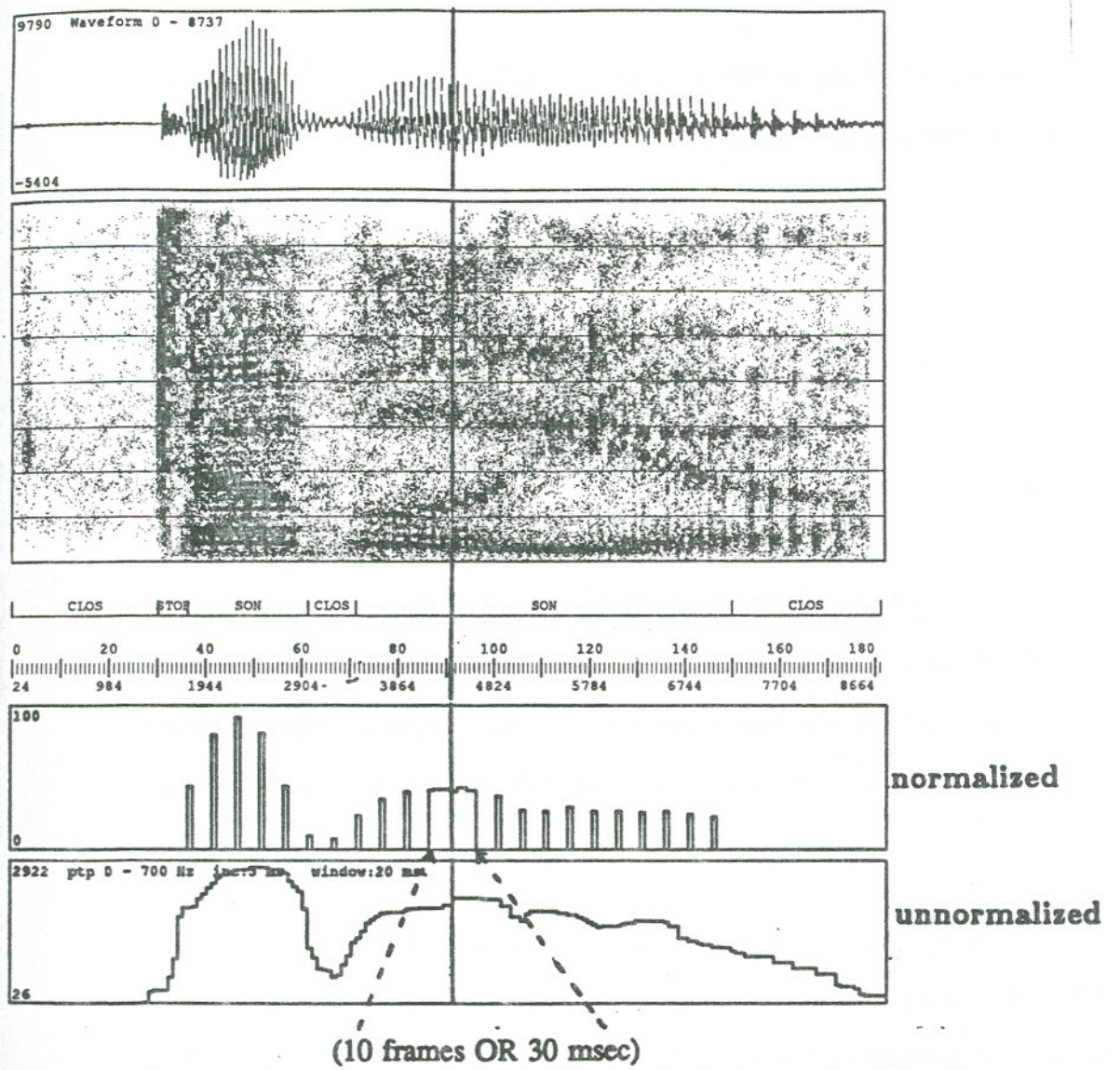
fundamental in the interpretation of the speech signal and it is very important to locate it with high accuracy.

2.4. Feature Measurement

Feature measurements are derived from the above parameters to provide pattern descriptors sufficient to classify each time frame as one of four broad phonetic categories. Figure 2.3 shows how the peak-to-peak 0 - 8000 Hz features are extracted from the corresponding representation. Feature measurements are taken from a window of 330 msec centered on the frame to be classified. Within this window, feature measurements are computed at two levels of resolution:

- **immediate context** – In the 30 msec surrounding the frame, features are sampled at every 3 msec frame. This produces 10 feature values for each parameter in the immediate context of the frame.

- **surrounding context** – In the rest of the region within the window, information is taken at a lower resolution of 5 frames (15 msec) per sample. The way the information is chosen in this low resolution region depends on the parameter from which the sampling is done. For the features extracted from waveform parameters except pitch the average of the parameter values in the 15 msec region is taken as the sample value. For pitch, the sample at every fifth frame is taken to be the representative sample of the region covered by the five frames. Since spectral difference is used only to determine whether or not there is a large spectral change in a region, the maximum value over 15 msec is taken as the sample value. This produces 20 feature values for each parameter – 10 for each 150 msec on either side of the immediate context of the frame.



SELECTION AND NORMALIZATION OF PARAMETERS

(eg. Peak - to peak 0 - 8000 Hz)

Feature computation

Figure 2.3

In this way we get 30 samples from each of ptp 0-8000Hz, ptp 0-700Hz, zc, pitch, SD 0-8000Hz and SD 0-700Hz which totals up to 180 features. The spectrum at the frame to be classified adds 64 more features totaling 244 features in all.

2.5. Normalization

Normalization of feature values is necessary to train the neural network and to accentuate differences among the categories. Normalization is performed by examining histograms of the normalized feature values and adjusting the normalization to optimize discrimination among categories. At this stage, knowledge of acoustic phonetics guides the research. For example, we know that zc should discriminate Fricatives from Sonorants and Closures, SD 0-8000Hz should discriminate Stops from the rest, and ptp 0-8000Hz should discriminate Closure from the rest, while ptp 0-700Hz should discriminate Sonorants from the rest. This knowledge engineering is an essential feature of the approach.

Each feature is normalized differently. In order to generalize the procedure so as to easily extend to natural continuous speech, a window is defined for normalization. The window is 250 msec (83 frames) ahead of (after) the frame to be classified and 300 msec (100 frames) behind the frame to be classified. The numbers were chosen based on psychological observations [22] on human short term memory. The methods used to normalize each feature are described below. Note, pitch is not normalized as it is either 0 (absence of pitch) or 1 (presence of pitch).

- **peak-to-peak** (both frequency ranges) – The maximum and minimum in the window are determined. The values are then normalized according to the formula (value - min) / (max - min). Here, normalization helps to remove differences due to loudness of speech. To protect against improper accentuation of silence when the

whole window is inside silence and the max & min values are almost the same, a threshold is used. If (max - min) is less than the threshold, the threshold is used as the divisor. The threshold was chosen after performing some statistical studies on the values from the utterances in the training data. The threshold is different depending on the variation of ptp used.

- **zero crossing count** – The minimum in the window is determined. The values are then normalized according to the formula (value - min) / DIVISOR. DIVISOR was chosen after studying the statistical variations of the values from the utterances in the training data. The zc is the same irrespective of loudness as mentioned in section 2.3.1. Therefore, it is not normalized with respect to the maximum value in the window.

- **spectral difference** (both frequency ranges) – The values are normalized according to the formula: value / DIVISOR. DIVISOR was chosen after studying the statistical variations of the values from the utterances in the training data.

- **spectrum** – The mean and standard deviation of all the coefficients within the window are determined. The maximum value for the window is set at (mean + 2 * standard deviation) and the minimum value for the window is set at (mean - 2 * standard deviation). The values are then normalized according to the formula (value - min) / (max - min). Though normalizing within the frame helps accentuate frequency bands better, it is found to cause problems in weak signal areas and in areas where the signal is absent (silence). To protect against improper accentuation of silence when the whole window is inside silence and the max & min values are almost the same, a threshold is used. If (max - min) is less than the threshold, the threshold is used as the divisor. The threshold was chosen after performing

some statistical studies on the values from the utterances in the training data.

2.6. Neural Network Classifier

A fully connected feed forward network was trained using the error back propagation [23] with conjugent gradient [24] optimization. The network has three layers including the input and output layers. The output layer has four neurons, one for each class. The number of neurons in the hidden layer was empirically determined to be 16. The number of neurons in the input layer is equal to the number of features in the input space.

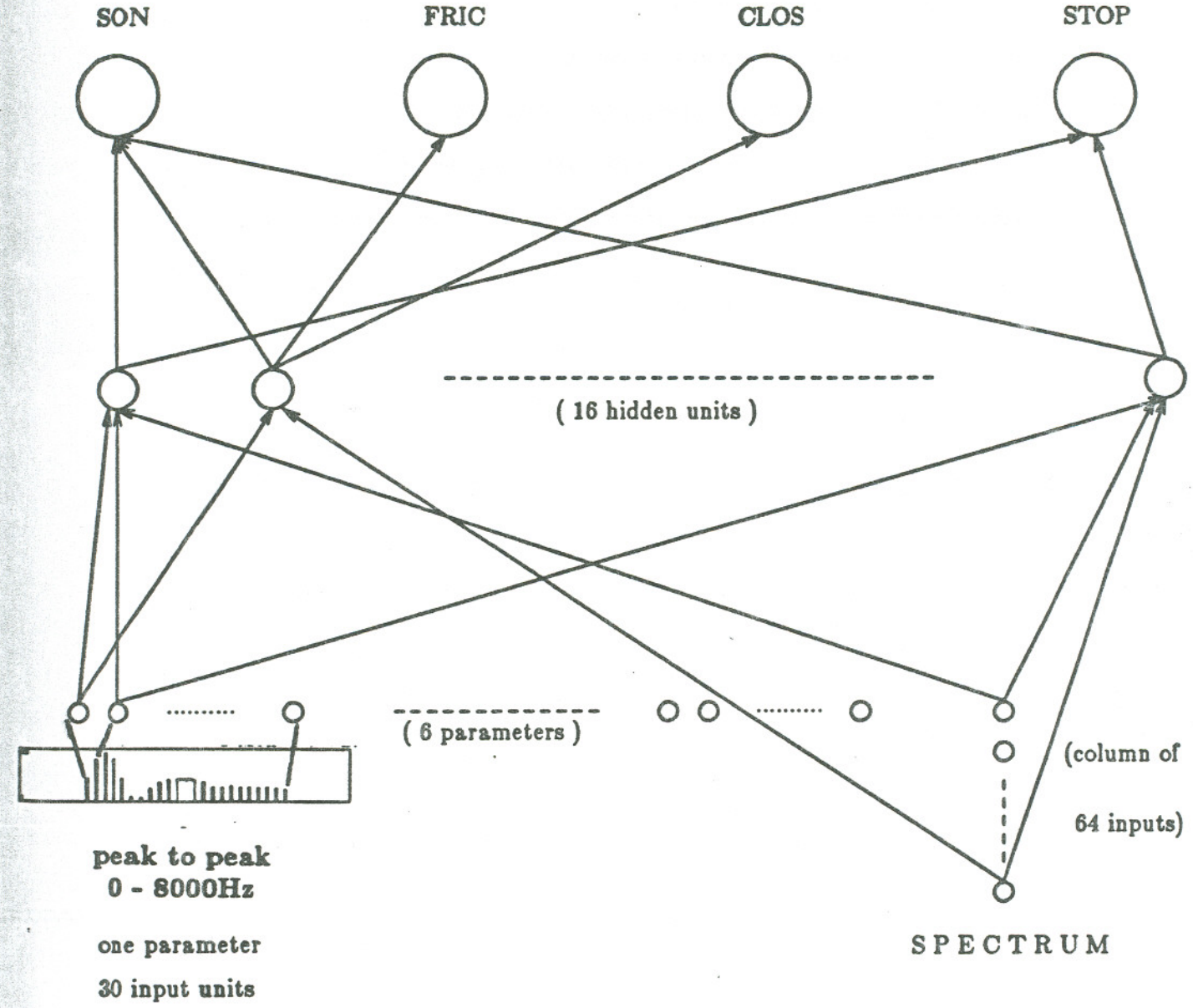
For each frame in the utterance, feature vectors were computed with that frame considered as the center frame in the window, and the output activations of the output layer neurons were recorded in memory. Figure 2.3 shows the computation of one of the features and figure 2.4 shows how the features are input to the neural net.

2.7. Output Processing, Boundary Placement & Labeling

Figure 2.5 shows the output responses of the network at each time frame. A 5-point median smoothing is done separately on the activations of each output neuron. Then, for each frame, the label corresponding to the neuron with the highest activation is chosen as the label for that frame. Finally, a pass is made over all the frame labels and segment boundaries are placed wherever there is a change in the label from one frame to the next.

