1991

Computer classification of stop consonants in a speaker independent continuous speech environment

Michael R. Campanelli

Rochester Institute of Technology
School of Computer Science and Technology

Computer classification of stop consonants in a speaker independent continuous speech environment

by

Michael R Campanelli

A thesis submitted to
The Faculty of the School of Computer Science and Technology.

Approved by:

Al Biles

James Hillenbrand

Peter Andrews

Computer Classification of Stop Consonants In a Speaker Independent Continuous Speech Environment

By

Michael R Campanelli

Date January 4, 1991

Acknowledgments

The author would like to thank Prof. James Hillenbrand and Prof. Al Biles for their linquistic and editorial comments. Thanks also to Rob Gayvert for unending assistance with the database and signal analysis work. And many thanks to my wife for her continued support and encouragement.

This work was supported by Rome Air Development Center and the Air Force Office of Scientific Research as part of the Northeast Artificial Intelligence Consortium (contract F3060285-C-0008) and by Xerox Webster Research Center Redirection program.

Abstract

In the English language there are six stop consonants, /b,d,g,p,t,k/. They account for over 17% of all phonemic occurrences. In continuous speech, phonetic recognition of stop consonants requires the ability to explicitly characterize the acoustic signal. Prior work has shown that high classification accuracy of discrete syllables and words can be achieved by characterizing the shape of the spectrally transformed acoustic signal. This thesis extends this concept to include a multi-speaker continuous speech database and statistical moments of a distribution to characterize shape. A multivariate maximum likelihood classifier was used to discriminate classes. To reduce the number of features used by the discriminant model a dynamic programming scheme was employed to optimize subset combinations. The top six moments were the mean, variance, and skewness in both frequency and energy. Results showed 85% classification on the full database of 952 utterances. Performance improved to 97% when the discriminant model was trained separately for male and female talkers.

Table of Contents

Chapter 1
Introduction

Artificial speech recognition has been a much sought after goal by researchers for over 30 years. One aspect of speech recognition is the phonetic classification of stop consonants, i.e., /p,t,k,b,d,g/. In conversational English, stop consonants account for over 17% of all phonemic occurrences [MINE78]. Therefore, accurate classification of stops across all phonetic contexts becomes a significant element of a complete recognition system.

The classification problem can be separated into three distinct stages. The first, data capture and signal processing, involves analog-to-digital conversion and, in most cases, some type of spectral analysis. The second stage involves extracting information to identify the phoneme; for example, peak frequencies, formant transitions, voice onset time, etc. The third stage involves the application of some type of classification algorithm, such as, Gaussian, Bayesian, Maximum-Likelihood, K-nearest neighbor, neural network, etc.

This work entailed the classification of stop consonants in a speaker independent continuous speech environment. The approach here consisted of extracting information about the consonant from a previously hand-marked segment in a database. Analysis of the token was performed using running spectra. Parametric representation of each spectrum was done using spectral moments as measures of gross spectral shape. These moments are taken as discrete features and sent to a maximum likelihood algorithm for classification.

Chapter 2 outlines the fundamental recognition problem and describes the recognition environment for this work. Also discussed are acoustic-phonetic invariance problems and principles of speech production and perception. Chapter 2 concludes with a discussion of past automatic recognition systems.

Chapter 3 describes the automatic stop consonant classification system for this work. This includes a description of the research objective, database, signal processing, features, classifier, and performance metrics.

Chapter 4 describes the results of classification experiments as using a multi-speaker continuous speech database. Multiple analysis formats are used and the results tabulated. Supplemental analysis is performed followed by training and testing results.

Finally, Chapter 5 concludes with a summary of the results and prospects for future work.

Chapter 2
Background

2.1 Speech Recognition

The goal of a general speech recognition system has not yet been realized. There are many factors that control the complexity of the speech recognition problem. Some fundamental considerations for a system are:

- Speaker independent or speaker dependent,
- Large vocabulary or small vocabulary
- Continuous speech or discrete words.

These considerations, among others, dictate the overall complexity of the speech recognition system. The remainder of chapter 2 highlights information gathered for the design of a phonetic recognition system for stop consonants in a multi-speaker, large vocabulary, continuous speech environment.

2.2 Approaches to Automatic Speech Recognition

Present recognition systems are divided into two primary categories: template matching and acoustically based segmentation and phonetic labeling. Characteristics for each type of system are discussed below.

2.2.1 Template Matching

Template matching is done by parametrically representing an unknown word and comparing it to a library of known words. This approach works well for relatively small vocabularies, isolated words, and a single speaker. However, because the acoustic characteristics of words vary depending upon sentence environment, it is difficult to generate stable templates. This problem becomes

more difficult when considering the acoustic variations of multiple speakers. Another problem for word recognition by template matching is that the English language has approximately 60,000 words and 10,000 syllables [AHD81]. The flow diagram of Figure 2-1 illustrates concepts of template matching.

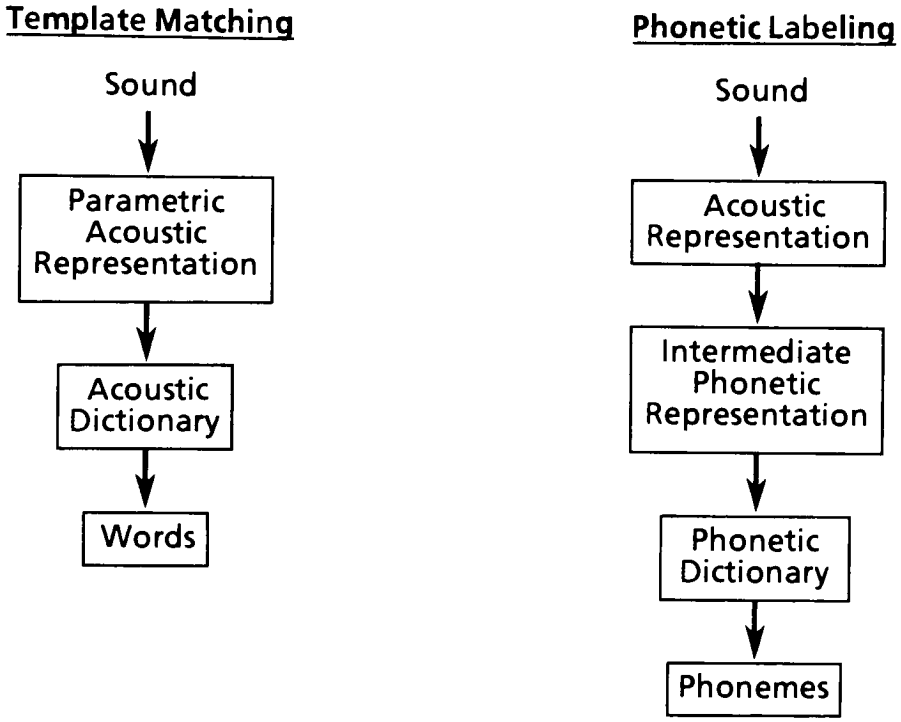


Figure 2-1 Graphical illustration of template matching and phonetic labeling.

2.2.2 Segmentation and Phonetic Labeling

The second category of recognition systems attempts to recognize phonemes directly and words indirectly as sequences of phonemes. English has approximately 46 phonemes [GLEA61]. The goal of segmentation and phonetic labeling is to convert an acoustic signal into a phonetic representation, followed by a search of a phonetically organized dictionary. Theoretically, phonetic labeling would be more robust than word level template based systems. This is primarily do to vocabulary

sizes, i.e., the word template vocabulary size is large relative to the phonetic vocabulary size. The flow diagram of Figure 2-1 illustrates concepts of phonetic labeling.

Segmentation and phonetic labeling is not without problems. The three areas of difficulty are segmentation, phonetic classification, and word identification. For connected speech, both approaches have the difficulty of word or phonetic segmentation. However, since the output of a phonetic labeling system is a sequence of phonemes, the additional component of word identification is required.

2.3 Acoustic-Phonetic Invariance Problem

The problems encountered in practical speech recognition are many. One that appears early in the recognition process is acoustic-phonetic noninvariance, i.e., the inability to determine what various instances of a phoneme have in common. Some problem areas of acoustic-phonetic noninvariance are context dependence, segmental duration, and talker variation.

Context dependence, or "coarticulation" as it is often termed, refers to neighboring phonemes affecting acoustic characteristics of the phoneme in question. What frequently happens is that each phoneme is considered as a target at which the vocal organs aim but may never reach. When the target has been approached enough to be intelligible to the speaker the destination is changed and a new target aimed for. The effects of coarticulation are illustrated in the speech spectrogram of Figure 2-2. Note that the burst frequency for the stop consonant /t/ in the word "Two" is lower than for the /t/ in the word "ten". This is due to anticipating the following vowels, /u/ and /ɛ/ respectively. Also, note that the second formant for the vowel /ɛ/ varies considerably for the words "seven", "less", and "ten". From this example, we can also observe the information structure for various

phonetic types. For example, stop consonants appear in small time durations with little frequency structure, while vowels are longer in duration and show formant frequencies and their dynamic motion.

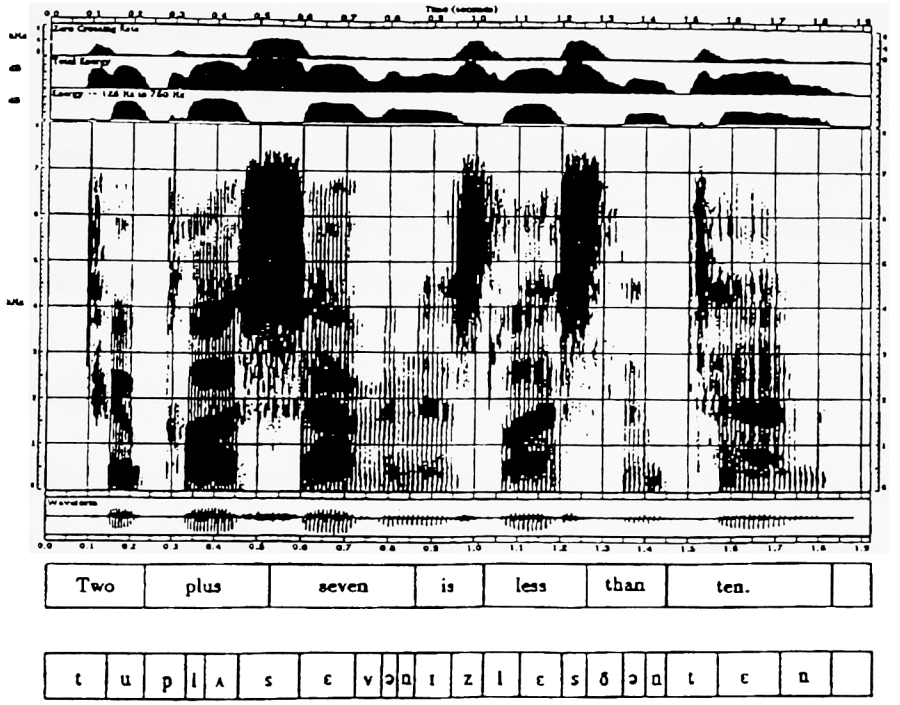


Figure 2-2 Speech spectrogram of "Two plus seven is less than ten.",to illustrate the effects of coarticulation in continuous speech. (Masters thesis of J.W. Delmege , RIT, 1988)

Segmental duration refers to the time allotted to produce the phoneme. Durations will vary widely in continuous speech depending upon context, speaking rate, and stress [KLAT76]. Klatt contends that If a change in these features is perceptible, then it potentially carries useful information from the speaker to the listener.

Talker variation refers to the fact that no two people sound alike. The speech signal contains talker-dependent variables which allow us to identify individuals based on voice quality alone. But these talker dependent variables complicate the task of developing robust classification algorithms that perform well with multiple speakers.

In addition, talker variations in individual speech include carelessness. Carelessness refers to the fact that people do not talk any more distinctly than necessary. Many short words in the English language frequently reduce to nothing more than grunts [PARS87]. Although correct pronunciation helps, it does not solve the problem. Finally, depending upon the amount of energy an individual is willing to expend for speech, the loci of formant frequencies as well as their transitional duration may change over time [PARS87]. All of the above problems contribute to the difficulties of phonetic classification.

2.4 Production of Stop Consonants

The operation of speech production is divided into two functions, excitation and modulation. Excitation of the speech production system primarily takes place at the glottis (the space between the vocal cords is called the glottis). Modulation is done by the various organs of the vocal tract, e.g., by the position of the tongue and lips.

Excitation of speech production can be done in several ways, i.e., phonation, whispering, frication, constriction, and vibration. The interaction between the multiple excitation points, e.g., phonation and constriction, cannot be neglected for stop consonants because the spectrum envelope of the source will affect the intensity levels of spectral peaks [FANT70]. Only phonation and constriction will be considered, since they are the most important excitation characteristics associated with stop consonants.

Phonation refers to the oscillation of the vocal cords as air from the lungs is being forced through the vocal cords. This oscillation is a quasi-periodic vibration that emphasizes the low frequency vibrations of the passing air.

An occlusion occurs when the vocal tract is completely shut off while the talker continues to exhale. As the occlusion is removed, the air release will be controlled by

both the pressure drop across the constriction and the size and shape of the opening at the constriction. The pressure drop across the narrow opening results in a turbulent air flow that generates acoustic noise [MINI73]. As the opening is enlarged, both the pressure drop and turbulent air flow are reduced. Minifie states that "The duration and intensity of the noise is dependent upon the time varying size of the opening and the rate at which the lungs resupply air" [MINI73, pg 265]. Upon release a small explosion occurs near the point of constriction. The combination of a short silence followed by a short burst is characteristic of a stop or plosive sound.

Modulation refers to the filtering of sound sources by the vocal tract. The vocal tract has natural resonant frequencies which vary depending on its length and shape. These resonant frequencies are called formants and are thought to be very important sources of information for identifying both consonants and vowels.

Vocal tract configurations determine the resonant characteristics of the speech. Vocal tract length controls the spacing between formants, thereby differentiating males, females, and children by having larger formant spacings and higher average formant frequencies respectively. Formant frequency patterns vary over time resulting in changes in the speech sound. It is the combination of these excitation and modulation functions which form speech.

2.5 Acoustic Characteristics of Stop Consonants

The stop consonants in English consist of /p,t,k,b,d,g/. Acoustically, a stop appears as a short period of low signal energy followed by an abrupt release. This release appears as a large body of noise in the stop spectrum. Unlike vowels, which have long duration dynamic formant frequencies, consonantal attributes associated with the burst usually occur within approximately 20 to 40 msec [KEWL83[b], STEV73].

The following sections will discuss the two primary articulatory features of stop consonants, place of articulation, and voicing.

2.5.1 Properties of Place of Articulation

Stops can be classified based on the primary constriction point and phonation type. Place classes refer to anatomical structures in the vocal tract, which include labial (/p,b/), alveolar (/t,d/), and velar (/k,g/). Table 2-1 shows the relationship between place and voicing for English stops, while Figure 2-3 visually represents place features in a cross section of the human head.

	Labial	Alveolar	Velar
Voiced	/b/	/d/	/g/
Unvoiced	/p/	/t/	/k/

Table 2-1 Place and voicing relationships for English stop consonants.

2.5.1.1 Labials

Labial stops (/p,b/) are formed with the primary point of constriction at the lips. They are generally characterized by a rapid spectrum change in which the spectrum burst has an energy concentration that is lower in frequency than the spectrum sampled a few milliseconds later [STEV73].

Blumstein and Stevens [BLUM79,STEV78] developed a *diffuse-falling* description for the burst spectrum bilabials. A spectrum would be diffuse-falling if a line drawn through the frequency peaks of an LPC spectrum, sampled at consonant release, has a negative slope (see Figure 2-4). The burst is generally weak and spread over the entire frequency range; at times the burst may not be visible on a spectrogram

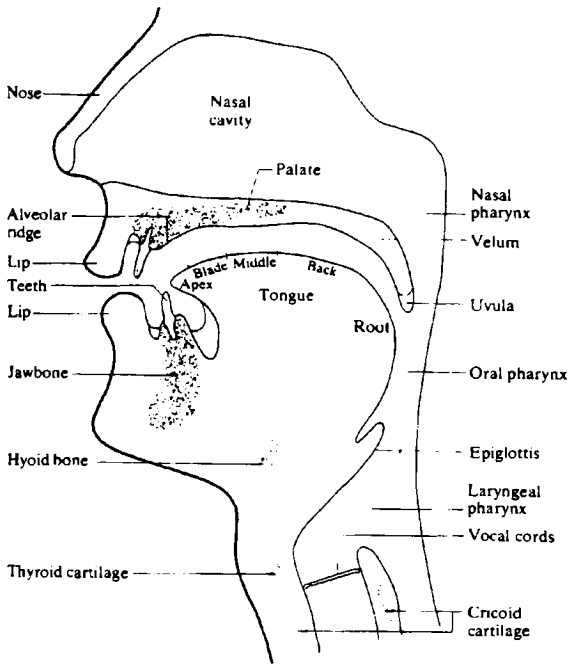


Figure 2-3. Cross section of the human head, showing the principle organs of speech (From PARS87, pg 63).

[ZUE85]. The primary concentration of energy is in the low frequency range (500-1500Hz) [HALE57].