2.8. Post Processing

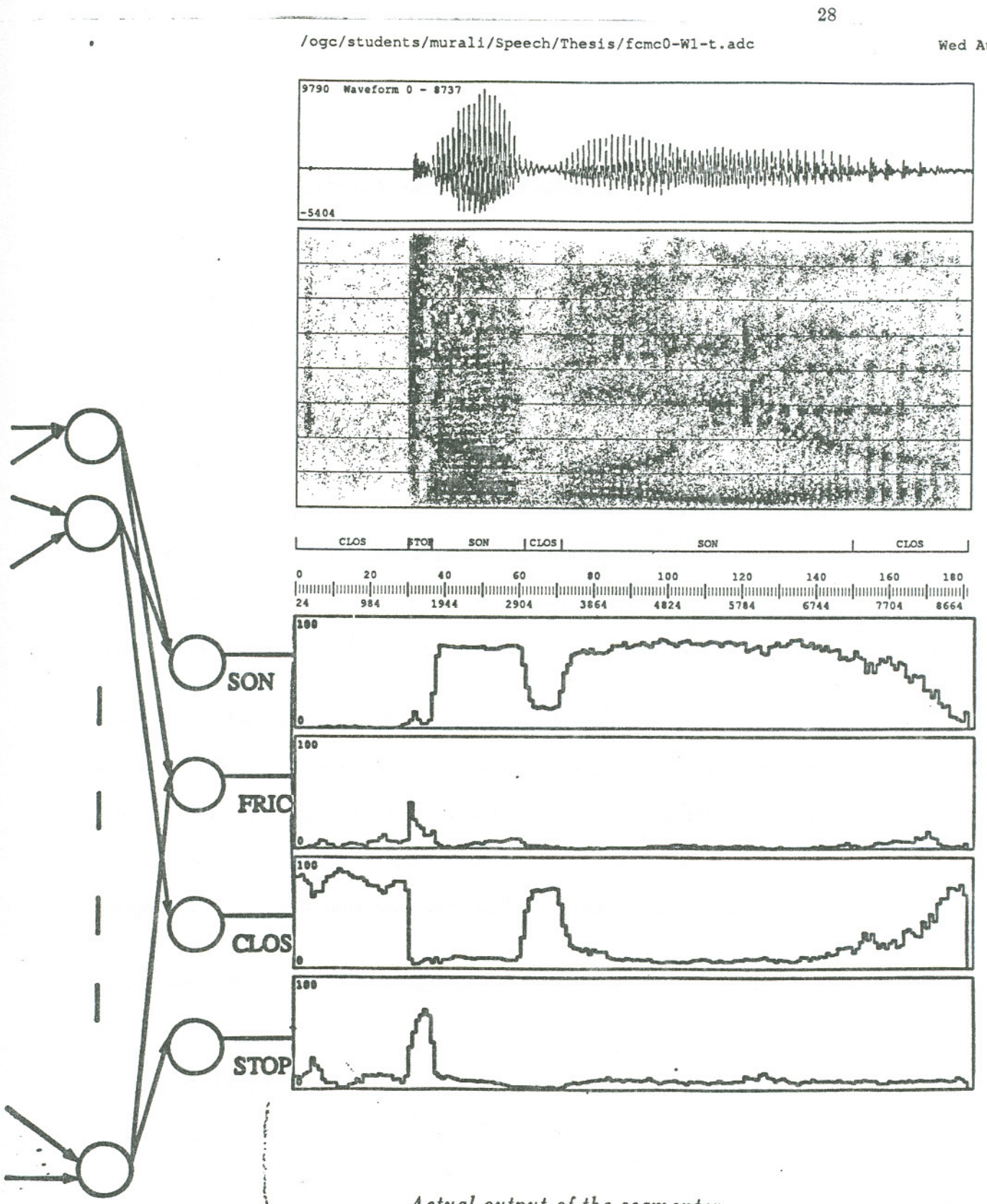
Sometimes, especially near the boundaries, certain frames get misclassified. In all these cases it has been noticed that the activation from the neuron representing the correct label is close to the maximum. This aberration is therefore due to transients in the



Features input to the neural network

Figure 2.4

signal, and the simplicity of the rule used to choose a single class for the frame. Therefore, the segmentation is cleaned up using simple rules based on the statistics of the data. Alternations of short Sonorants and Closures occuring mostly near the end of a Sonorant segment are merged into that segment. Short Closures and Fricatives occuring anywhere in the utterance, especially as insertions in other segments, are removed. The minimum lengths of valid Sonorants, Closures and Fricatives are determined from the hand-segmented training data.



Actual output of the segmenter

Figure 2.5

CHAPTER 3

Isolated Letter Segmentation

This chapter describes the initial research, using isolated letters, that led to the segmentation algorithm described in the previous chapter.

3.1. Database

The training data consists of spoken letters from the **ISOLET** database [25] , recorded in the Speech Laboratory of the Department of Computer Science and Engineering at the Oregon Graduate Institute. It consists of recordings of 150 speakers saying each letter of the English alphabet twice. There are an equal number of male and female speakers in the database. All the speakers are native speakers of English.

In order to train the neural network it is necessary to define two data sets, one for training and another for testing to check for generalization. The utterances from the training set are used to generate features to train the neural net using back-propagation as the learning technique. The test, or cross-validation set is used to determine the amount of training that provides the best generalization on new data.

Training set *isolet1*, consisting of 10 male and 10 female speakers, was used for training the network.

The following 20 letters (utterances) were chosen

A, B, D, F, G, H, K, L, M, N, O, P, R, S, U, V, W, X, Y, Z.

for each of the following 20 speakers

fcmc0, fcmg0, fdcf0, fec0, fet0, fews0, fjw0, fka0, fkh0, fmb0, mjc1, mjfv0,

mjp0, mjrs0, mnjh0, mnre0, mrmh1, mrs0, msa0, mtdw0.

from the isolet1 database.

For the **test set**, we attempted to use the same list of letters as above from *isolet0* and *isolet2*. However, the isolet0 database being incomplete, some utterances were missing from some speakers. Following is the list of speakers with a list (if any) of missing letters given in parenthesis after the speaker.

fbjt0 (D), fbpr0, fcch0 (Z), fcm1 (K,N), fcr0 (G), fdc0, fdec0, fdw0, mdf0, mhhp0 (M), mjh0 (B), mjms0 (N), mlht0 (D), mmcm0, mmr0.

In order to compensate for the missing letters and bring the test set to size, these letters were taken from another speaker.

mpak0 – B, D, G, K, L, M, N, P, V, Z.

All the utterances in both the sets were hand-labeled by the author and verified by an expert labeler. Table 3.1 shows the distributions of labels for the 400 utterances in the training set. It can be seen that all the utterances contain *SON* and *CLOS*; *STOP* and *FRIC* occur in less than half of the utterances.

Label	Speaker	Utterance	Frames
SON	20	400	2101
FRIC	20	176	1050
CLOS	20	400	2562
STOP	20	180	685

Distribution of labels in the training set

Table 3.1

3.2. Preliminary Experiment

The goal of the first experiment was to investigate neural network configurations, feature measurements and normalization procedures using a small subset of the data. The goal in subsequent experiments was to determine features and related normalization methods that produce the best results on isolated spoken letters. Figure 3.2 shows the output of the neural network arrived at after all the experiments.

The network was trained with three parameters:

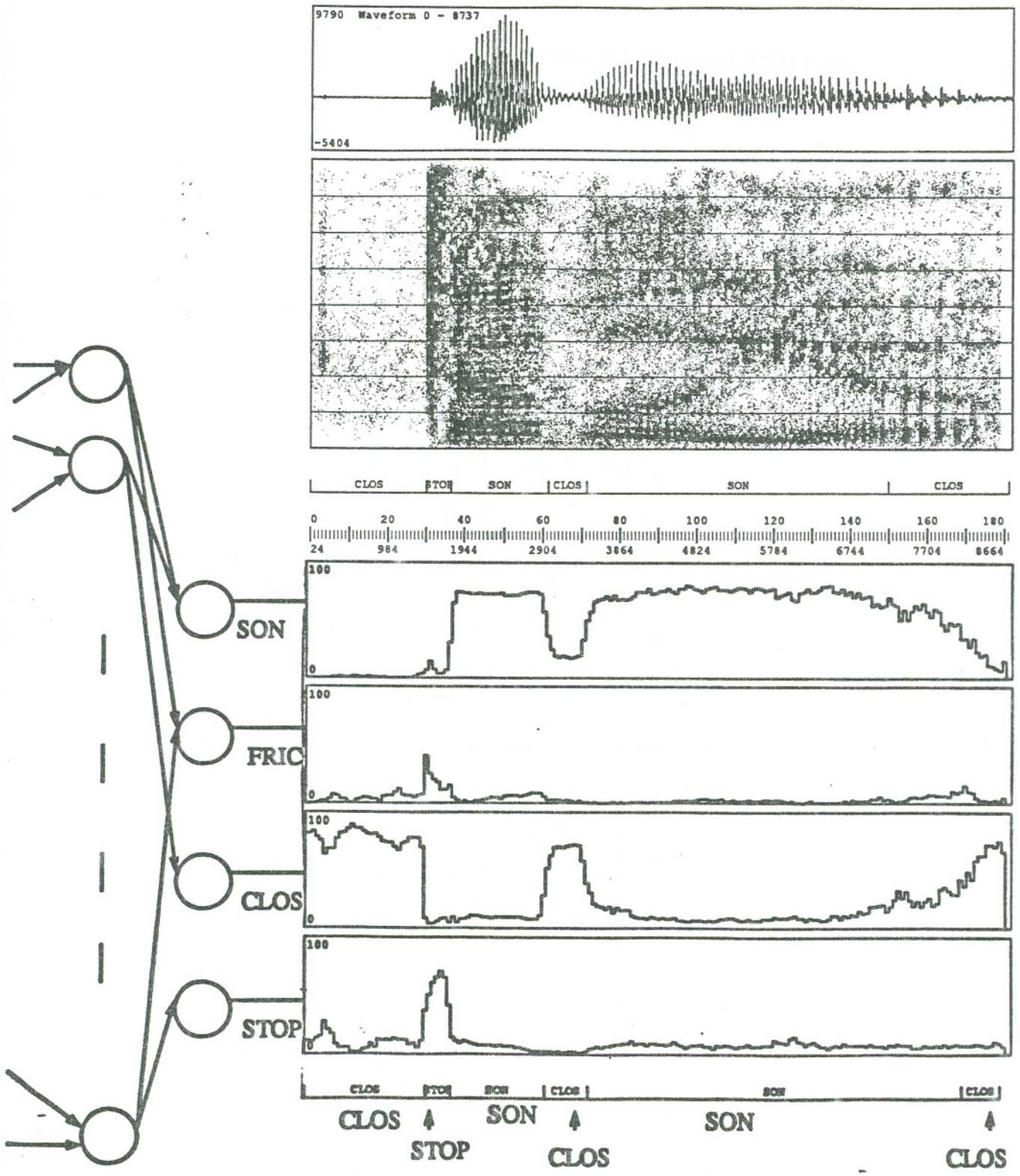
(a) Peak to peak 0 - 8000Hz,

(b) Peak to peak 0 - 700Hz, and

(c) Zero crossing rate 0 - 8000Hz.

Features were computed from these parameters on a 30 msec window centered around the frame to be classified. Taking all the frames from each training utterance provides too many feature vectors to train neural networks in a reasonable amount of time. Thus we sampled a subset of the frames in each utterance based on a hand segmented and labeled version of it. After some statistical studies, the optimum number of samples to be taken from segments of each class was determined and kept constant for all future experiments. In order to take into consideration boundary conditions, for each segment one sample was taken from near each of the two boundaries. One vector was generated for each sample taken from the training or test data.

Statistical methods were used to determine how well the features discriminate among the target classes. For each feature, histograms of feature values for each category were generated as shown in tables 3.2 and 3.3. From these tables we can clearly see how ptp 0-700Hz helps identify Sonorants while zc 0-8000Hz helps identify Fricatives.



Output of the isolated letter segmenter

Figure 3.2

Stops too have some high zc values due to the frication immediately after the burst and before the Sonorant, but they can be distinguished from Fricatives by their duration which is much smaller than that of Fricatives. Based on these histograms the feature selection and normalization techniques for each feature were then modified to help discrimination.

Histogram of ptp 0-700 taken from the center frame

feat 15	SON	FRIC	CLOS	STOP
= 0.0	0	7	79	0
< 0.1	22	726	1965	227
< 0.2	98	195	318	172
< 0.3	210	64	136	103
< 0.4	206	25	39	89
< 0.5	229	12	14	39
< 0.6	220	8	5	38
< 0.7	240	6	4	14
< 0.8	255	1	1	1
< 0.9	252	2	0	0
< 1.0	317	1	1	1
= 1.0	52	3	0	1

Table 3.2

Histogram of zc 0-8000 taken from the center frame

feat 75	SON	FRIC	CLOS	STOP
= 0.0	11	2	208	0
< 0.1	393	18	1432	26
< 0.2	832	47	529	163
< 0.3	561	69	223	194
< 0.4	213	105	123	118
< 0.5	59	98	29	69
< 0.6	16	142	13	49
< 0.7	16	138	4	26
< 0.8	0	105	1	10
< 0.9	0	88	0	18
< 1.0	0	75	0	7
= 1.0	0	163	0	5

Table 3.3

Classifier

The neural network simulator was initialized with small random weights. The network was run for N iterations, where N was determined from the experiments. After each set of iterations the data was tested on both the training and test data to check for convergence and generalization respectively. After the first training run, subsequent runs were performed with the weights obtained from the previous run - i.e. the neural net continues training from where it left off. The overall performance of the net when tested on the training and test data after each training run were plotted. The curve for the performance on the test data is expected to rise initially and then fall after peaking at some point. This point is where the net generalizes best. In the runs after this peak, the net begins to overlearn. The trained network obtained at the point of maximum generalization was chosen as the best network from that experiment. After each experiment, the network was studied using a visualization tool to check whether it had indeed learned to make the discriminations expected from the feature(s) introduced in that experiment. The tool was also used to make sure that each new feature did not work at cross-purposes with the features introduced earlier. Figure 3.3 shows the excitations and inhibitions of the activities of the different neurons in the hidden layer that are needed in order to identify a sonorant. The activations from the input that excite one of the neurons in the hidden layer which excites the sonorant neuron in the output layer more than others is also seen in that figure. Since this neuron in the hidden layer also activates closure to some extent, we can say that this neuron probably encodes information to recognize a sonorant near the offset of the sonorant. This is vindicated by the fact that peak-to-peak 0 - 700 Hz is excited for the current frame and for the past frames, but is inhibited for the frames that come later. Peak-to-peak 0 - 8 KHz is inhibited strongly in the immediate context while it is excited in the far context that is nearer to the frame being classified. Spectral difference is excited in the

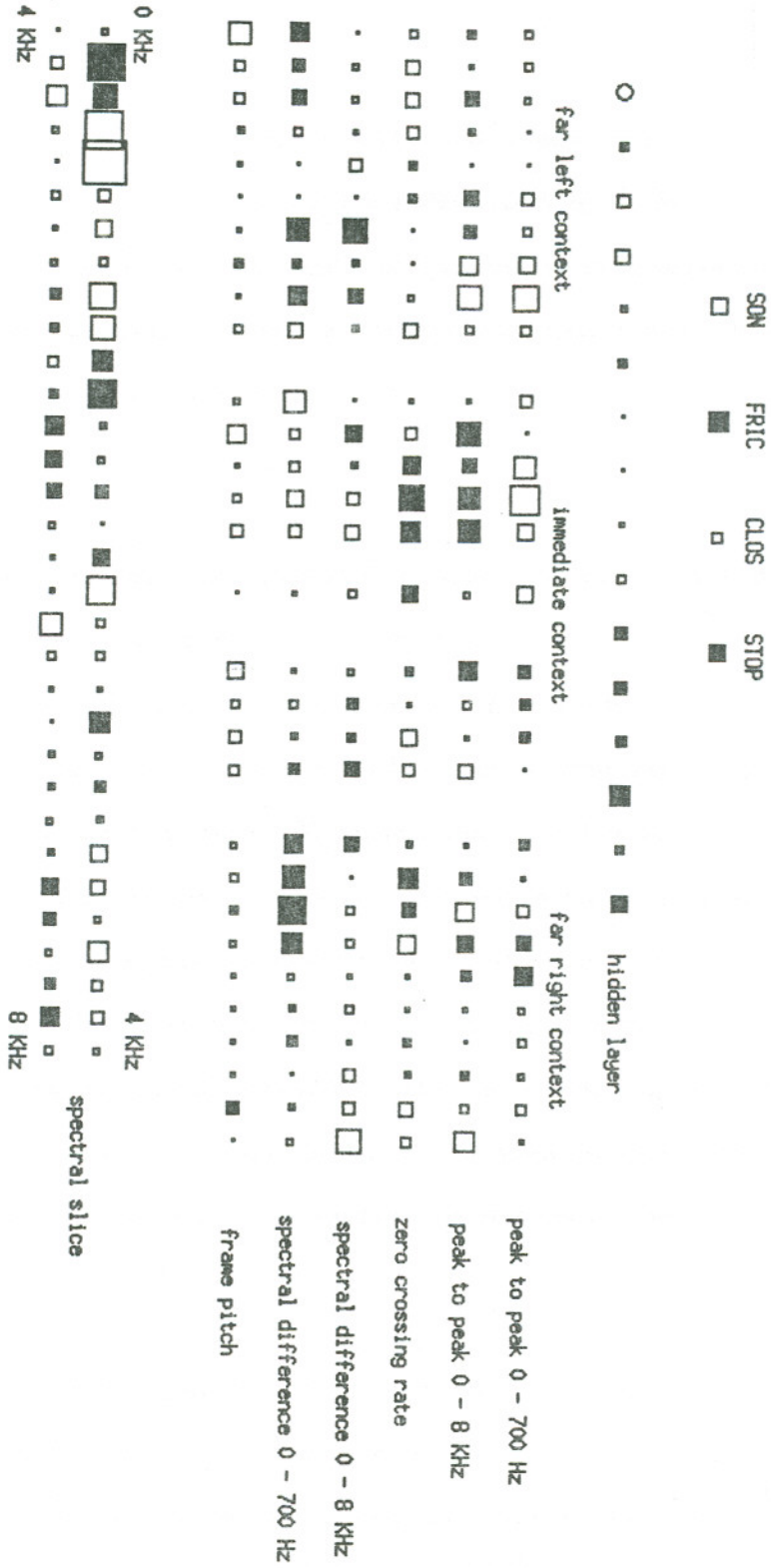


Figure showing which hidden units excite and which units inhibit the classification of a SON; and activation patterns of the input units to one of the hidden units that excites SON more

Figure 3.3

near context, and the spectrum values are excited in the formant regions, especially the for first formant. Figure 3.4 shows only the activations from the input layer to another neuron in the hidden layer that excites only sonorants. Peak-to-peak in the low frequency and the full frequency ranges are excited in the near context, and so is spectral difference in the low frequency range.

Results

The neural network thus obtained was evaluated on random utterances drawn from a test set not previously seen. The preliminary experiment showed that the best network was the one with 16 hidden units in one hidden layer. The simulator uses conjugent gradient optimization, and the information needed to keep the gradients conjugent is not stored at the end of the N iterations. This implies that on each restart, the search in the simulator is at first identical to simple gradient descent and is therefore slow. The speed of optimization is thus lost at every restart. Hence N needs to be large enough for the number of restarts to be fewer before learning is complete. However, we need to keep N sufficiently small to be able to detect more accurately the point at which the net generalizes best. A cycle of 80 iterations between test check points was found to be optimal. The normalization methods for the features are described in the previous chapter.

Another important observation made at this point was that the percentage of correctly classified frames was given out by the simulator could only be used as a rough indicator of performance. The reason for this is that certain kinds of errors are more acceptable than others. For example, if the offset boundary of a Sonorant is in error by 10 frames, the count of errors in frames immediately goes up. However, such an error can be tolerated as almost always, this error occurs at the boundary with a Closure. This

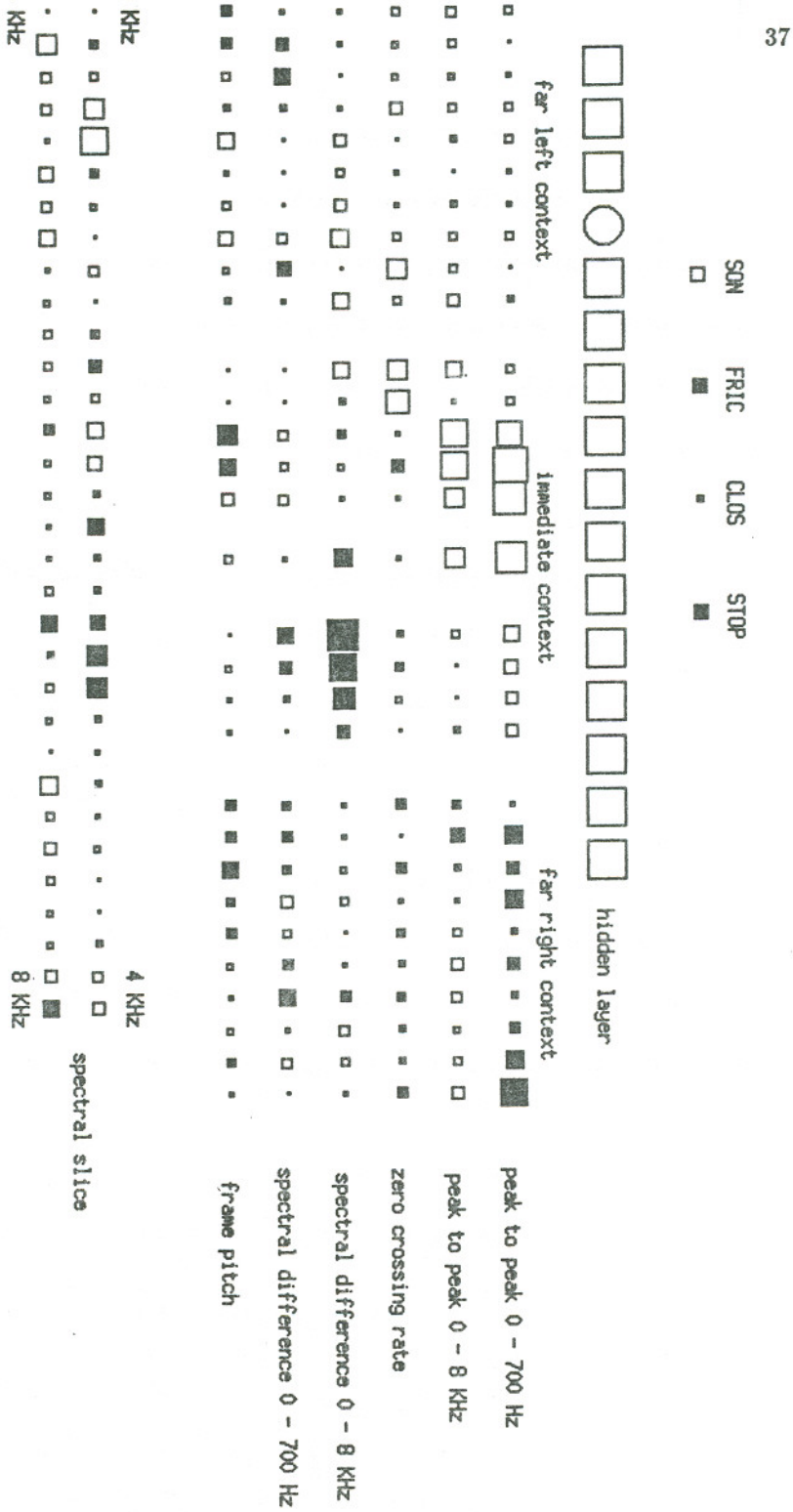


Figure showing activation patterns of the input units to another of the hidden units that excites SON

Figure 3.4

Closure is also, most often, the pause between letters. The Sonorant offset extending into the Closure does not really affect recognition, as most of the information is taken from near the beginning of the Sonorant. Hence, an error in the onset is far more serious than one in the offset. Stops get misclassified too. But it is imperative that we find a /*b*/ or a /*d*/ wherever there is one, whereas if we miss a *glottal* stop, the error is hardly serious. The letter can be identified in spite of such an error. For this reason, in addition to the confusion matrix output by the simulator, the segmenter is run on a separate set of test data and the output is analysed for the kinds of errors made. This process is crucial in our evaluation of the success or usefulness of a strategy or a feature that we may have used while building the system.

3.3. Experiment 1: Minimum features

The above experiment was then repeated on a larger scale with the full complement of speakers and utterances as described earlier under section 3.1 of this chapter. The best result is shown in table 3.4. Each row in table 3.4 represents the classification of the network for vectors belonging to the class representing the label in the first column.

Best overall result on test data after convergence

overall % correct = 88.28					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1483	22	46	24	94.2
FRIC	36	661	73	34	82.2
CLOS	167	18	1718	23	89.2
STOP	63	42	20	417	76.9

Experiment 1

Table 3.4

The network converged after 720 iterations and had an overall percentage correct rate of 88.28% on the test data and 91.64% on the training data. On running the segmenter program on some randomly chosen test utterances it was found that Closures were being introduced in the middle of Sonorants whenever there was an amplitude dip. In addition, Stops had a very low classification percentage. Furthermore, there were a large number of boundary errors. Examination of visual displays of the segmentation errors revealed that the network did not have sufficient information about the context to the left and to the right of the frame to be classified.

3.4. Experiment 2: Adding Context

The goal of segmentation is to locate boundaries between regions having different characteristics. Most segments are long with very little change in signal characteristics inside the segment. For example a Sonorant may have formant movement inside of it, but will not contain a burst or high frication anywhere inside it. A window as small as 30 msec can therefore not capture sufficient information about previous segments or later segments and can therefore not use the contextual information in classification unless it is near a boundary.

This prompted the use of broad context information in a 150 msec window on either side of the 30 msec immediate context window as described in the previous chapter. The number of samples per parameter was thus increased from 10 to 30. The same network parameters were found to be sufficient.

Results

The best results are shown in table 3.5. On running the segmenter program over random utterances using the new neural net, it was found that the problem of Stops

Best overall result on test data after convergence

overall % correct = 89.75					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1494	11	40	30	94.9
FRIC	37	685	45	37	85.2
CLOS	159	30	1717	20	89.1
STOP	61	12	15	454	83.8

Experiment 2

Table 3.5

persisted, but the number of spurious segments within long segments of Sonorant or Fricative had been substantially reduced.

3.5. Experiment 3: Adding Spectral Difference

As seen from the previous experiments stops are not being classified well. The reason is that Stops are too short to be represented by steady state parameters in such a wide window. In order to detect Stops and Sonorant onsets better, a difference or derivative like parameter is needed to indicate points of sharp change.

Two new features were added, spectral difference (or change) in the full range of frequencies (0 - 8000Hz) in order to detect Stops as can be seen from table 3.6, and spectral difference in the range 0 - 700Hz in order to more precisely locate Sonorant onset boundaries. The same set of utterances was used. As this feature was intended to detect sharp changes in the signal, averaging in the low-resolution context regions may average out any sharp changes detected in that region. Therefore, the maximum value is chosen as the representative sample.

Results

Histogram of avg. specdiff 0 - 8000Hz

feat 105	SON	FRIC	CLOS	STOP
< 0.0	0	0	0	0
$= 0.0$	0	0	0	0
< 0.1	1541	662	1943	39
< 0.2	214	146	214	43
< 0.3	98	79	116	53
< 0.4	79	41	101	66
< 0.5	47	17	64	71
< 0.6	19	19	44	77
< 0.7	27	18	29	59
< 0.8	14	17	18	61
< 0.9	13	12	10	38
< 1.0	15	8	6	29
$= 1.0$	34	31	17	149
> 1.0	0	0	0	0
> 2.0	0	0	0	0

Table 3.6

The best results on the test set are shown in table 3.7. Stops were found to be better classified, and insertion of spurious segments was substantially reduced. However, the locations of the Sonorant onset and offset boundaries were underestimated. There are two possible ways in which this problem may be fixed, and both introduce a new feature.

Best overall result on test data after convergence

overall % correct = 90.26					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1503	18	31	23	95.4
FRIC	31	697	45	31	86.7
CLOS	145	29	1733	19	90.0
STOP	63	27	10	442	81.5

Experiment 3

Table 3.7

3.6. Experiment 4: Adding Pitch

One approach to the problem found in the results of the previous section was to add pitch. The bounds of consistent pitch defines and in some cases overestimates the limits of the Sonorant in that region. When combined with the peak to peak 0 - 700 Hz and spectral difference 0 - 700 Hz, Sonorant boundaries could become fairly accurate.

A new feature was added, the presence or absence of consistent pitch in a particular frame. This being an on-off feature, it doesn't require any normalization, and sampling every fifth frame in the wide context region is a very good approximation to averaging over every five frames.

Results

The best results on the test set are shown in table 3.8. The boundaries were found to be fixed near the extremities of pitch. However, voiced Fricatives like /*v*/ and /*z*/ were affected by the insertion of Sonorants inside the Fricative. This also affected Sonorant onsets whenever they were preceded by glottal Stops with a substantial burst, as the boundaries were placed so as to include the glottal stops as part of the Sonorant segment rather than separating the two as different segments.

Best overall result on test data after convergence

overall % correct = 90.06					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1474	24	47	30	93.6
FRIC	19	709	40	36	88.2
CLOS	149	25	1725	27	89.6
STOP	55	14	16	457	84.3

Experiment 4

Table 3.8

3.7. Experiment 5: Spectrum instead of Pitch

The alternative approach to the problem described under the results in section 3.5 was to add spectrum to help the spectral difference feature to set the boundaries only where there really was a change. This suppresses the importance of the amplitude and enhances the recognition of frequency bands. This should help distinguish the voiced non-sonorants from the sonorants.

Spectrum at the frame to be classified and in the near context was added after normalizing with respect to some broad region around the frame. Though normalizing the values in each frame with respect to the minimum and maximum values within that frame helps accentuate frequency bands better, it was found to cause problems in weak signal areas and in areas where the signal is absent (silence).

Results

The best results on the test set are shown in table 3.9. Improvements were seen for /*v*/ and /*z*/ as well as glottal Stops, but the Sonorant offsets were often way off the mark. In addition, amplitude dips at nasal boundaries seemed to be getting encoded as spectral changes too, and Closures were being inserted in those places.