2.5.1.2 Alveolars

The constriction point for alveolar stops lies between the palate and upper teeth (see Figure 2-3). Alveolars (/d,t/) have a broad (diffuse) burst spectrum with a large proportion of energy above 4 kHz [ZUE85, STEV73, HALE57]. There also may be a small concentration of energy in the low frequency range between 500 and 1500Hz [HALE57], but in general, the amplitude of high frequency peaks will be larger than for low frequency peaks. Blumstein and Stevens [BLUM79, STEV78] devised the definition of *diffuse-rising* for the burst spectrum of alveolars. If a line were drawn

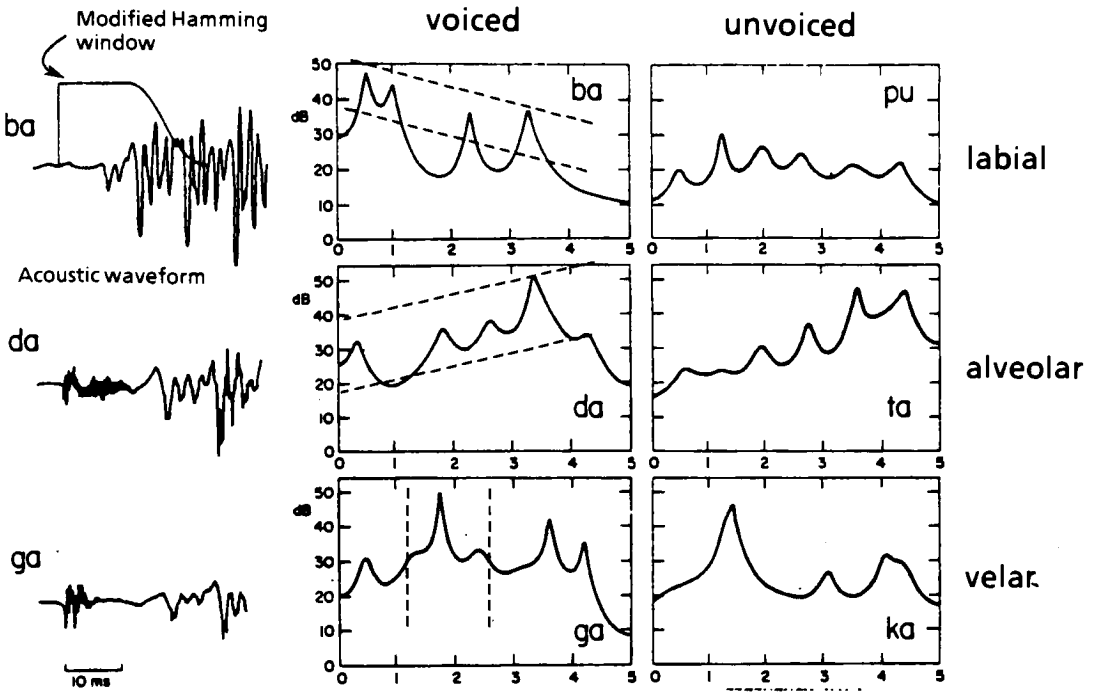


Figure 2-4 Linear predictive waveform examples to correspond to the diffuse-falling (labial), diffuse-rising (alveolar), and compact (velar) burst spectrum analysis of Blumstein and Stevens. Superimposed on ba, da, and ga are templates aimed at capturing the gross spectral shape. (From BLUM79, pg 1004)

through the frequency peaks of burst spectrum alveolars, it would have a positive slope (see Figure 2-4).

2.5.1.3 Velars

The point of constriction for velar stops (/k,g/) lies approximately near the soft palate or velum (see Figure 2-3). Velar stops have a compact burst spectrum with a small number of well-defined peaks [ZUE85, BLUM79, STEV78, STEV73]. Compactness indicates the relative predominance of energy centrally located around the second and third formant regions (see Figure 2-4). The literature also indicates a strong dependence on the adjacent vowels. For example, the front velar has a

spectral peak slightly above F2 of the following vowel. This spectrum can be confused with a rounded /t/ unless vowel context is considered [ZUE85].

2.5.2 Properties of Voicing

The term phonation commonly refers to oscillation of the vocal cords. Another term to describe phonation is voicing. Stops mainly appear as wideband noise, while voicing is associated with low frequency energy (generally below 250 Hz) [SEAR79, LISK64]. Lisker and Abramson [LISK64] performed extensive studies that measured voice onset time (VOT), which refers to the time from the release of the stop closure to the beginning of vocal cord vibration, for several languages. Their study found that word-initial English stops could be differentiated into two categories (voiced and unvoiced) based on measures of VOT. In general, voiced stops have a VOT less than approximately 35 msec while unvoiced stops have a VOT greater than approximately 35 msec. Another generality indicates that VOT for velars are longer than alveolars which are longer than labials [EDWA81].

2.6 Perception of Stop Consonants

Listener perception has been a tool used by several researches in attempts to isolate invariant cues in acoustic waveforms. The following studies serve to highlight some of the techniques and results found in perception research.

Cole and Scott [COLE74] used a tape splicing method to test perceptual invariance for stop consonants. The tape splicing technique was used to transpose the entire energy spectrum from one consonant phoneme in one syllable to another. The stops /b,d,g,p,t,k/ in their experiments were recorded before vowels /i,a,u/.

In the first experiment, they transposed vowels between the original consonants for all stops, e.g., the steady state /i/ from /bi/ was switched with the steady state /u/ in /bu/ and read as /b$_i$u/. For the initial CV combinations (40 each) with no transposing, the percent recognition scores never fell below 98%. For the transposed vowels, recognition results were generally greater than 82%. However, the transposed vowels for the velar stops /k$_i$u/ and /g$_i$u/, had relatively poor results, 54 and 21 percent respectively. They hypothesized that poor invariance came from the high energy in the second format of the /i/ (above 3 kHz). To test this, they low pass filtered the syllables to eliminate energy above 2 kHz. Listeners correctly identified (general accuracy greater than 80%) the low-pass filtered /k$_i$u/ and /g$_i$u/ in addition to the originally transposed syllables.

In the second experiment, Cole and Scott extended the results of experiment one to syllables produced before /a/. Following a similar procedure to the first part, consonants were transposed between /i/ and /a/ and between /u/ and /a/ for unvoiced stops. The results were 89% correct recognition for /i/-/a/ transpositions and 91% for the /u/-/a/ transpositions.

Cole and Scott concluded that a high degree of invariant perceptual information exists in the entire energy spectrum for stop consonants of labial and alveolar place of articulation. Although they resolved difficulty with /k/ and /g/ before /i/, they believe that some context dependence still exists for velar stops. In summary, they believed that no primary cues exist to allow for definitive perception, but rather that all cues available are analyzed.

The perception studies of Wally and Carrell [WALL83] were aimed at establishing the importance of the onset spectrum as a primary cue for place identification, and that formant transitions offer secondary cues in voiced stop consonants (Stevens and Blumstein [STEV81]). This study was based on the work of Stevens and Blumstein [STEV78], and Blumstein et al. [BLUM79, BLUM82]. Wally and Carrell's experiments mismatched the burst spectrum with the formant transitions of another consonant, e.g., the rising formant transition associated with the consonant /b/ of /ba/ replaced the same information in the syllable /da/. It was hypothesized that the primary cue would override secondary information. In the example, the syllable /da/ still should be identified as such even though it has an erroneous formant transition.

The stimuli (three control and six mismatched) were synthetically generated via resonators connected in parallel, which allowed for better control over formant amplitudes than resonators connected in a cascading fashion. The formant duration and amplitudes were varied to correlate with different place of articulation, e.g., first formant durations varied by 20, 35, and 45 msec and the second and third formant starting frequencies varied between 900/2000Hz, 1700/2800Hz, and 1640/2100Hz for /ba/, /da/, and /ga/ respectively. The attached vowels were /a/ and /u/. Thirty six listeners were divided into two groups, where one group heard variations on /a/ the other heard variations on /u/. The control and experimental stimuli were played in random order, ten times each for the controlled and five each

for the experimental. Listeners were told they would hear syllables of /ba/, /da/, and /ga/, or /bu/, /du/, and /gu/ and asked to record their response as quick as possible.

The results showed that both groups achieved high recognition for the control stimuli and appeared to classify experimental stimuli slightly more by formant transition than by onset. They concluded, in agreement with Stevens and Blumstein, that the two cues, i.e., onset and formant transition, coexist in CV syllables and that neither one is completely sufficient for place identification.

Kewley-Port, Pisoni, and Studdert-Kennedy [KEWL83[b]] have done extensive work in analyzing acoustic cues for place of articulation. Early work by Kewley-Port was concerned with establishing time varying cues to place of articulation [KEWL83[a], KEWL81]. Her most recent work, with Pisoni and Studdert-Kennedy, involves perception studies to establish timing information for place identification. Their study consisted of three experiments. The purpose of the first experiment was to determine the time necessary to classify natural CV waveforms. The second and third experiments used synthetic speech to determine whether static or dynamic acoustic cues are used to identify place of articulation.

For experiment one, 30 CV syllables spoken by two male talkers were extracted from the carrier phrase "Teddy said CV". The syllables covered all combinations of voiced stops /b,d,g/ and vowels /i,e,a,o,u/. To verify the data set, the stored CV syllables were played to six new listeners who correctly identified 99.8% of the stop consonants. The CV syllables were separated from the carrier phrase and edited into five new CV segments. The segments contained the burst only, burst plus one, three, five, and seven pitch periods. Because of the difficulty in isolating the burst only for /b/'s, the segment of burst plus two pitch periods was substituted for the burst only spectrum. The resulting 150 segments were randomized and played to ten new listeners who were asked to classify by letter, i.e., b, d, g, p, t, k.