Best overall result on test data after convergence

overall % correct = 90.16					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1487	14	46	28	94.4
FRIC	33	697	42	32	86.7
CLOS	145	30	1718	33	89.2
STOP	48	17	9	468	86.3

Experiment 5

Table 3.9

3.8. Experiment 6: A net with all features trained on ISOLET

From the results of the previous experiments we conclude that spectrum as well as pitch were required in order to have a better classification. The pitch helps by broadly demarcating the Sonorant boundaries and improving the accuracy of the Sonorant offset.

In the final experiment, the entire set of features were used to train a neural network using the set of utterances described in section 3.1. The complete set of features and normalization schemes used is described in chapter 2.

3.9. Results

The net was near convergence when it performed at above 95% correct on the training data. Best generalization was achieved after convergence when it had been trained for 160 iterations as seen in figure 3.1. In the figure, the solid line represents the results on test data while the dotted line represents the training data.

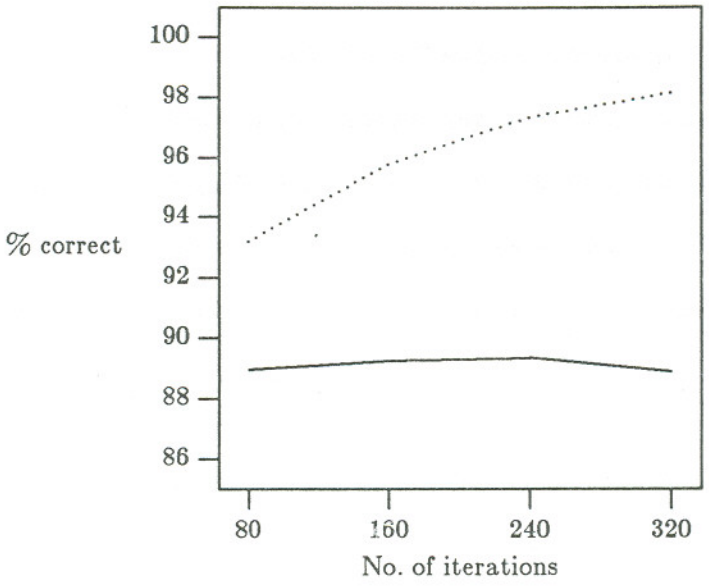


Figure 3.1

Table 3.10 shows the individual confusions and percent correct for each of the labels.

Best overall result on test data after convergence

overall % correct = 89.33					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1482	19	43	31	94.1
FRIC	26	696	49	33	86.6
CLOS	165	29	1710	22	88.8
STOP	65	17	18	442	81.5

Experiment 6

Table 3.10

3.10. Discussion

Let us first analyze the results in this table. Sonorants were rarely confused with others categories and it is easily understood when we look at how well ptp 0 - 700Hz (table 3.2) and pitch discriminate the data. Closure should have a higher classification accuracy, but when we look at the data and what we call Closure, it quickly becomes

apparent where the errors come from. Door slams, and other noise spikes make it look like a Stop, and when extended make it look like a Fricative. Closure gets confused mostly with Sonorants because we have included prevoicing, and low-frequency breath noise and other such periodic signals with Closure whereas they actually have Sonorant like characteristics. Unfortunately there are too few examples of these patterns for us to put them in a separate class. Stops are the most confused kind of segments. Furthermore, they are confused most with Sonorants.

Now let us move on to a subjective evaluation by running the segmenter on random utterances never seen before. We find that most of the errors are due to glottal Stops before Sonorants. These Stops usually have a lot of the characteristics of the following Sonorant, eg. formants. The spectral difference of the energy below 700Hz – the feature used to separate formant onsets from Stop bursts – cannot distinguish the burst from the formant onsets which happen to be part of the burst for these Stops .

Except for the insertion of glottal Stops, most errors could be corrected by simple rules. Therefore a general rule-based post processing stage is introduced to eliminate these simple errors. Only two rules are required, one to remove or merge small spurious Sonorant segments, and another to deal with insertions of other labels inside Sonorants, Fricatives and Closures.

3.11. Evaluation in a system

The segmenter is tested in a system to recognize letters in the */iy/*-set (B, C, D, E, G, P, T, V, Z) taken from the English alphabet. It is contrasted with the best rule-based segmenter built in-house which is an improved version of the segmenter built by Cole and Hou [15] at CMU. The performance figures are given in table 3.11.

In another test in a system to classify all 26 letters of the English alphabet [3] the version with the neural net based segmenter showed a classification accuracy of 95.26%, comparable to the accuracy of the version with the rule-based segmenter which was 95.89%.

Comparison of Rule-based and Neural net Segmenters
Classification of /iy/-set

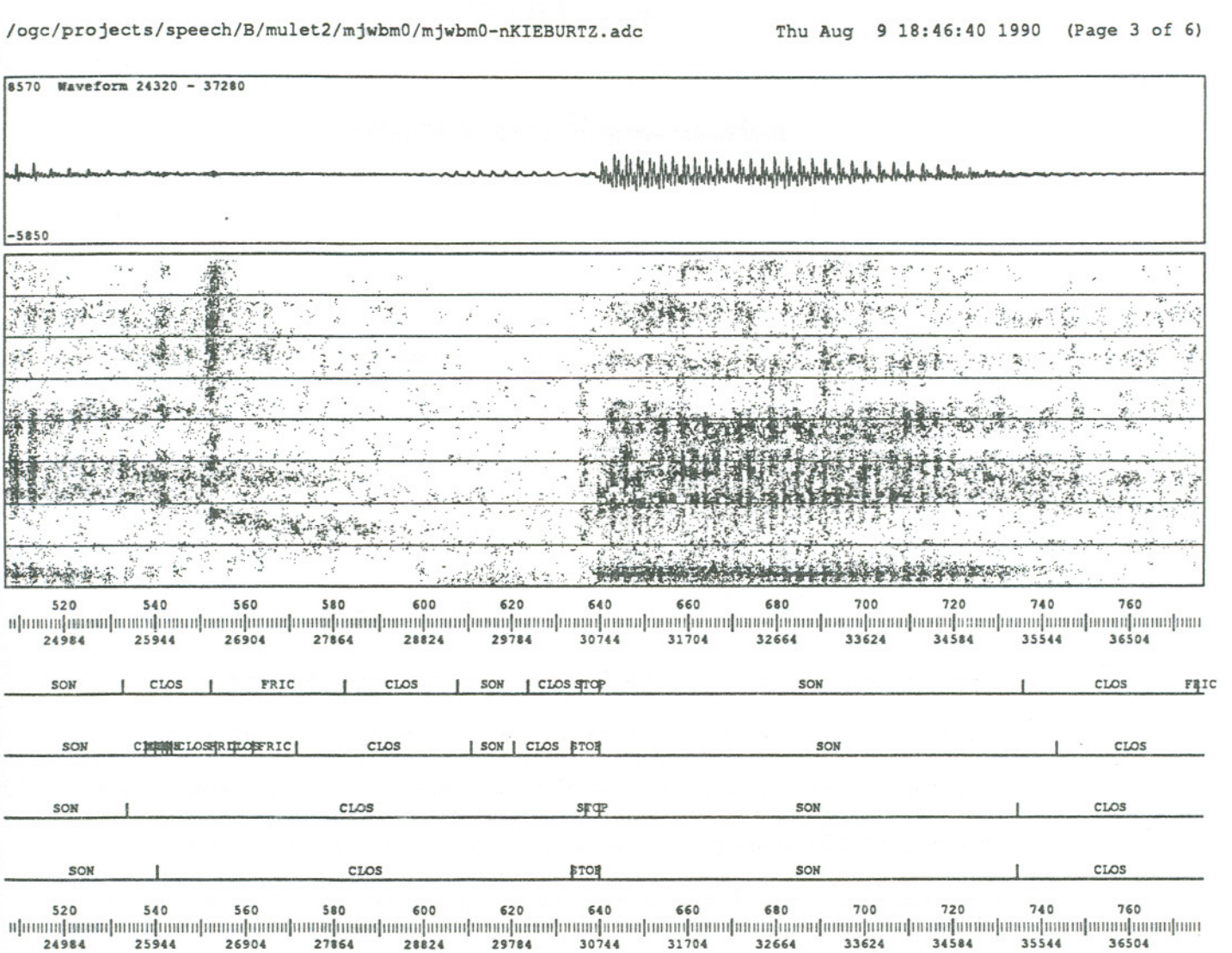
Rule-based segmenter

overall % correct = 93.33										
Letter	B	C	D	E	G	P	T	V	Z	% correct
B	104	.	3	5	.	4	.	4	.	86.7
C	.	117	.	.	.	1	.	.	2	97.5
D	1	.	115	1	1	.	2	.	.	95.8
E	1	.	1	116	.	2	.	.	.	96.7
G	1	.	.	.	117	.	2	.	.	97.5
P	1	.	1	3	2	109	.	3	1	90.8
T	.	1	3	.	2	3	111	.	.	92.5
V	7	.	1	1	.	1	.	107	3	89.2
Z	.	3	.	.	1	.	.	4	112	93.3

Neural network segmenter

overall % correct = 92.49										
Letter	B	C	D	E	G	P	T	V	Z	% correct
B	105	.	5	4	.	2	.	3	1	87.5
C	.	117	.	.	.	.	.	.	3	97.5
D	1	.	112	3	3	1	.	.	.	93.3
E	.	.	1	111	2	3	.	1	2	92.5
G	.	.	.	.	116	1	3	.	.	96.7
P	.	.	3	1	2	105	.	6	2	88.2
T	.	.	5	.	2	4	109	.	.	90.8
V	3	1	2	1	.	.	.	110	3	91.7
Z	1	2	.	.	.	.	.	4	113	94.2

Table:3.11



Effect of training on errors

Figure 4.1 a

CHAPTER 4

Multiple Letter Segmentation

Now that we have a full set of features and a network architecture, we can move on to the problem we sought to solve at the start, ie. building a segmenter for multiple letters. Multiple letters utterances are sequences or strings of spoken letters separated by clear pauses. We shall start out with a description of the method we used and follow that with descriptions of the experiments. The performance of the system will be discussed in the next chapter.

4.1. Training on errors scheme

Since we wish to have the net perform on multiple letter sequences, we need to train it on such utterances. However, there are no new features in these utterances except letters following one another. The assumption that there was only one letter per utterance is invalid now. This affects – actually increases – only the possible left and right contexts that can appear for a given segment class, making the task a little more complex. However, this change could be captured by the scheme with which we represent context. Therefore, there is no necessity to change anything in the training procedure. In fact, the set of vectors used for training the isolated letter segmenter can be reused. Our strategy for this is to start with the network trained on isolated letters and work forward. The success of this scheme can be seen from figure 4.1.

The sets of utterances are chosen as before. However, instead of selecting frames from each segment in the utterance for extracting feature vectors, the best network obtained from the previous stage is run on the training set and those frames not in

agreement with the hand-labeling are selected for generating vectors for retraining. These vectors are then added to the set used to train the isolated letter segmenter. This modified procedure we call *training on errors*. This was done so as to give the network instances it has not learned at the same time saving space consumed by the training data due to want of memory, and time spent in generating data from frames that are labeled correctly. To compensate for the multiple letter utterances in the training set, some vectors are generated from the multiple letter utterances and added to the cross validation test set. The training procedure from here on is the same in all respects as the final one arrived at the end of the previous chapter for isolated letters.

4.2. Database

The training data consists of a database of speakers, called the **mulet** database, recorded in the Speech Laboratory of the Department of Computer Science and Engineering at the Oregon Graduate Institute and created in a way similar to the *isolet* [25] database. It consists of recordings of speakers spelling some names and some random sequences of letters of the English alphabet with pauses between the letters. There are a total of 60 speakers in the database, 30 of them female and 30 male. All the speakers are native speakers of English.

In order to train the neural network it is necessary to define two data sets, one for training and another for testing to check for generalization. Each set of speakers is chosen such that there is equal representation of male and female speakers. A set of utterances is chosen from each speaker.

For this project *mulet1*, a subset of the *mulet* database was used for training the network while *mulet2*, another subset was used to test multiple letter segmentation. For each phase of training, 3 utterances from each of 10 speakers totaling 30 utterances

were chosen. For the testing data, another such 30 utterances were chosen from *mulet2* and added to the test set from *isolet*. A table of the utterances used for the test set is given in table 4.1.

Male speakers		Female speakers	
Spkr.	Utterance	Spkr.	Utterance
mdcd0	CURRIE	fcah0	CAZIER
mdcd0	RANDY	fcah0	ELDON
mdcd0	SIMPSON	fcah0	VAROZ
mdls0	DAVIS	fceb0	DOBER
mdls0	RODGERS	fceb0	MARY
mdls0	STEVEN	fceb0	WALKER
mfes0	HAZELTINE	fdml0	BAUMANN
mfes0	RHONE	fdml0	JOHN
mfes0	STEVE	fdml0	SANDERS
mglt0	COHEN	fgn0	MICHAEL
mglt0	MOORE	fgn0	ROBERTS
mglt0	WENDY	fgn0	SWENSON
mrvs0	GREENSTREET	fkrm0	LOLITA
mrvs0	RON	fkrm0	MCMILLAN
mrvs0	WOLDRICH	fkrm0	TISTADT

Test Set Utterances

Table 4.1

Table 4.2 shows the distributions of labels among speakers, utterances and frames in the test set.

Label	Speaker	Utterance	Frames
SON	10	30	970
FRIC	10	24	294
CLOS	10	30	693
STOP	10	30	472

Table 4.2

The details of the training data sets are given in the respective discussions below.

4.3. Training on errors: Phase I

A table of the utterances used for training on errors is given in table 4.3. The isolated letter segmenter is run on multiple letter utterances listed in table 4.3, and the errors it made are noted by hand and categorized as insertions or deletions of, or substitutions for certain segments, or errors at boundaries, etc. The largest number of errors were found to be Fricative insertions in Closure. The reason for this is that the long noisy Closures found between letters were not seen in isolated letter utterances and the noise is picked up as frication. These errors add up to about 40% of the total errors found. Approximately 30% of the errors were due to Stops and Closures affecting the Sonorant onset while 15% were due to Sonorant offsets extending into the following

Set of speakers and utterances for training

Male speakers		Female speakers	
Spkr.	Utterance	Spkr.	Utterance
mbv0	BERENBERG	fbls0	CURTIS
mbv0	CHARLES	fbls0	GOODMAN
mbv0	MALMSTROM	fbls0	NEVILLE
mgws0	CUSHING	fcmd0	DALE
mgws0	GEORGE	fcmd0	SMITH
mgws0	STRUBLE	fcmd0	TATE
mjgh0	INGRID	fdm0	BRANT
mjgh0	OATES	fdm0	GORDON
mjgh0	WINANS	fdm0	PRICE
mji0	KAGAWA	fdr0	LAURA
mji0	KAWAMURA	fdr0	PAYTON
mji0	LEIALOHA	fdr0	SCOTT
mjjs0	ROGER	fglh0	BRIAN
mjjs0	SANGREY	fglh0	GINA
mjjs0	SCROOGE	fglh0	YORK

Phase I

Table 4.3

Closures. The rest were Stop insertions in Closures near the Sonorant onset boundary and some other kinds of errors.

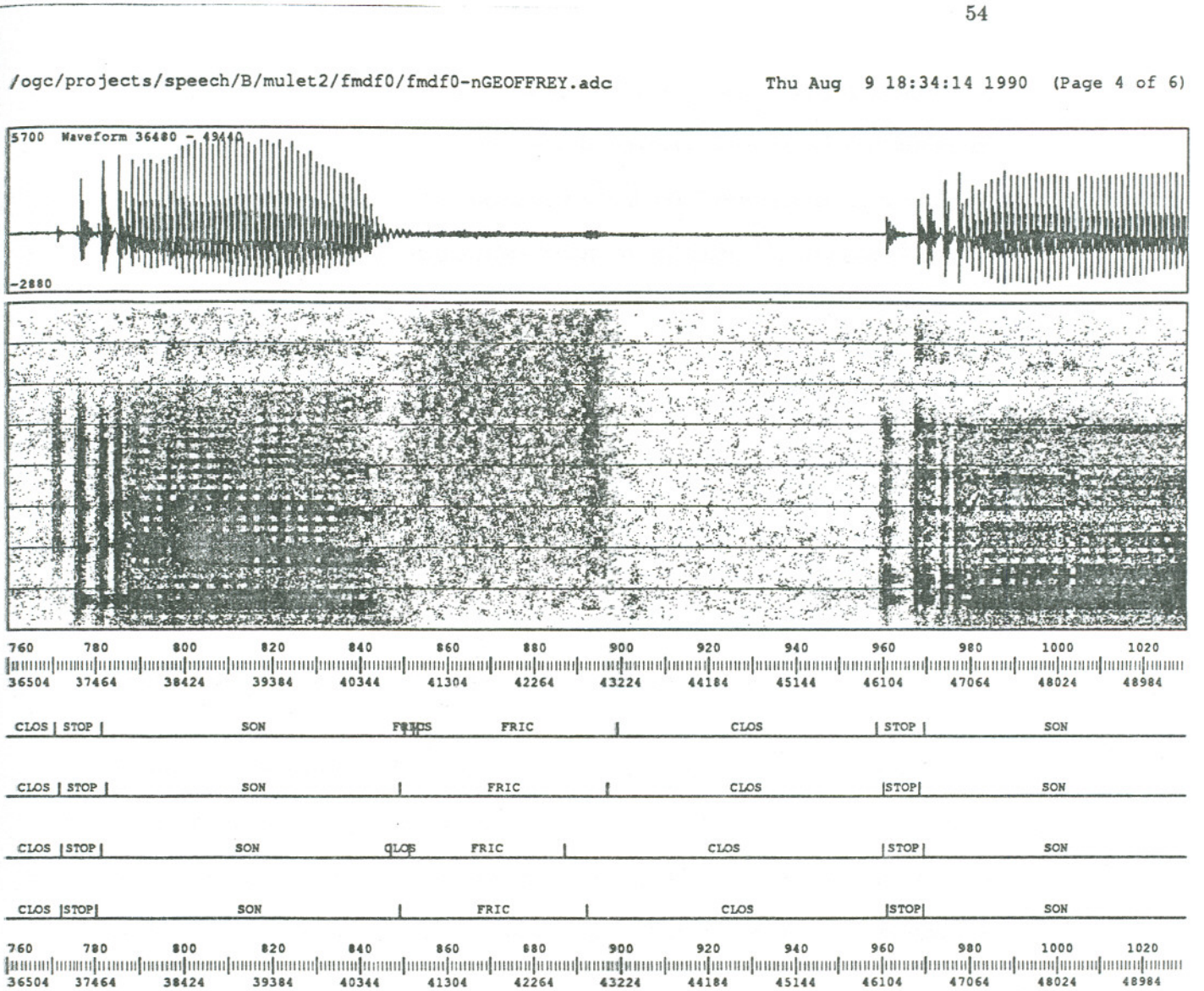
Since 70% of the errors were from the first two cases, it seems best to concentrate on eliminating or at least substantially reducing them. Hence, most of the error vectors for training are chosen from these errors. This results in 434 vectors from wrongly classified frames for a total of 6833 vectors in the final training set.

The network converged after 160 iterations and had an overall percentage correct rate of 89.21% on the test data and 94.79% on the training data. On running the segmenter program on utterances in the training set for the next phase it was found that 25% of the errors were due to Fricatives inserted in Closures and only 15% from Stops or Closures affecting Sonorant onsets, both numbers down from 40% and 30% prior to training on errors. Though this seems encouraging, the errors of Stops inserted in Closures was up at 35% of the total and again up at 25% were Sonorant offsets extending into the following Closures. However, overall the results were encouraging because the number of errors were fewer with the greater percentage of them (about 60%) being those which do not really affect both letter boundary determination and letter recognition. Examples of the improvement in segmentation and the errors may be seen in the second labeling as compared to the first in figure 4.1a. The Closures being more or less

Best overall result on test data after convergence

overall % correct = 89.21					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1471	16	48	40	93.4
FRIC	28	678	63	35	84.3
CLOS	158	22	1725	21	89.6
STOP	51	17	24	450	83.0

Table 4.4



Effect of training on errors (contd.)
Removal of new error: CLOS at SON-FRIC boundary

Figure 4.1 b

in the right place would cause little problems for letter boundary determination, and the more or less accurate Sonorant onset boundaries, fewer misclassifications of Sonorants and fewer errors in the Consonants would help letter classification to a fair extent. However, 40% of the errors being serious, is still high. The number of correct classifications of Sonorants is not sufficient and so also the accuracy of determination of Sonorant onset boundaries. These two errors need to be especially low for good letter classification percentages. After all, the segmenter is but one step in a larger system. So further training on errors is necessary for greater improvement in the system.

4.4. Training on errors: Phase II

A table of the utterances used for training on errors is given in table 4.5.

While it seems best to concentrate on the one or two errors making up the largest percentage of the total, it is prudent not to ignore the kinds of errors found in the previous stage when selecting error vectors for retraining. Hence, for this stage there are 697 vectors from misclassified frames for a total of 7530 vectors in the final training set.

The network converged after 240 iterations and had an overall percentage correct rate of 89.31% on the test data and 93.25% on the training data. On running the segmenter program on utterances in the training set for the next stage it was found that the total number of errors was drastically reduced, even though only a small increase in the overall percentage correct rate was observed. The reason was that qualitatively, there was vast improvement in segmentation while the number of frames correctly classified increased by a very small amount. This difference may be observed

Set of speakers and utterances for training

Male speakers		Female speakers	
Spkr.	Utterance	Spkr.	Utterance
mjw0	BIGGAR	fjbc0	AIKIN
mjw0	KIRTI	fjbc0	BAYARD
mjw0	SHAH	fjbc0	JOHN
mmr0	CAROL	fjmr0	BYRON
mmr0	COLE	fjmr0	DAFOE
mmr0	WEATHERILL	fjmr0	HOWARD
mnjh0	CONSTANTINO	fjms0	BRANDY
mnjh0	HARRIS	fjms0	LUTHER
mnjh0	SERENITY	fjms0	WELBORN
mpdn0	ARMSTRONG	fkma0	BILL
mpdn0	CRAIG	fkma0	COWAN
mpdn0	JUSUS	fkma0	MATHER
mrac0	BISCHEL	flkm0	FASSNIDGE
mrac0	PEGGY	flkm0	JOANN
mrac0	WEINSTEIN	flkm0	RIES

Phase II

Table 4.5

Best overall result on test data after convergence

overall % correct = 89.31					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1474	15	55	31	93.6
FRIC	19	677	76	32	84.2
CLOS	164	17	1725	20	89.6
STOP	54	13	22	453	83.6

Table 4.6

in figure 4.1a when comparing the second and third labelings in that figure. Most of these qualitatively determined errors were Stops inserted in Closures, and Sonorant offsets extending into the following Closures. The other errors were very few except for one. Small Closures were inserted between the Fricative and Sonorant in letters like *C, P, K, F* and *S* as shown in the second and third labelings in figure 4.1b. It looked like training on some of the previous errors had resulted in over correction. It is therefore necessary to train further on these errors.

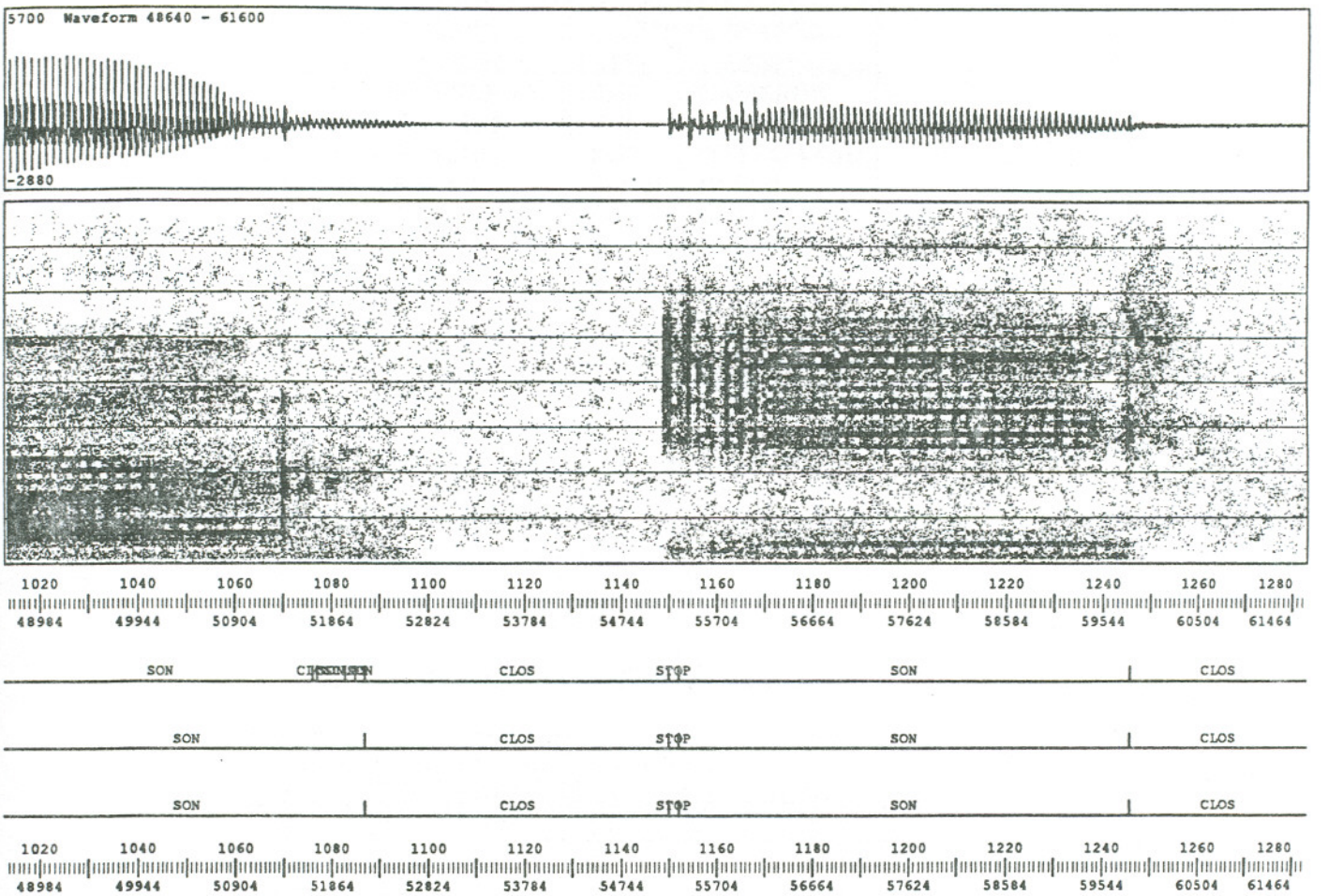
4.5. Training on errors: Phase III

A table of the utterances used for training on errors is given in table 4.7.

Though emphasis is necessary on the new errors found in the previous phase, it is important not to underplay other errors and to remove as many errors as possible. Furthermore, as the number of errors are far fewer, it is possible to sample more errors and do so more frequently. This results in 356 vectors from wrongly classified frames for a total of 7886 vectors in the final training set.