Results showed that approximately 95% recognition was achieved with burst plus two pitch periods for labial, burst plus one pitch period for alveolar, and burst plus three pitch periods for velar stops. However, by going as high as burst plus seven pitch periods, both labial and alveolar recognition increased to 100%, while velar recognition stabilized at approximately 95%. Upon further analysis, it was found that velar stops exhibited much more context dependency than either labial or alveolar stops. Figure 2-6 summarizes the percent correct recognition at duration times of 20 and 40 msec.

consonant	20msec	40msec
/b/	96	99
/d/	94	98
/g/	73	90

Figure 2-6 Recognition results from perception experiments for voiced stops. [KEWL83[b]]

The second experiment was similar to the first except that synthetic, instead of natural, speech was used. For the static spectrum, i.e., burst only, the burst spectrum was synthesized according to the templates suggested by Blumstein and Stevens [BLUM79], while the dynamic, or running spectral analysis, were modeled after natural speech. Both were synthesized with the Klatt digital synthesizer [KLAT80]. The stops /b,d,g/ were each paired with vowels /i,a,u/. Durations were set to either 20, 30, or 40 msec.

The results for natural speech, synthetic running spectra, and synthetic static spectrum were 94%, 78%, and 68% respectively. In addition, a confidence measure was established for each subject, which categorized guess, sure, and very sure ratings. These values showed that the static spectrum was guessed at three times more than natural or running spectra, and the very sure rating was 15% higher for both natural and running spectra.

Kewley-Port et al. concluded that information for place of articulation recognition is recognized better by use of running spectral analysis and that the duration necessary for identification is 20 to 40 msec relative to the burst.

In summary, Cole and Scott, and Wally and Carrell [COLE74, WALL83] support the idea that a high degree of invariant information exists in the entire spectrum for CV syllables, i.e., the release burst plus contextual information. Finally, from Kewley-Port et al., place information is better recognized using running spectral analysis with the relevant information occurring in the first 40 msec relative to the burst.

2.7 Automatic Classification of Stops

This section describes studies which developed and tested algorithms for classifying stop consonants according to voicing and/or place of articulation. The classification studies are separated into two groups, those that analyze a static spectrum, and those that perform dynamic analyses.

2.7.1 Static Classification

Edwards [EDWA81] studied several different features to assess their relative importance in automatic classification of voicing and place of articulation. His database was generated by ten adult talkers (five males and five females) using intervocalic English stops in the sentence "Please say huh/CVt/ again". The consonants were /p,t,k,b,d,g/, and the vowels were /i,ɛ,a,u/. The initial data set was hand segmented into the sound classes voicing, aspiration, and silence. Voiced segments were divided into individual pitch periods while aspirated segments were divided into 10 msec intervals. This was followed by Fourier spectral analysis. Classification was based on the maximum *a posteriori* probability to separate voicing and place information.

Edwards summarized his findings with two tables that rank ordered the features for both voicing and place. Eleven features were quantified for voicing ranging in correct decision probability from .62 to .99. The top four were:

VOT (knowing correct place of articulation)	.99
VOT (not Knowing correct place)	.97
Percent voicing during closure	.92
Duration of voicing during closure	.89

Ten features were quantified for place of articulation ranging in correct decision probability from .54 to 1.00. The top three were:

Double burst release (not always seen)	1.00
F/B ratio (formant freq divided by burst release freq)	.81
Burst frequency	.74

Edwards concluded that although the results were encouraging, interdependent features would be required for accurate recognition of stop consonants in word-medial position.

2.7.2 Dynamic Classification

Searle, Jacobson, and Rayment [SEAR79] developed a system for stop consonant classification based on studies of auditory physiology and perception. For the input of their system, they devised a bank of eighteen $\frac{1}{3}^{rd}$ octave filters followed by envelope detection to simulate auditory bandwidth sensitivity. The filter center frequencies ranged from 125Hz to 6.3KHz, with the octave base frequencies of 125, 160, and 200Hz. The signal was then passed through log amplification followed by A/D conversion. Characteristics extracted from the filter outputs were:

The abrupt onset (release burst), measured in the high frequency outputs (630Hz - 6.3KHz),
- VOT,
- Location and shape of spectral peaks close to the burst,
- Location and shape of spectral peaks in formant transition region, and
- Formant slopes measured during the first 100 msec after the burst.

These characteristics were used to determine where and when to measure features. Classification was done using a discriminant analysis program. Their database consisted of isolated words spoken by 15 adult males and females.

One of the unique algorithms implemented by Searle et al. was release burst detection, which formed the basis of all other feature information. According to Searle et al., the release burst energy was readily observed in the upper filter

channel outputs. Their algorithm summed across channels (630Hz-6.3KHz) and registered release when the average rate of change exceeded an empirically determined threshold of 17 dB in 3.2 msec. Burst onset was correctly identified for 97% of the database. A similar algorithm was used to detect voicing onset. Voicing onset was detected when the derivative sum of the low frequency channels exceeded an experimentally determined threshold. VOT could then be determined by measuring the interval from release to voicing onset.

For their experiments, they arbitrarily divided the data set for training and testing. The test set also contained new speakers and utterances. Before final classification, they made a high level sort decision based on VOT. They found that VOT sorted voice class with nearly 100% accuracy. The classification results are shown in Figure 2-6.

stops	recognized / total	% correct recognition	test criterion
/b,d,g/	69/77	90%	training set
/b,d,g/	72/92	78%	test set
/p,t,k/	67/71	94%	training set
/p,t,k/	45/56	80%	test set

Figure 2-6 Classification results from Searle et al [SEAR79].

Overall performance on 148 test tokens (including burst detection and classification errors) was 79%.

They concluded that classification performance equaled or exceeded other systems, and claimed their design to be theoretically more resistant to extraneous inputs, e.g., noise contamination, multiple reflections, and other distortions. They also, to the best of my knowledge, are the first to demonstrate burst detection and classification in the same system.

Kopec [KOPE84] used LPC spectra with a variable number of frames and three classifiers (K-nearest neighbor (KNN), minimum vector quantization distortion (VQ), and maximum likelihood (ML)) to classify voiceless stop consonants. "In KNN recognition each stop was represented using a very large collection of templates. The VQ recognizors were an attempt to reduce the number of templates by clustering the training spectra of each stop category into a small number of subclasses. Each subclass was represented using a single template" [KOPE84]. His database (936 tokens) was generated by 13 adult males and 13 adult females. Each speaker contributed 36 tokens, 18 from "/CVb/ is the word" and 18 from "Please say huh/CVb/ again". The stops were /p,t,k/ and the vowels were /i,e,æ,a,o,u/. The burst locations were hand marked. The signal was processed with 12 pole LPC (no preemphasis) and a 25 msec Hamming window. Kopec tested spectral sequences consisting of 1,3, or 12 frames. The 1- and 3-frame sequences started 5 msec before release and were non-overlapping. The 12-frame sequence started 15 msec before release and shifted at 5 msec intervals. The feature for the KNN classifier was the average distance from the unknown to the k ($k = 5$) members nearest to it. The distance between two tokens was the sum of the gain-optimized Itakura-Saito distance [PARS87]. VQ classification was used on 1, 3, and 12 spectra (4 codebooks each), VQ plus maximum likelihood on 1, and KNN for 1 and 12 (a total of 15 classification systems were tested). Each stop was broken into four subcategories, male or female and front or back vowels, yielding a codebook of twelve classes. Both classifiers were trained using twelve tokens from each speaker.

For cross study comparisons, the results were tabulated using a codebook size of 78 for the KNN classifier (total tokens divided by codebook classes). The results for single spectrum analysis are shown in Figure 2-7.

Classifier	Codebook Size	# Itakura-Saito Calculations	% Correct
KNN	78	900	96
VQ	1	12	87
	2	24	90
	4	48	90
	8	96	90
VQ + ML	1	12	90

Figure 2-7 Overall recognition results from single spectrum analysis [KOPE84].

Although this work only applied to voiceless stops in CV context and is computationally expensive, Kopec concludes by contending that template matching can be a viable strategy for high performance stop classification.

Demichelis et al. [DEMI83] developed a computerized classification system based on multiple acoustic cues (e.g., transitions, bursts, and timing), rules, and fuzzy set theory. The features, such as, formant loci, formant frequencies, rates of change, and burst spectra, were combined in context dependent rules and possibility distributions ("possibility distributions are fuzzy sets defining the meaning of linguistic variables", [DEMI83, pg 363]). Classification was accomplished by scheduling rules from statistical decision criteria to allow competing hypotheses to continue processing.

Their database consisted of syllable segments extracted from sentences generated by four adult males and one adult female (no other specifics were discussed about the database). After the rules were established, the original speakers, plus new speakers, produced a set of new utterances. To facilitate classification, the decision for voicing mode was made immediately. Although

classification of place of articulation knowing voicing mode was slightly higher, the overall recognition rates never fell below 90%.

They concluded that recognition rates improved when more rules were added to the system and that important rules for new cues would offer more improvement than refining algorithms to match the data sets. They also claimed that their results exceeded those of other designs and that this approach facilitates modifications for improvements.

Forrest et al. [FORR87] took an approach to quantify spectra via the moments of a distribution in both Bark and linear frequency scales. Since recognition can be achieved knowing shape characteristics of spectra ([STEV78, BLUM79, KEWL83[a]]), the mean, skewness, and kurtosis were used as features to capture spectral tilt and compactness.

Their database was generated by five adult males and five females and consisted of thirty one words containing voiceless stops (/p,t,k/) and fricatives in initial position. The speakers, upon being played their initial recorded speech, repeated the words in the carrier phrase "I can say ____, again". With the aid of a graphical computer program for speech, the burst and third pitch period of transition (consistent with KEWL83[b]) were hand marked.