The network converged after 240 iterations and had an overall percentage correct rate of 89.33% on the test data and 92.00% on the training data. On running the segmenter program on some randomly chosen test utterances it was found that the errors

/ogc/projects/speech/B/mulet2/fmdf0/fmdf0-nGEOFFREY.adc Sun Aug 12 19:28:36 1990 (Page 5 of 6)



Effect of post processing rules
Rule to collapse short alternations of Sonorant
and Closure into one Sonorant

Figure 4.2 a

Set of speakers and utterances for training

Male speakers		Female speakers	
Spkr.	Utterance	Spkr.	Utterance
mras0	KOPELMAN	flvl0	CAMERON
mras0	RAYMOND	flvl0	OSKIERKO
mras0	STAEHLI	flvl0	OVERGAARD
mrmh0	KELLY	fmbd0	DINAH
mrmh0	MYRTLE	fmbd0	KELLER
mrmh0	RALF	fmbd0	WOODWARD
mtgd0	BENNETT	fmlv0	KOVARIK
mtgd0	JAMES	fmlv0	NELSON
mtgd0	RIVIN	fmlv0	VLASTA
mtkl0	BURTON	fske0	ERICHSEN
mtkl0	FROST	fske0	SMITH
mtkl0	RUSSELL	fske0	TIANA
mtlr0	LINUS	ftlj0	DICKSON
mtlr0	MODLICH	ftlj0	GINA
mtlr0	WOLFINGER	ftlj0	HIPA

Phase III

Table 4.7

Best overall result on test data after convergence

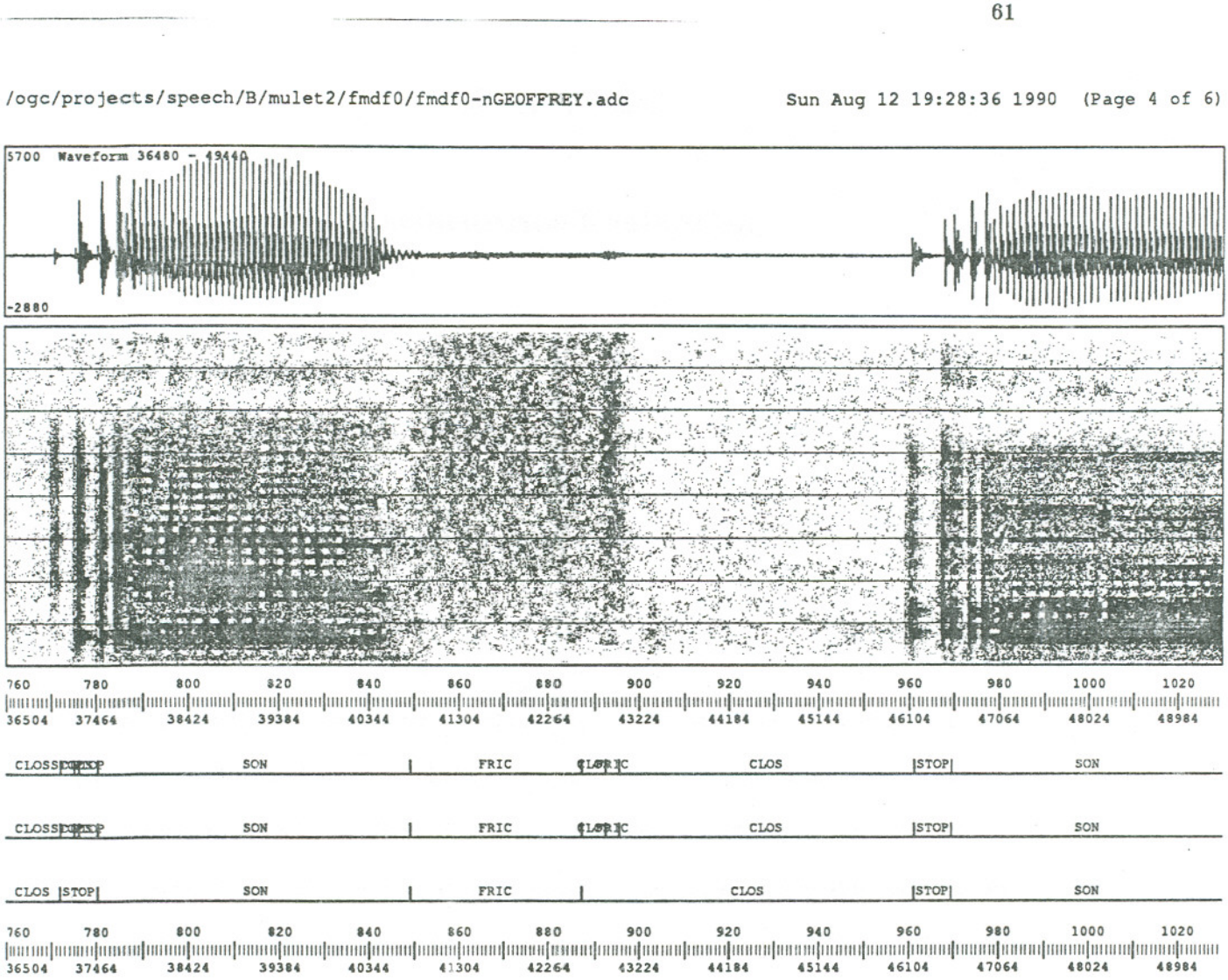
overall % correct = 89.33					
Label	SON	FRIC	CLOS	STOP	% correct
SON	1457	14	77	27	92.5
FRIC	24	678	70	32	84.3
CLOS	127	19	1756	24	91.2
STOP	72	11	20	439	81.0

Table 4.8

introduced in the previous stage of training were removed and no new errors were added as seen from the third and fourth labelings in figure 4.1b. The errors were few and not serious enough to warrant more training. Most errors were Stops inserted in Closures before the Sonorant onsets, and Sonorant offsets extending into the following Closures. It is possible to remove the few wrinkles left using some simple rules as done in the previous section, by incorporating knowledge of the possible segment class sequences.

4.6. Rule-based post processing

This being a neural network based segmenter and not a rule- based one, only a few simple rules which take advantage of the structure of broad class segments in spoken letters are used to augment the overall classification. The small Sonorant segments in the Closure near the Sonorant offset boundary are cleaned up by relabeling them as Closure and then merging them into one segment so as to get a more accurate Sonorant offset boundary. Next, small Closures and Fricatives inserted in various places are also removed. The range of durations of these classes of Segments are determined by constructing histograms from the hand labeled data used for training. The rules seem to work fairly well as seen in Figure 4.2. A more objective evaluation of the performance of the segmenter is found in the next chapter.



Effect of post processing rules (contd.)
Rule to remove short segments of certain types

Figure 4.2 b

CHAPTER 5

Performance Evaluation

5.1. How to measure performance?

In the previous chapter we have seen some of the kinds of errors made by the segmenter. Examination of visual displays revealed that the errors were not serious considering the application of recognition of spoken English letters.

Thus far, we have measured performance in two ways; examination of confusion matrices, and subjective evaluation of visual displays by experts. Subjective evaluation by experts is a fairly useful method but, as with all subjective judgements, it is limited by biases. This method has been used very effectively to build the segmenter and bring it to its "final" form. It has been discussed in earlier chapters and shall not be repeated here. Suffice it to mention that it is necessary to introduce some objectivity into measuring performance. Hence, two other methods have been used: performance as compared to independent human labelers, and performance within a more complete system.

5.2. Comparison with experts

We turn to another method of performance measurement, comparison to expert labelers. Humans are the best speech recognizers, and they provide the best measuring standards. A previous study has shown that expert spectrogram readers can locate up to 97% of phonetic segments in spectograms of unknown utterances. a set of random utterances from different speakers is chosen. For this thesis, two experts were used. They independently labeled a set of randomly selected utterances and the agreement

between them was determined. The segmenter is then run on these utterances and the output saved. One expert must be used as the standard and his labeling assumed to be correct. The other expert's labeling was then compared with that of the first expert, and so was the output of the neural network segmenter. The errors were determined and tabulated for both comparisons. The number and kinds of errors made by the neural network segmenter was compared with the disagreement in labeling between the second labeler and the first. If the corresponding numbers are more or less the same, then it can be argued that the neural network segmenter disagrees with experts to approximately the same extent as the disagreement among the experts.

Errors can be classified as *insertions*, *deletions* or *substitutions*. In addition, we scored *inaccurate boundaries*. The determination of the first three is done by counting the numbers of such errors, which is used as the performance – the fewer the better. The boundary errors are rather difficult to interpret. As mentioned before, while Sonorant onsets need to be fairly accurate, it is not the case with Sonorant offsets. This can be seen by the extent of disagreement among human labelers as regards these boundaries. Further, Stops being fairly short, need to be accurately detected. Therefore, it was decided to present the accuracy as a histogram of the number of 3 msec frames by which the boundaries are in disagreement.

Two tables are presented, one with the first human labeler as the standard and the other with the second labeler as the standard. The errors in the segmenter can thus be compared with the agreement between the labelers giving us an indirect performance measure for the segmenter.

5.2.1. Database

The utterances used are from the *mulet2* database, and have not been used or seen in earlier experiments. Five speakers are chosen with three utterances from each. Table 5.1 lists the utterances used for performance testing.

Male speakers		Female speakers	
Spkr.	Utterance	Spkr.	Utterance
mdem0	DEWITT	fcch0	EPSTEIN
mdem0	DIDIER	fcch0	RUBY
mdem0	MACKE	fcch0	SILLS
mmwp0	ETOL	fkw0	BROWNLOW
mmwp0	LILLARD	fkw0	BUKOJEMSKY
mmwp0	TUEY	fkw0	MARLIN
		fmer0	DAVIS
		fmer0	DEBRA
		fmer0	WALSH

Table 5.1

5.2.2. Insertions, Deletions and Substitutions

As mentioned earlier, for insertions, deletions and substitutions only the the count of errors need to be compared. The number of these kinds of errors is presented for each of the four labels. Apart from how good or bad the segmenter is overall, this tells us which kinds of labels are more difficult to classify. These numbers are shown in tables 5.2a and 5.2b for *evaluator 1* and *evaluator 2* respectively as the standards.

5.2.3. Boundary misalignment errors

The histograms for the boundary errors are grouped by segment class (or label). This gives us an idea of which segment boundaries deserve further work, and which to leave because the disagreement amongst the human labelers is comparable to the error rates. These histograms are cumulative histograms, and are shown in Figures 5.1 and 5.2 for labelers 1 and 2 respectively as standards.

Label names	evaluator B		NN Segmenter	
	insertions	deletions	insertions	deletions
SON	0	0	3	0
FRIC	0	0	0	0
CLOS	2	2	3	3
STOP	6	15	11	5

Reference evaluator A

Table 5.2a

Label names	evaluator A		NN Segmenter	
	insertions	deletions	insertions	deletions
SON	0	0	3	0
FRIC	0	0	0	0
CLOS	2	2	2	2
STOP	15	6	16	1

Reference evaluator B

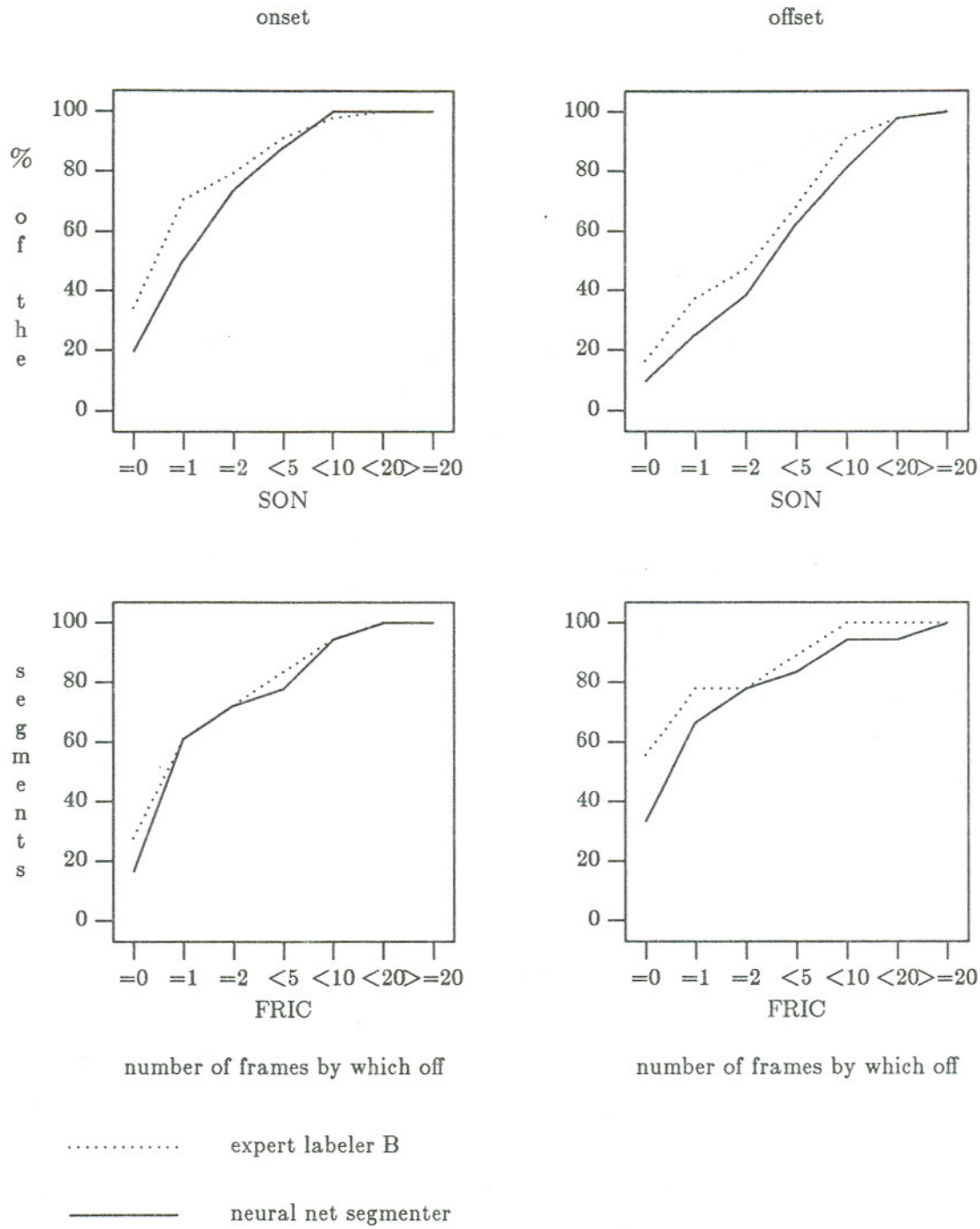
Table 5.2b

5.3. Performance as a sub-system

It has been mentioned a number of times when discussing the errors that the performance analysis of the segmenter is to a large extent determined by the application in which it will be most used. The system in which this segmenter is used is a directory assistance system which has as part of it a letter recognition module. The speakers were required to pause between letters when spelling the name. One important performance measure is the reliability of the Closures between letters so that their approximate boundaries may be correctly determined. Another equally important measure is the reliability of those segment boundaries within the letter which are critical in the generation of features for letter identification. In the in-house directory assistance system the boundaries are the Sonorant onset and the Stop onset and offset. These two need to be determined with maximum accuracy. The Fricative onset and offset boun-

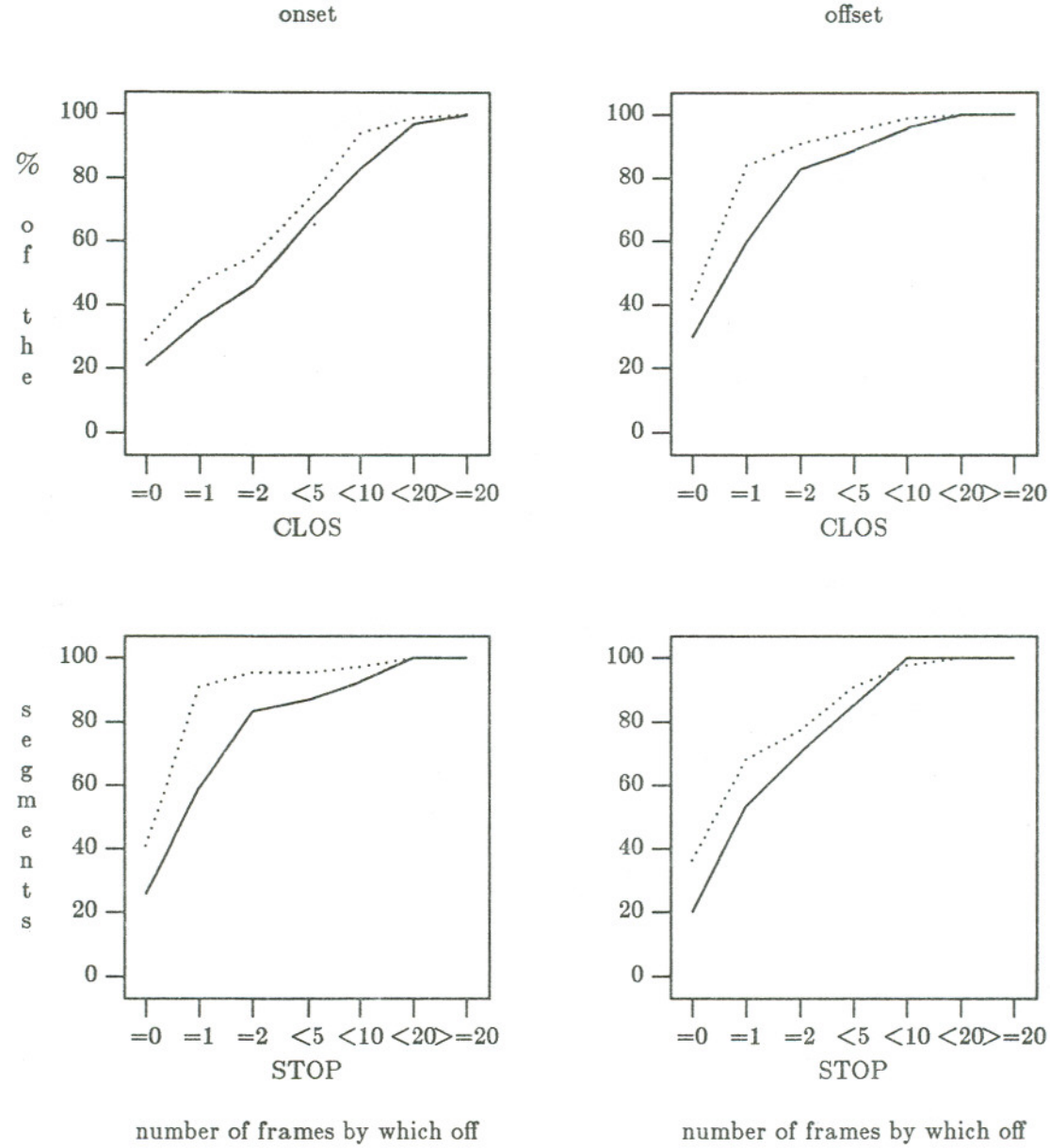
daries may be allowed a little more room for error. The Sonorant offset before a Closure is allowed maximum lattitude with a single rigid constraint that it contain the Sonorant.

The performance of the segmenter in the system is compared to that of a system with the best rule-based segmenter developed in-house. The system was tested on 422 utterances by speakers spelling names or random sequences of letters not used for training the system. The output of the segmenter was used to segment the utterance into letters for further classification. The number of errors in segmenting into letters with the output of the neural network based segmenter was 16 (3.8%) as compared with 48 (11.4%) with the output of the rule-based segmenter.



Reference labeler A

Figure 5.1 a



Reference labeler A

Figure 5.1 b

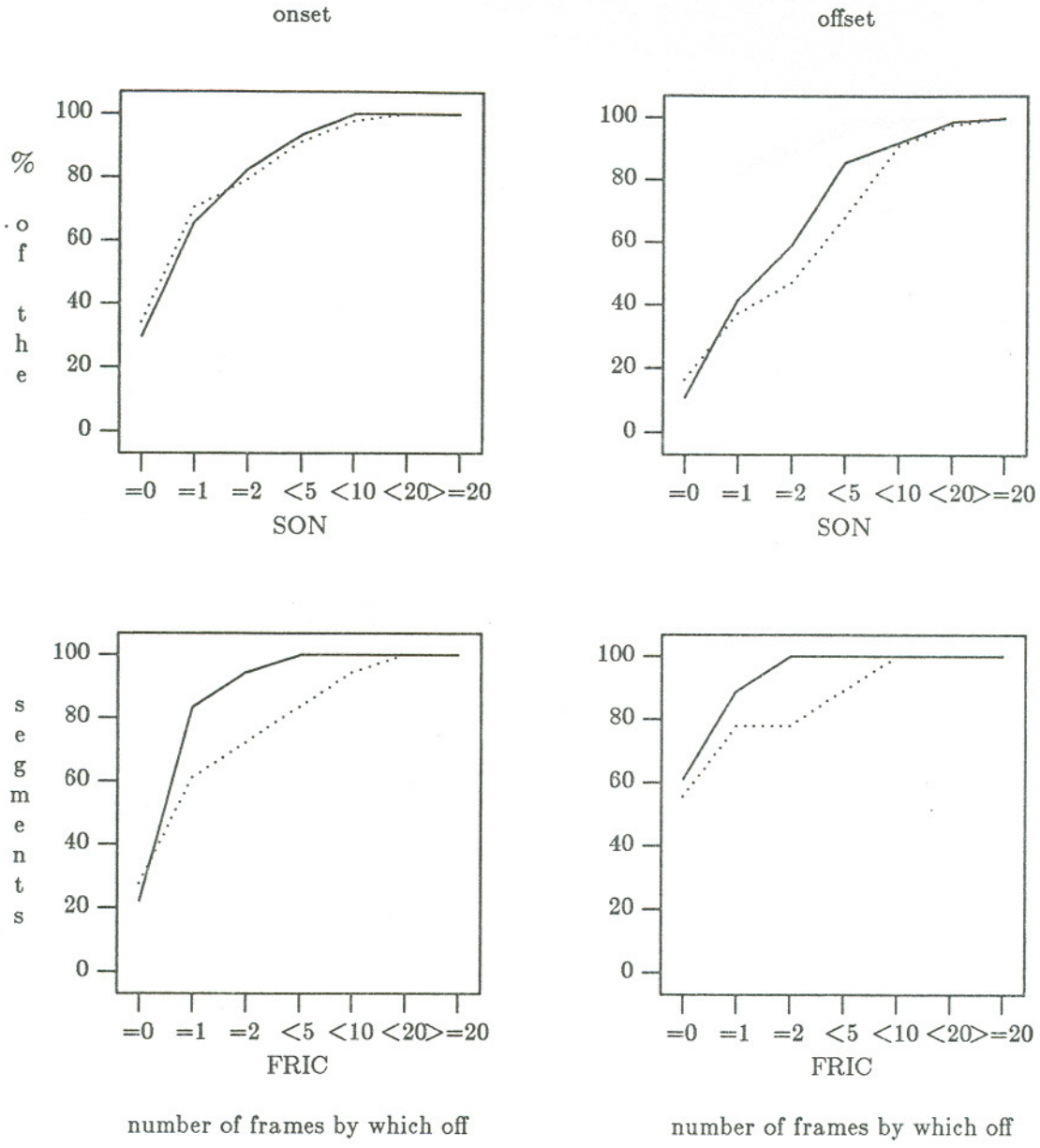
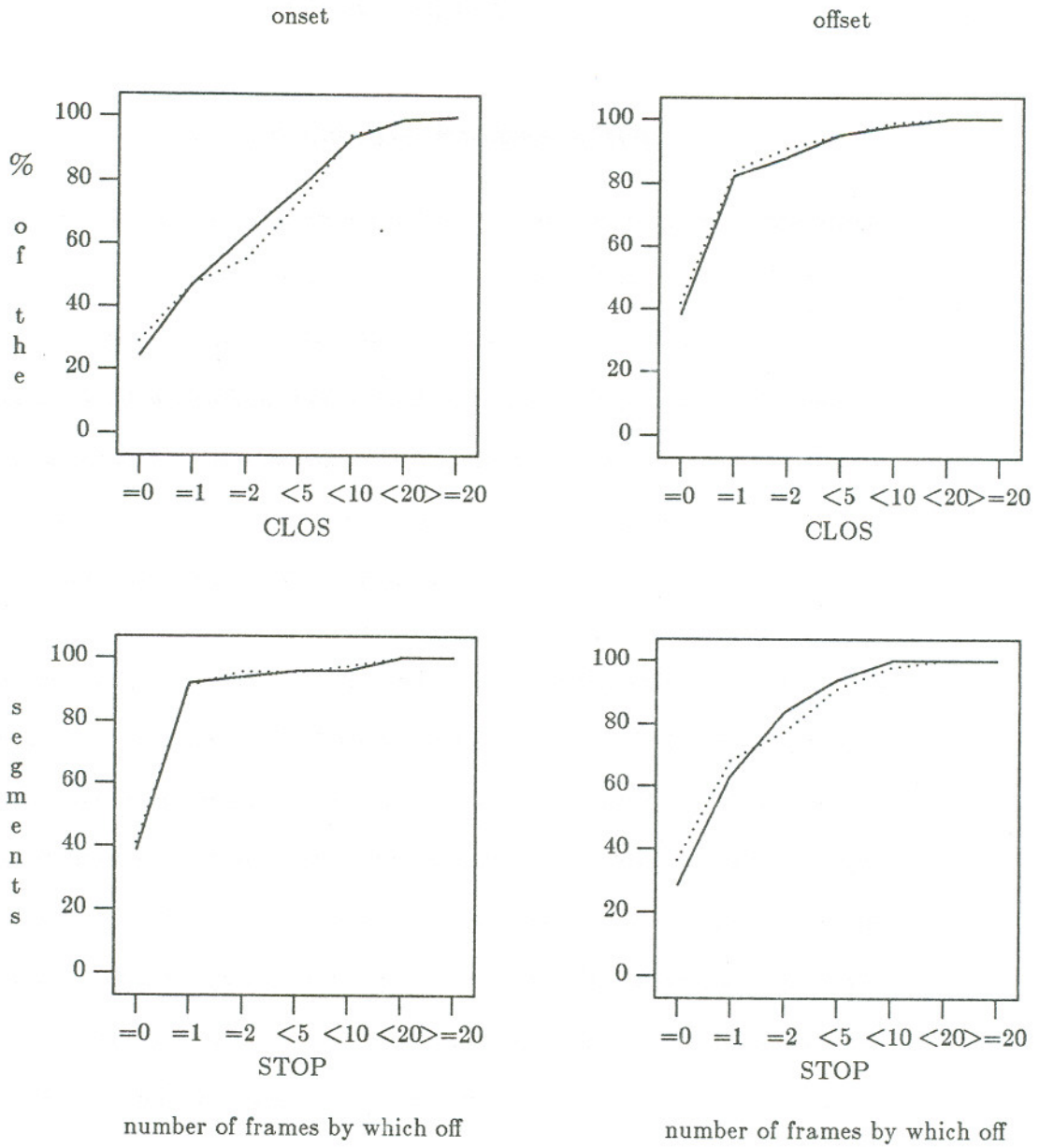

onset
offset
% of the segments
100
80
60
40
20
0
=0 =1 =2 <5 <10 <20 >=20
SON
FRIC
number of frames by which off
expert labeler A
neural net segmenter
Reference labeler B

Figure 5.2 a



Reference labeler B

Figure 5.2 b

CHAPTER 6

Conclusions and Future work

Neural network segmentation systems are a viable alternative to speech segmentation and classification into broad phonetic categories. Neural network hardware will greatly enhance the speed of such systems. The frame-by-frame classification method, keeping only a window of information around, lends itself nicely to an on-line segmentation system, making it more like the way humans process speech. Now that digital signal processing (DSP) and neural network hardware with these capabilities are being developed, such systems can actually be realized.

Knowledge does help segmentation and classification to a fair extent. We have seen how peak-to-peak amplitude in the frequency range 0 - 8 kHz effectively differentiates Closure from the other categories, while pitch and peak-to-peak amplitude low-pass filtered to 700 Hz identify Sonorants. Zero-crossing count is useful for locating Fricatives while spectral difference is necessary for Stops. Finally, we also found that the spectrum itself cannot be ignored as it contains valuable information about the distribution of the energy in frequency. Though it doesn't perform as well on its own [18] , spectrum helps fine tune the classification and improve performance. Despite the goodness of the features one chooses, one cannot avoid the use of large amounts of data. We do manage to improve the efficiency of training by training on errors as described earlier. But, we cannot overlook the fact that more data does improve generalization. Current techniques to train such large networks require fast processors and large amounts of memory. We expect this problem to all but disappear once neural network hardware is readily available at reasonable cost, but that is still a few years away.