The speech waveform was preemphasized by differencing, followed by a 400 point (20msec) Hamming window, and 512 point Fourier transform. The power spectrum were normalized prior to feature extraction. The spectral shift interval was 10 msec. Classification was done using stepwise discriminant analysis.

The experimental results showed that the second moment (variance) did not add to stop discrimination. After examining linear scale graphical plots of mean, skewness, kurtosis, and time interval of calculation, they concluded that "... labial and alveolar stops are distinct in terms of mean and skewness, while the velar stops

are distinguished by their kurtosis". The discriminant analysis system was trained on the males' data set and tested with the females' data set. The best results came by using three running spectra starting at burst release, i.e., the first 40 msec. The percent correct results are shown in Figure 2-8.

	/p/	/t/	/k/
males (trained & tested)	95.4	88.0	92.6
females (tested)	90.5	96.5	93.6

Figure 2-8 Recognition results from spectral moments in three running spectra [FORR87].

Forrest et al. concluded that although the results are preliminary, they present this quantitative procedure as a new approach for voiceless stop consonant classification. They also claimed this test to be the first demonstration of cross gender classification without the use of waveform (acoustic or spectral) normalization techniques.

Yoder and Jamieson [YODE86, YODE87] addressed the area of signal processing to improve classification accuracy using template based features. The signal processing parameters that were studied included the number of frames used, overlapping versus non-overlapping data windows, data window size, amplitude normalization, and the type of transforms taken. Their database was extracted from the multi-dialect Texas Instrument (TI) database. The words were chosen to represent all CV combinations of stops /p,t,k,b,d,g/ and vowels /i,ɪ,ɛ,æ,ʌ,ɜˆ,u,a,ɔ/. The speakers contained an equal number of male and female adults and children. The database was high-pass filtered with the cutoff at 6.25 kHz and sampled at 12.5

kHz. The data were classified using a K-nearest neighbor (K = 5) algorithm, testing all tokens as unknown and training on a random half of the tokens.

Based on initial testing, recognition improved when different data window sizes were chosen for the two voicing modes and when three consecutive spectra were used with a 4 msec overlap for voiced and non-overlapping for voiceless. They used a VOT threshold to differentiate voicing mode and then assigned the window lengths of 8 msec and 12 msec for voiced and unvoiced respectively. Both amplitude and energy normalization was employed in the various experiments. In attempts to reduce the influence of signal variations associated with age and gender, they implemented a signal normalization scheme that involved taking multiple transforms, i.e., Fourier and Mellin, in what they called the Mellin-Fourier Homomorphism (MFH).

Results achieved, from their most recent work, for both voicing and place of articulation, are summarized with respect to vowel context in Figure 2-9.

/CV/	front	central	back	Total
/p/	95	97	98	97
/t/	92	81	88	89
/k/	91	100	98	95
/b/	93	81	97	93
/d/	83	50	79	77
/g/	77	91	90	84

Figure 2-9 Recognition results for vowel position using three running MFH spectra [YODE87].

They achieved 85% and 94% recognition for voiced and unvoiced respectively. The overall recognition was 89%.

Yoder and Jamieson concluded that the MFH was slightly better than Fourier analysis, the VOT threshold was useful in establishing the frame length and degree

of overlap, better results were achieved with multiple versus single spectra, and some form of normalization is better than no normalization.

In summary, several studies have been reviewed which suggest guidelines for the configuration of a stop consonant classification system. From the perception studies, there is general agreement that enough information exists in the first 40 msec following the burst, and that recognition is based on dynamic rather than static spectral information. The classification studies suggest that spectral shape holds a strong cue to recognition. The approach that seemed most promising, from the invariance issue of age and gender, is that of Forrest et al. [FORR87] which attempted to capture the important shape characteristics via spectral moments. Further development of their work is the focus of this thesis.

Chapter 3
Implementation

3.1 Research Objective

The objective of this thesis is to classify stop consonants according to place of articulation and/or voicing in a speaker independent, continuous speech environment. This objective is pursued by extending the past research of Forrest et al. [FORR87]. Although their work achieved high recognition rates, the database, signal processing, and feature implementation can all be expanded. The database is expanded from prevocalic voiceless stops to all stops, and includes stops from continuous speech instead of a fixed word in the same carrier phrase. Signal processing is expanded to test a variety of options that allow control over data window size, spectral techniques (FFT versus LPC), data preemphasis, etc. The features are expanded to include mean versus median as a measure of central frequency and variance versus average deviation as a measure of diffuseness.

3.2 System Overview

The objective is to establish a computer program, or sequence of programs, to classify stop consonants. Figure 3-1 illustrates the program flow for the stop consonant analysis program (subsequently referred to as SAP). Flexibility in parameter selection is built into SAP to facilitate experimental changes. The controlled loop of 'Spectral Analysis', 'Feature Extraction', and 'Shift Position', allows extensive flexibility for running spectral analysis. The 'Feature Extraction' module stands alone and will be discussed in detail later in this chapter. In order to minimize the number of features for classification, several options in a 'Feature Compression' module are available and are also discussed later in this chapter.

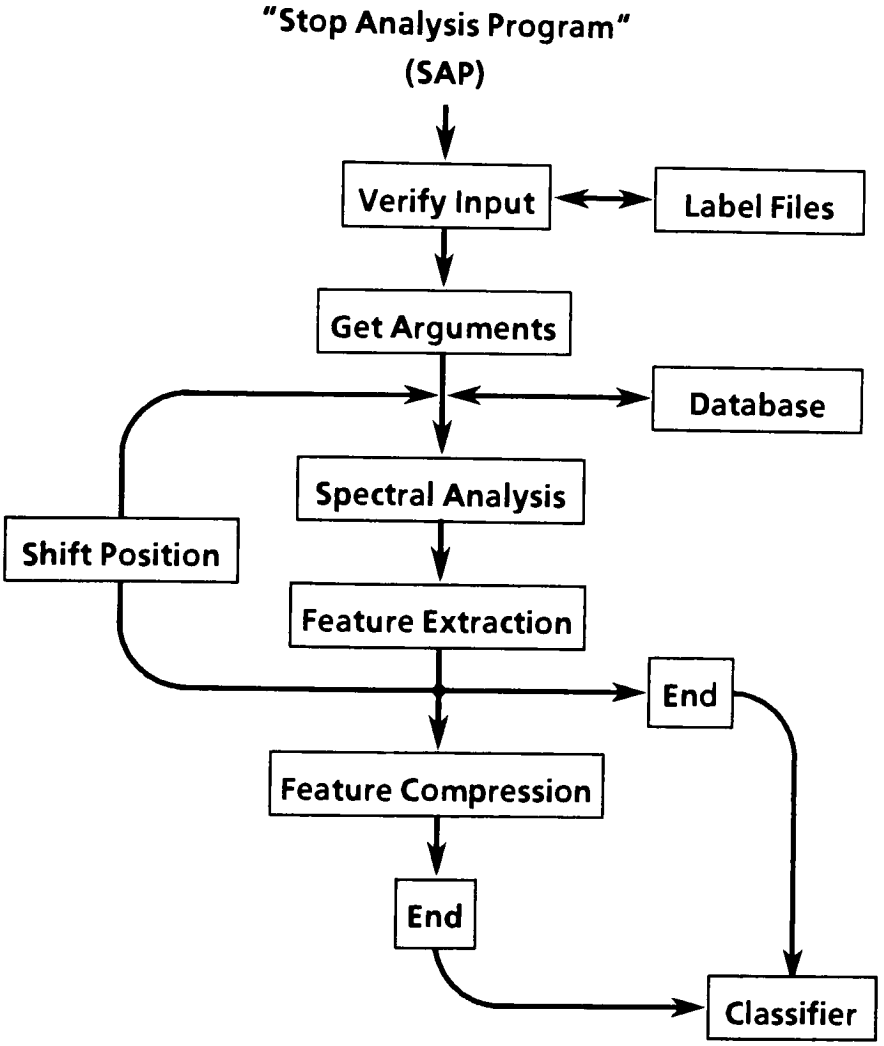


Figure 3-1 Block diagram of the stop analysis program (SAP).

Database

Speech signals for training and testing were taken from the Carnegie Mellon University (CMU) continuous speech database. The vowel database (approximately 300 tokens) was generated by four male and three female speakers uttering ten sentences each. SAP is designed to access the database label files (provided by CMU) that contain hand-marked start and stop points for all phonetic segments of speech. The waveforms were originally sampled at 16 kHz with 12 bits of amplitude

resolution, and subsequently lowpass filtered at 6 kHz and downsampled to 12.8 kHz. During supplemental experimentation, tokens were included from the CMU stop dense database (approximately 1140 tokens) . This section of the database was generated using ten male and eight female speakers uttering 10 sentences each.

Signal Processing

As with past research, experiments here will analyze the difference between static and running spectra for different signal processing parameters. The data window was set to 10, 15, 20, 30, or 40 msec. The data window was a cosine modified Hamming, i.e., the left side of the data window was not scaled. The number of spectra analyzed varied from one to six with a shift size equal to fifty percent of the data window. The FFT was 512 points and zero padded if not filled. For the static spectrum, both 256 and 1024 point FFT's were taken to determine the effect of frequency resolution. The spectral analysis consisted of Fourier analysis or Linear Predictive Coding (LPC). Features were extracted from either the log or normalized power spectrum.

Four variations of the frequency spectrum were compared. They were the FFT, LPC, smoothed FFT, and FFT that included high frequency preemphasis by simple differencing. Finally, the aforementioned processing was applied to three stop consonant categories, i.e., stops preceded by a closure segment, stops not preceded by a closure segment, and all stops together.

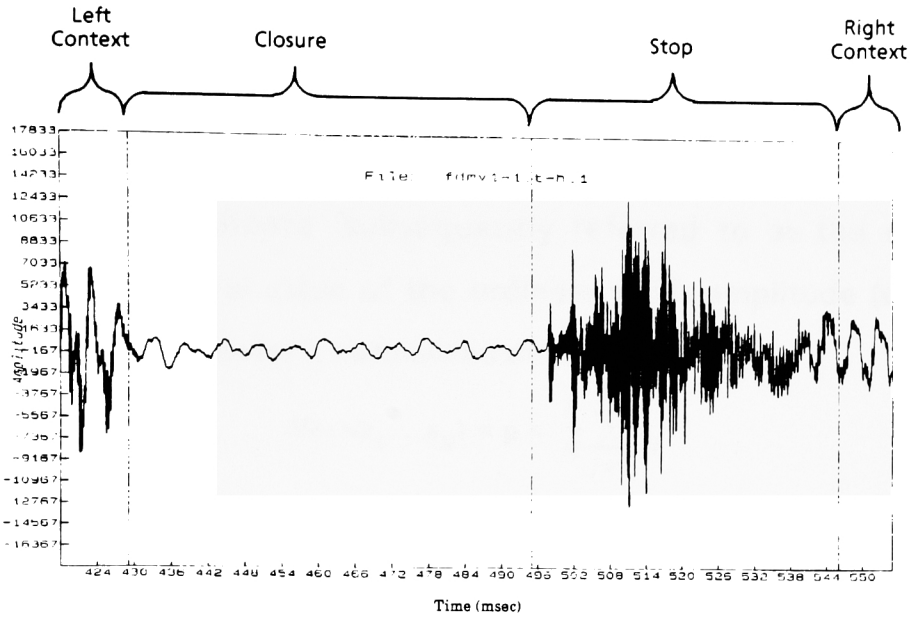
One of the identifying features of stop consonants is the silence (from closure) before the burst release. In general, there has not been a lot of work classifying stops that are not preceded by a closure segment. For example, in the continuous speech database, the phrase "The auctioneer accepted the bid." has the /d/ in the word "accepted" marked as a simple d-closure. In the phrase "He left them...", the /t/ from the word "left" was also marked as a t-closure. Subsequent classification

attempts to recognize the closure segment. If stops do not specifically contain a preceding closure segment, they are marked as not preceded by a closure. Since the distribution of stops not preceded by a closure in the CMU database constitute more than $\frac{1}{3}$ of all stop tokens, the experiments should be complete in representing stops in all positions, i.e., word-initial, word-medial, and word-final.

Since algorithmic accuracy in finding the burst release (start) and voice onset (stop) for stop consonant segments is less than 100%, two data processing procedures were considered. The first attempts classification based on both the start and stop times, i.e., from burst release to voice onset. In this procedure, the number of frames depends on the duration of the consonant. This procedure attempts to establish whether totally marked discrete stops can be classified. It is well documented that phonetic segmentation is a difficult problem. However, if classification is possible in this mode, results would suggest that more work needs to be done with segmentation and not classification. The second approach attempts classification using the start time boundary only, i.e., burst release. In this case, multiple frames simply run into whatever data follow the token in question. The final point concerns context, i.e., the speech data preceding and following the token in question. Context frames precede and follow the stop token and are fixed at 10 msec each. Figure 3-2 illustrates the data format as viewed by SAP.

3.2.1 Feature Descriptions

In order to characterize the spectral shape, Forrest et al. [FORR87] proposed the moments of a distribution: mean, variance, skewness, and kurtosis. SAP takes the following features from the spectra in attempts to measure their relative importance for classification:



Data from stop token "fdmv1-1.t-h.1".

Figure 3-2 Shows the data format as viewed by SAP. The stop data includes a 10msec spectral slice at each end to encode context dependent information.

The central frequency (or central clustering) of a distribution is usually measured by the first moment (mean). For a normalized distribution, the mean is only relevant for the x-axis. Since the signal processing includes the log power spectrum, the mean is established for both the x (abscissa) and y (ordinate) axes. An alternative measure of the central frequency is the median. The three estimators for central clustering are:

- Mean of the abscissa (subsequently referred to as the center frequency). The distributions central value of the abscissa (x-axis) which results in units of frequency. (The mean value from [FORR87] is the center frequency for this work.)

$$Centf(x_1 \ldots x_N) = \frac{\sum_{j=1}^{N} (x_j * j)}{\sum_{j=1}^{N} x_j}$$

- Mean of the ordinate (subsequently referred to as the mean). The distributions central value of the ordinate yields amplitude (energy). This value is meaningless for a normalized distribution.

$$Mean(x_1 \ldots x_N) = \mu = \frac{1}{N} \sum_{j=1}^{N} x_j$$

- Median. The median of a probability distribution function p(x) is the value for which larger and smaller values of x are equally probable.

$$Med(x_1 \ldots x_N) \rightarrow \int_{-\infty}^{x_{med}} p(x)\,dx = \frac{1}{2} = \int_{x_{med}}^{\infty} p(x)\,dx$$

If the values x_j $j=1,\ldots,N$ are sorted into ascending order, then the formula for the median is:

$$x_{med} = x_{(N+1)/2} \qquad N\,odd$$

$$x_{med} = \frac{1}{2}(x_{N/2} + x_{(N/2)+1}) \qquad N\,even$$

Forrest et al. discovered that in the normalized spectrum the second moment (variance) did not distinguish stops according to place. Therefore, in addition to variance, the mean absolute deviation or average deviation is used.

- Variance. The dispersion around the mean.

$$VAR(x_1 \ldots x_N) = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \mu)^2$$

- Mean absolute deviation (MAD). The dispersion around the mean. MAD is claimed to be a more robust estimator than the standard deviation (robust meaning, ". . . estimation for broad distributions with significant numbers of "outlier" points." [PRES88]).

$$MAD(x_1 \ldots x_N) = \frac{1}{N}\sum_{j=1}^{N} |x_j - \mu|$$

- Skewness. Characterizes the degree of asymmetry (tilt) of a distribution around its mean. The value σ is the standard deviation.

$$Skew(x_1 \ldots x_N) = \frac{1}{N}\sum_{j=1}^{N} \left[\frac{x_j - \mu}{\sigma}\right]^3$$

- Kurtosis. A measure of the relative peakedness or flatness of a distribution.

$$Kurt(x_1 \ldots x_N) = \left\{\frac{1}{N}\sum_{j=1}^{N} \left[\frac{x_j - \mu}{\sigma}\right]^4\right\} - 3$$

Note that when using the normalized power spectrum, moments that are a function of the mean (center frequency) produce features based on frequency. For the log based power spectrum, moments were a function of the y-axis mean and produce features based on energy.

3.3 Classification

Classification was done using a statistical distance (sd) program written by Robert Gayvert (RIT Research Corp.). The program performs quadratic discriminant analysis for two or more groups given one or more features. Discrimination was measured by the generalized squared distance assuming that the groups distributions are approximately multivariate-normal. The distance equation is given by:

$$D_i(\mathbf{f}) = (\mathbf{f} - \overline{\mathbf{f}}_i)^T \mathbf{R}_i^{-1} (\mathbf{f} - \overline{\mathbf{f}}_i) - ln|\mathbf{R}_i|$$

were $\mathbf{f}_i$ is the vector of m features associated with word i, $\overline{\mathbf{f}_i}$ is the mean of all $\mathbf{f}_i$, and $\mathbf{R}_i$ is the covariance matrix for feature i. This equation is also considered to be the basic maximum-likelihood criterion.

3.4 Data Separation

As discussed in section 3.2, data were separated into two classes: those that utilize the start (burst release) and stop (voice onset) time boundaries and those that considered just the starting time boundary. The second level of separation is to apply the two experimental classes to the various data groups, i.e., those with preceding closure segments, those without preceding closure segments, and all stops. Finally, results are established for each subset category of stop consonants, i.e., phonation type, place of articulation, and discrete stops. This hierarchy of data separation is shown in Figure 3-5.