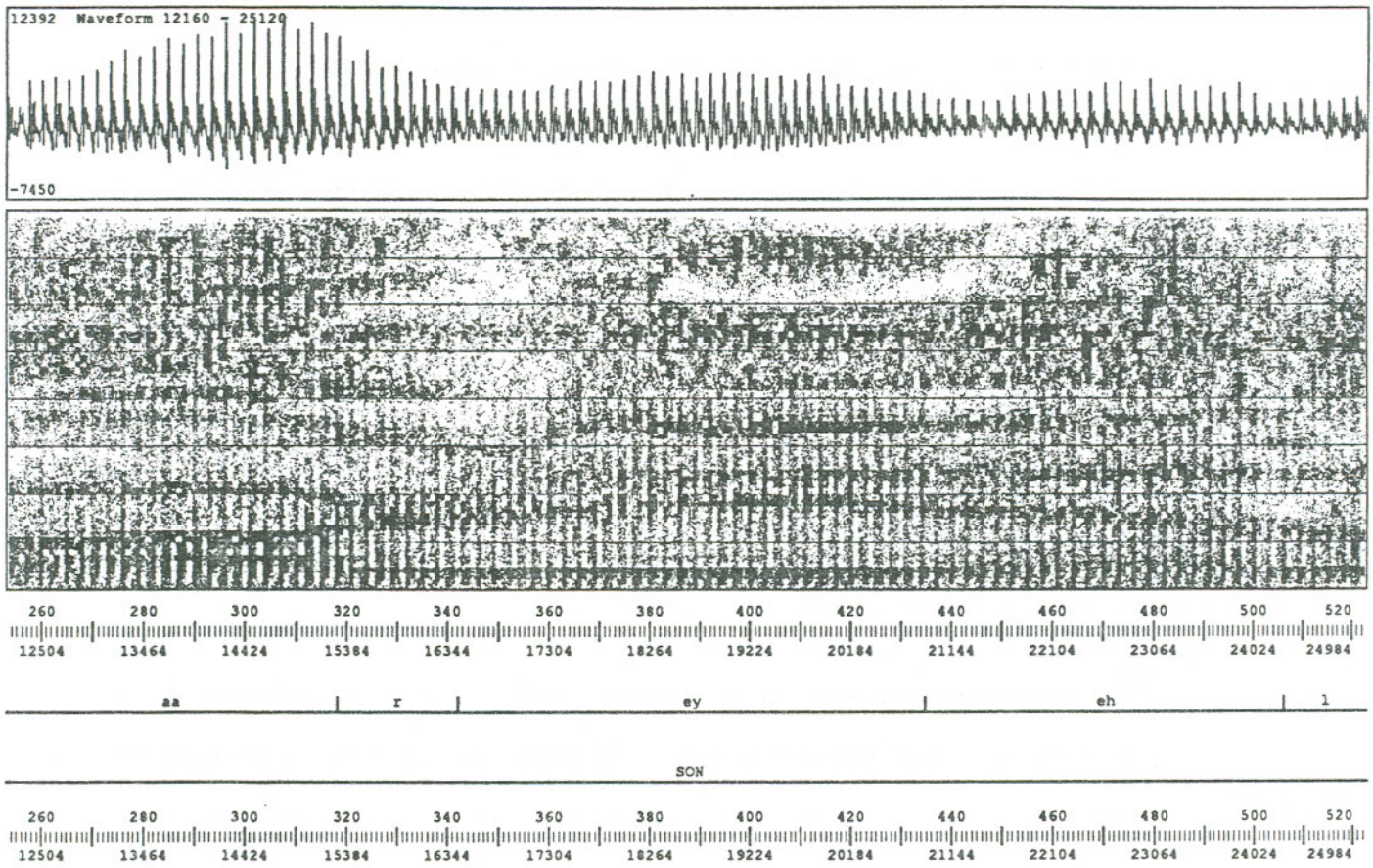
Until then, we are "caught between a rock and a hard place" so to speak. We need more data, but we also need more resources and must allow for long training times. We could also look for more efficient ways of training a back propagation neural network.

Frame-by-frame classification followed by median smoothing is fine but we need a more advanced technique to clean-up and improve the segmentation. Rules will work only in a limited sense. Once we move on to more complex systems where the segmentation is refined, rules will start breaking down. A time-dependent probabilistic network could be a good start.

As seen earlier, the segmenter works fairly well even when taken from the problem of spoken isolated letters to spelled words. The ease with which the system could be extended has increased our confidence in the approach.

The success of a word recognition system where the word is spelled shows that segmentation is helpful in identifying word boundaries. That the neural network segmenter performs better than the best rule-based segmenter developed in-house further vindicates our approach.

A project to extend the system to connected letters, ie. spelled names or words without pauses, is on going. One problem that is seen in this kind of speech is that Sonorants of adjacent letters, sometimes from three or four consecutive letters, are not separated by any other segment, as seen in figure 6.1. The question is how to find the boundaries between the Sonorants of consecutive letters. Sometimes, when making the transition from one letter to the next, glottalization is introduced (figure 6.2), and if detected the inter Sonorant boundaries could be detected. But what about the transitions at which



Section of speech showing
no segment separating successive SONs

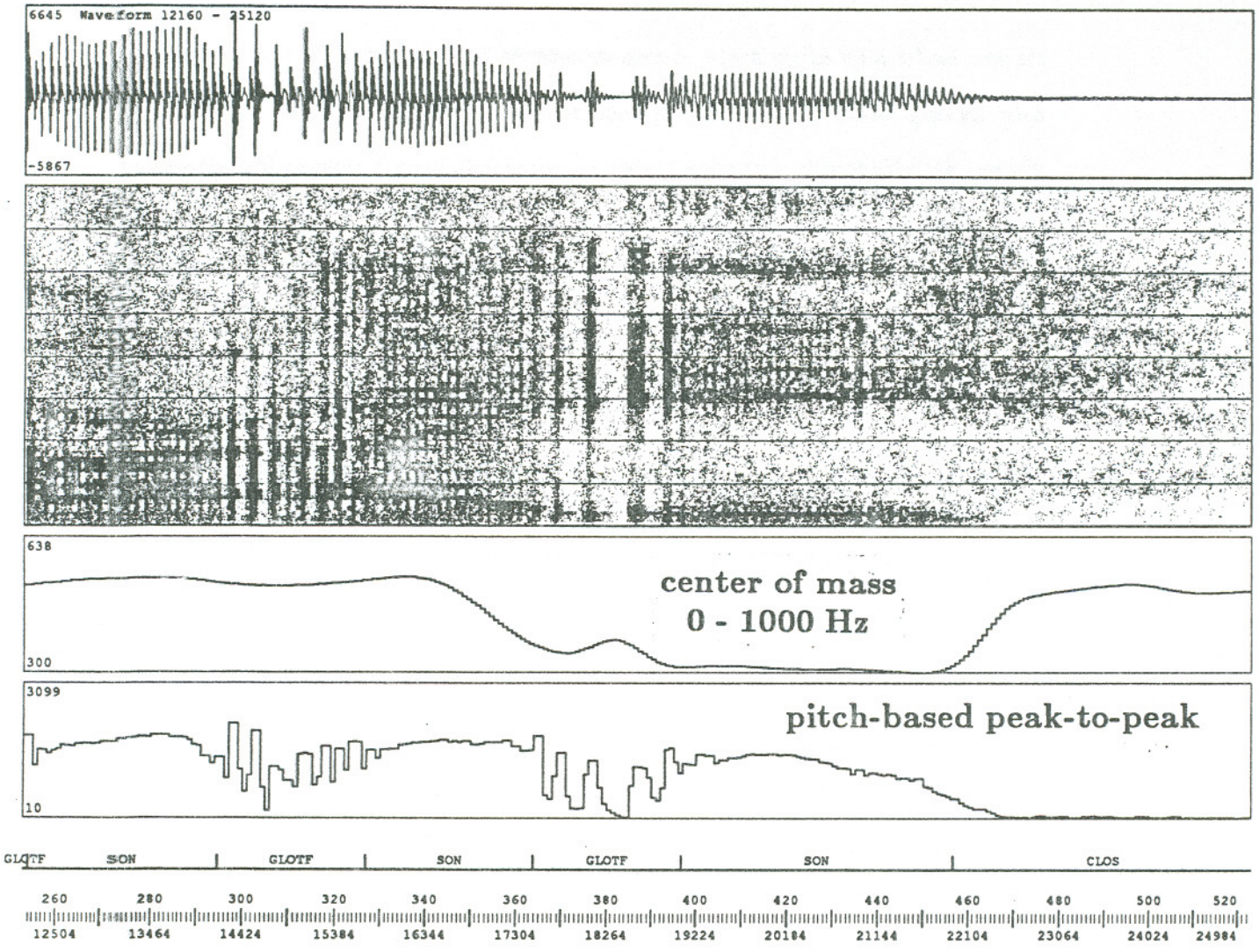
Figure 6.1

glottalization is not perceived?

There are two possible approaches. The first is to build a larger neural network with probably more features to recognize finer phonetic categories than Sonorant or Fricative. The same methods as used for the current segmenter can be applied to search for features to perform this finer classification. A problem that such a system could run into is that some differences like those among Sonorants and among Fricatives are much smaller than others like those between Fricatives and Sonorants. The larger differences – across broad categories – could at some point squash the smaller ones – within broad categories – and make them appear insignificant. This could cause confusion in trying to distinguish among Sonorants, for example, brought about because of competition. An alternative approach is to refine the segments determined from the broad classification system. That means smaller networks performing the subclassification need to be built, each with its own set of special features to best bring out the differences amongst the classes in its set. This seems to be a good solution, but it has a weakness in that if the segment given to a specialized network has been misclassified at the previous stage, the error cannot be corrected, and the system could go hay wire.

In either of these approaches, the possibility of confusions is such that it is necessary to hypothesize alternate segmentations and classifications and connect them into a network with probabilities. We then need to find the best path through this network, for which we shall have to go back to basic speech research results and use knowledge to constrain the search instead of multiplying probabilities together.

Further research must then be done to find ways to determine word boundaries, and to use the constraints in the language to restrict the choice of letters to look for at every



Section of speech showing glottalization separating SONs

Figure 6.2

stage.

The next step is of course, natural continuous speech, which opens up a whole new set of problems. We, however, are confident that neural network based systems with knowledge will provide a good alternative to todays systems and probably the systems of the near future.

References

1. R. Jakobson, G. Fant, and M. Halle, "," in *Preliminary to Speech Analysis,* **MIT** Press, Cambridge, MA (1963).

2. L. Lisker, "Voicing" in English: A catalogue of acoustic features signalling /b/ versus /d/ in trochees," pp. 3-11 in *Language and Speech,* (1986).

3. R. A. Cole and M. A. Fanty, "Spoken Letter Recognition," *Proceedings of the DARPA Speech and Natural Language Workshop,* (June 1990).

4. B. T. Lowerre, "The Harpy Speech Recognition System," PhD Thesis, Dept. of Computer Science, Carnegie-Mellon University (April 1976).

5. V. R. Lesser, R. D. Fennell, L. D. Erman, and D. R. Reddy, "Organization of the Hearsay II Speech Understanding System," *IEEE Trans. Acoustics, Speech, and Signal Processing* **ASSP-23**(1) pp. 11-24. (Feb 1975).

6. J. J. Wolf and W. A. Woods, "The HWIM Speech Understanding System," pp. 316-339 in *Trends in Speech Recognition,* ed. Wayne A. Lee,Prentice-Hall, New Jersey (1980).

7. R. A. Cole, R. M. Stern, M. S. Phillips, S. M. Brill, A. Pilant, and P. Specker, "Feature-based speaker-independent recognition of isolated English Letters," *Proceedings of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing,* pp. 731-734 (1983).

8. H. Kasuya and H. Wakita, "An approach to segmenting speech into Vowel and Non-Vowel like intervals," *IEEE Trans. Acoustics, Speech, and Signal Processing* **ASSP-27**(4) pp. 319-327 (Aug 1979).

9. C. K. Un and H. H. Lee, "Voiced / Unvoiced / Silence discrimination of speech by Delta Modulation," *IEEE Trans. Acoustics, Speech, and Signal Processing*

ASSP-28(4) pp. 398-407 (Aug 1980).

10. S. G. Knorr, "Reliable Voiced / Unvoiced decision," *IEEE Trans. Acoustics, Speech, and Signal Processing* **ASSP-27**(3) pp. 263-267 (June 1979).

11. L. J. Siegel and A. C. Bessey, "Voiced / Unvoiced / Mixed excitation classification of speech," *IEEE Trans. Acoustics, Speech, and Signal Processing* **ASSP-30**(3) pp. 451-460 (June 1982).

12. L. Wilcox and B. T. Lowerre, "Coarse classification using a hierarchial decision tree and top down parsing," *Proceedings of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pp. 73-76 (April 1986).

13. P. Regel, "A module for acoustic-phonetic transcription of fluently spoken German," *IEEE Trans. Acoustics, Speech, and Signal Processing* **ASSP-30**(3) pp. 440-450 (June 1982).

14. R. A. Cole, A. I. Rudnicky, V. W. Zue, and D. R. Reddy, "Speech as patterns on paper," in *Perception and Production of Fluent Speech*, ed. Ronald A. Cole,Lawrence Erlbaum Assoc., Hillsdale, NJ (1980).

15. R. A. Cole and L. Hou, "Segmentation and broad classification of continuous speech," *Proceedings of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (April 1988).

16. J. R. Glass and V. W. Zue, "Acoustic Segmentation and Classification," *Proceedings of the DARPA Speech Recognition Workshop*, pp. 38-43 (March 1987).

17. H. C. Leung, "The use of Artificial Neural Networks for Phonetic Recognition," PhD Thesis, Dept. of Electrical Engineering & Computer Science, Massachusettes Institute of Technology (May 1989).

18. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme Recognition using Time-Delay Neural Networks," *IEEE Trans. Acoustics, Speech, and Signal Processing* **ASSP-37**(?) pp. 328-339. (March 1989).

19. A. Aktas, O. Schmidbauer, O., Maier, K.H., Fiex, W.H., K. H. Maier, and W. H. Fiex, "Classification of coarse phonetic categories in continuous speech: statistical classifiers vs. temporal flow connectionist network," *Proceedings of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing,* pp. 89-92 (1990).

20. R. L. Watrous, "Learning algorithms for connectionist networks: Applied gradient methods for nonlinear optimization," *Proceedings of the 1st Intl. Conf. on Neural Networks,* pp. 619-627 (1987).

21. E. Barnard, R. A. Cole, M. Vea, and F. Alleva, "Pitch detection with a Neural Net classifier," *IEEE Trans. Acoustics, Speech, and Signal Processing,* (Feb 1991 (to appear)).

22. A. I. Rudnicky and R. A. Cole, "Effect of subsequent context on syllable perception," *J. of Experimental Psychology: Human Perception and Performance* **4**(4) pp. 638-647 (1978).

23. D. E. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature* **323** pp. 533-536 (1986).

24. E. Barnard, "Optimization for training neural nets," *IEEE Trans. Pattern and Machine Intelligence,* (March 1989).

25. R. A. Cole, Y. Muthusamy, and M. A. Fanty, "The ISOLET Spoken Letter Database," Technical Report 90-004, Computer Science Department, Oregon Graduate Institute (1990).

BIOGRAPHICAL NOTE

The author was born on 13 November 1963, in Coimbatore, India. He received his Bachelor of Science degree from the University of Madras in June 1984.

After completing undergraduate studies in Computer Science and Engineering at the Indian Institute of Science, Bangalore in 1987, and obtaining a Bachelor of Engineering degree, he worked for a year as a Software Engineer at Wipro Information Technology Ltd., Bangalore, India. He then joined the Oregon Graduate Institute of Science and Technology for graduate studies, where he completed the requirements for the degree Master of Science in August 1990.