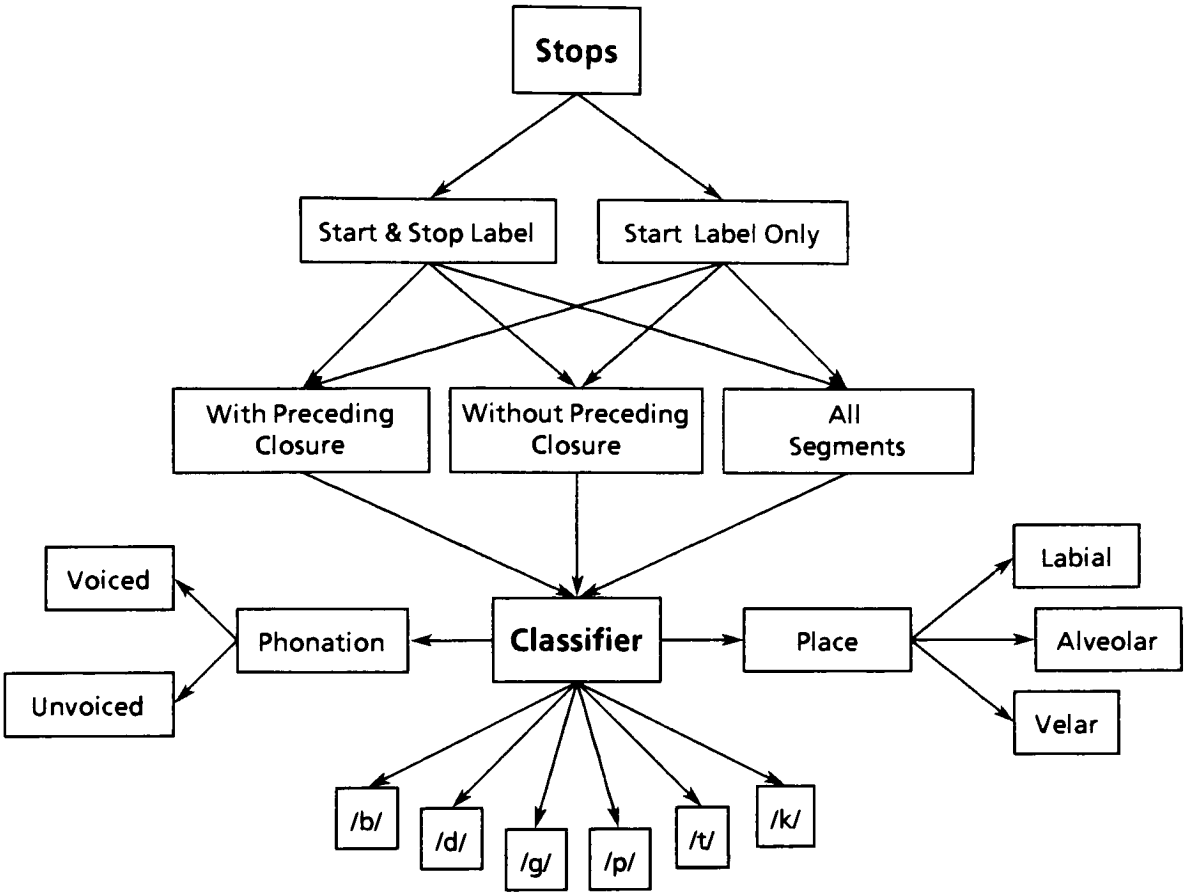


Figure 3-5 Categories of stop consonants classified via statistical description.

Chapter 4
Experimentation and Results

Initial testing was applied to the data groups discussed in section 3.4. These tests used only the CMU continuous speech vowel database. After careful scrutiny of the results, supplemental experimentation was devised to further study the implications of feature selection. Comparisons between features based on energy and frequency led to the use of dynamic programming to optimize subset feature combinations. The latter work was done using both the CMU vowel and stop databases.

4.1 Signal Processing Tests

The goal of preliminary recognition testing was to narrow the field of variables associated with signal processing. Relationships between the feature types used and signal processing variables were undetermined at the beginning of these tests.

The CMU vowel database was tested using the data group separation discussed in section 3.4. Data were broken up into three main categories: discrete stops (/b,d,g,p,t,k/), phonation type (voiced vs unvoiced), and place of articulation (labial, alveolars, velars). These tests, which ran over 6000 iterations of the classification program, showed that Fourier spectral analysis equaled or outperformed Linear Prediction, smoothed Fourier, and Fourier with preemphasis by simple differencing. It was also shown that 4 running frames (starting at burst release), with 15 msec data window and 7.5 msec shift size, yielded the best results for discrete stop classification using all tokens (stops with and without a preceding closure). Better performance was achieved when using an FFT window of 512 points versus 256 or 1024. Another noteworthy issue was the selection of data windowing filters, e.g., Hamming, Hanning, Welch, etc. Results showed that the running cosine windowing filter (modified Hamming) outperformed the more traditional windowing filters,

including no filter at all, consistently by 2 to 3 percent. In addition, it was found that the use of context (the surrounding phonetic environment) improves classification accuracy. Finally, in comparing similar feature types taken from both energy and frequency spectra, minimal classification differences were observed.

As a result of this previous work, the following parameters were fixed for the remaining tests. They include,

192-point data window (15 msec),

Shift size equals 50% of the data window (7.5 msec),

Cosine Hamming data windowing filter (applied to 192 points),

Fourier analysis using 512 points (zero padded if not filled),

4 running frames plus left context frame,

Log power spectrum for energy features,

Normalized power spectrum for frequency features.

A difficulty associated with early testing was the low number of tokens used. In many cases, particularly for the discrete stop classification, a class dropped out when the number of features used approached or exceeded the number of tokens in that class; this was a byproduct of the classification program. However, the features discussed in section 3.3 general performed well in discriminating stop consonants.

4.2 Discrete Feature Analysis

The goal here was to observe the relative importance of discrete features in variable frame sequences. In particular, those features calculated from energy and frequency. Because of the low number of tokens in the data group of non-closure prefixed and the difficulty of their automatic segmentation, the following results exclude this group.

As previously mentioned, the high order spectral moments from the log power spectrum were derived from the mean energy while the normalized spectral moments were derived from the mean frequency. All features for each format are shown in Table 4-1. The center frequency and median are the same for both spectral formats and the mean energy is not applicable for a normalized spectrum. To distinguish features between the two power spectrum formats, features were marked as follows: meany, meanx, mady, madx, skewy, skewx, kurty, kurtx, vary, varx, median. A 'y' suffix was attached to features as a function of energy while an 'x' was attached to those as a function of frequency.

Table 4-1 contains the discrete feature classification results for a variable frame sequence up to 4 frames long. To evaluate the sensitivity of left contextual information, the frame sequences were repeated including the left context frame (note that this sequence adds an additional set of features; 5 or 6 depending upon frequency or energy format respectively).

Results varied widely with feature type, number of frames, and context. The exception was that kurtosis in both spectral formats appeared to be a relatively poor classifier. In keeping with the knowledge that stop consonants can be perceived in the first forty milliseconds relative to burst release [KEWL83[b]], features of the four frame sequence with and without left context were rank ordered for both spectral formats. Table 4-2 contains this ordering. The main observation is that results increased for all features when the left context frame was added. This is not so surprising since classification was done with an additional feature set.

The following examines the rank ordering of features by their physical interpretation, i.e., spectral centers (mean, median), diffuseness (variance, MAD), tilt (skewness), and peakidness (kurtosis). General observations would suggest that energy diffuseness (MADy and vary) has a key roll in class separation. On the other

Energy spectrum # frames	Mean (y-axis)	MAD y	Var y	Skew y	Kurt y	Centf (Meanx)	Median (x-axis)
1	23.3	29.0	27.3	25.0	22.7	26.7	26.7
12	31.8	43.8	41.5	30.7	22.2	42.0	33.5
123	38.6	48.9	48.9	39.2	26.1	46.6	40.3
1234	50.0	55.1	54.5	S1.7	29.0	48.3	46.7
Lt1	31.1	31.3	38.6	30.1	21.0	35.8	30.7
Lt12	32.4	45.5	44.9	36.9	24.4	47.7	37.5
Lt123	42.6	51.1	53.4	43.2	25.6	53.4	42.6
Lt1234	52.8	59.1	56.8	48.3	33.0	56.8	48.9
Frequency spectrum # frames		MAD x	Var x	Skew x	Kurt x	Centf (Meanx)	Median (x-axis)
1		36.4	33.0	31.3	28.4		
12	na	42.6	43.8	41.5	35.8	same	same
123		47.7	44.2	48.9	42.6	as	as
1234		48.3	48.3	50.6	42.0	above	above
Lt1		35.2	34.7	36.9	26.1		
Lt12		42.0	42.0	46.6	39.2		
Lt123		48.3	51.7	51.1	43.8		
Lt1234		50.0	56.3	57.4	44.9		

Table 4-1 Classification results for discrete stops consonants (952 tokens) preceded by a closure segment for a feature sequence up to 4 frames (15 msec data window, 7.5 msec shift). Lt refers to the left context frame, while 1 thru 4 are sequential frames relative to burst release.

end, peakidness (Kurtosis) in both energy and frequency are poor class separators. Other than these two points, anything else would be speculation. Since classification is generally accomplished by feature combinations, a subset optimization scheme was required. This will be the topic of the next section.

Rank order	Energy and Frequency spectrum			
	4 frames without left c.		4 frames with left c.	
1	55.1	MADy	59.1	MADy
2	54.5	Vary	57.4	Skewx
3	51.7	Skewy	56.8	Vary,Centf
4	50.6	Skewx	56.3	Varx
5	50.0	Meany	52.8	Meany
6	48.3	MADx,Varx Centf	50.0	MADx
7	46.7	Median	48.9	Median
8	42.0	Kurtx	48.3	Skewy
9	29.0	Kurty	44.9	Kurtx
			33.0	Kurty

Table 4-2 Rank ordering of energy and frequency features from Table 4-1.

4.3 Dynamic Programming

Since no simple feature yields high results for discrete stop consonant classification, an optimized set of features was desired. The approach here used dynamic programming [PARS87] to find an optimal feature subset.

Two problems arise with discrete feature optimization (assuming many features). First, the number of permutations can be computationally very large. Second, the resulting feature optimization is only applicable for the database to which it was applied. The approach here used a feature type sequence, i.e., left context plus 4 frames, to represent one item in a subset of items. In this case, subset types for optimization were taken to be feature sequences of mean, MAD, variance, skewness, kurtosis, center frequency, and median.

The iterations required to optimize a subset of items is given by N(k-1)(N-k/2) [PARS87, p. 182]. The total number of items is N with the number of items in the subset represented by k. For example, given 6 items and requiring an optimum subset of 5 yields a maximum of 96 iterations (this number is actually less because of redundancy). An illustration of the algorithm is shown in Figure 4-1. It can be seen

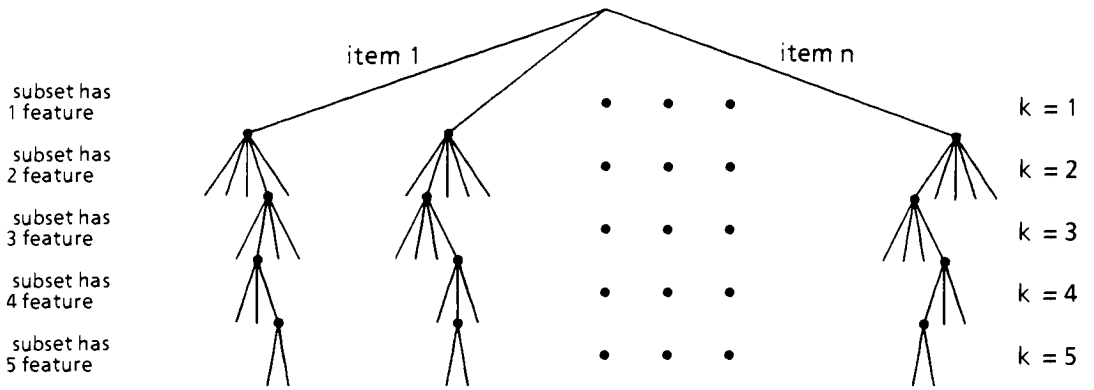


Figure 4-1. Parallelism illustration of Dynamic Programming algorithm for subset optimization.

that a true optimization occurs since the elements of the tree optimize item combinations in parallel, i.e., the best of all combinations continue processing at each subset level. When the desired subset level is reached, the subset with the best performance is the optimum set.

4.4 Feature Set Optimization

The dynamic programming algorithm was implemented for data groups of closure prefixed stops, all stops, closure prefixed males, and closure prefixed females with all features from the energy and frequency power spectra. Both the CMU

vowel and stop dense databases were used for this experimentation. The signal processing parameters were those discussed in section 4-1. Again, each item in the subset consists of a feature type sequence containing 5 discrete features (11 items for a total of 55 discrete feature values).

Table 4-3 shows the results of optimized classification as applied to all stop consonants, all closure prefixed stops, closure prefixed stops for males, and closure prefixed stops for females. For each data group, the optimized subsets are compared to those for each power spectrum format and to the energy format with the 'mean' feature removed (Energy (6)); this results in an equal set of feature type comparisons. In general, the optimized subsets outperformed the discrete energy or frequency feature sets when comparing equal numbers of feature items. Results also indicate that recognition improves when the overall data group is subdivided into smaller classes, e.g., gender separation.

Data Group (# of tokens)	Spectral Type			Variable Size Optimum Subsets		
	Freq (6)	Energy (6)	Energy (7)	Optimum (6)	Optimum (7)	Optimum (8)
All (1428)	---	---	---	71.5	75.9	77.7
Closured(952)	76.1	76.4	81.1	82.4	85.9	87.5
Females(413)	92.5	89.8	95.5 -/g/	95.6	98.1 -/g/	100 -/g/
Males(539)	89.1	89.6	92.0	93.1	95.4	97.0

Table 4-3 Percent correct classification comparison for various data groups, spectral formats and optimized subsets. The number in parenthesis following each data group is the number of tokens for that set. The number in parenthesis below the spectral type and optimum subsets indicates the number of feature types used. The /g/ after data indicates results tabulated excluding the /g/ class. Each data group used left context plus four 15 msec frames starting at burst release and shifted by 7.5 msec.

From additional data used for Table 4-3 (not shown), it was observed that classification failed when using all features. Therefore, a peek number of features must have existed for each data group. Figure 4-2 shows a plot of the percent recognition versus the total number of features used for each data group in Table 4-3. Here it is seen that recognition begins to fail when the number of features comes close to or exceeds the number of tokens in a class. The dip in percent recognition for the female curve between 30 and 35 discrete features was due to an artifact of the classification program (the /g/ class dropped out due to a low number of tokens). The data point for 55 features (11 feature types) does not specifically represent all the curves end points. Instead, it illustrates that recognition failed with results located entirely in one class.

As an indication of the potential for classification, the optimized subset for females was continued after the /g/ class dropped out. The dropout is observed at 35 discrete features (7 types). The next level optimized subset achieved 100% recognition for 378 tokens of /b,d,p,t,k/. The low data point at 40 features is due to the lost /g/ class tokens (35).

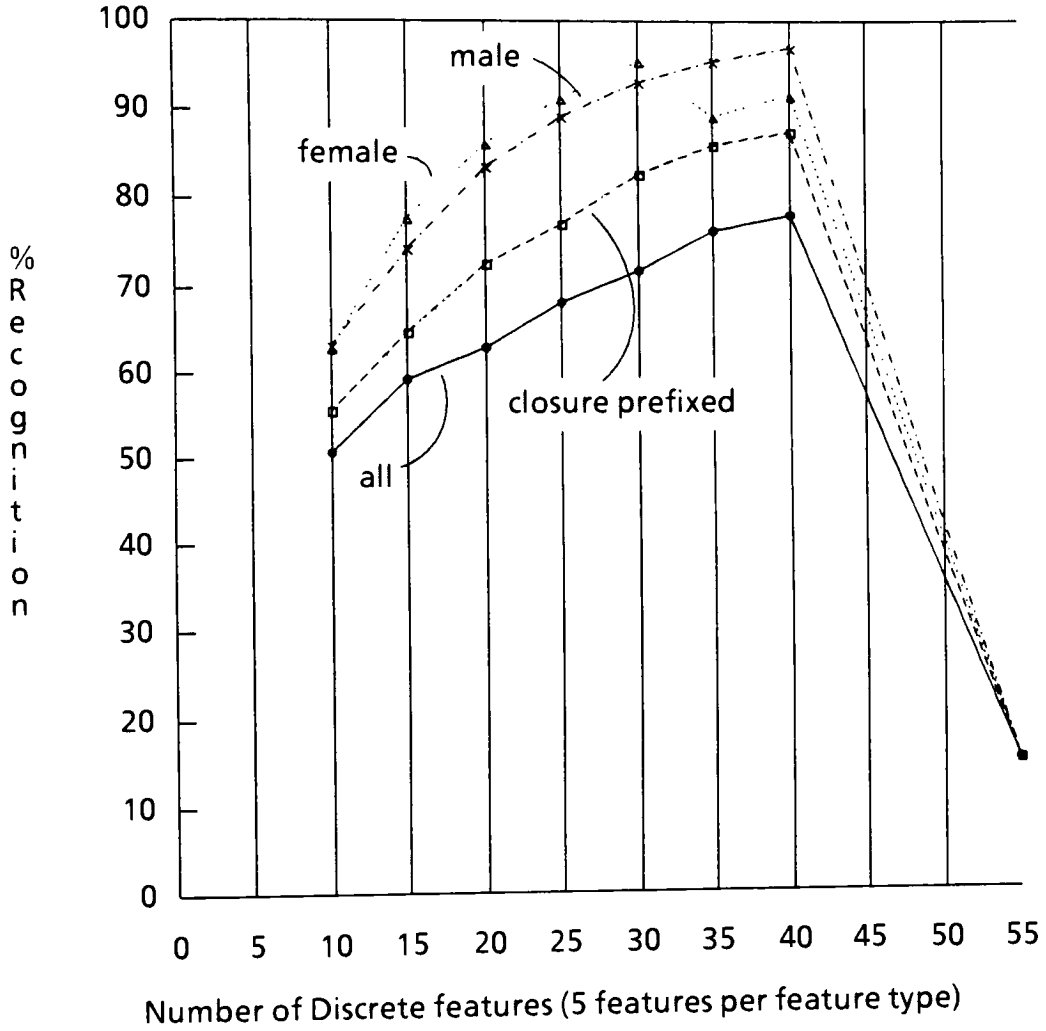


Figure 4-2 Percent recognition versus the total number of features for the optimum feature subset.

Train and Test results

The final set of experiments split various data groups to attempt classification through training. The three data groups used were closure prefixed, closure prefixed males and closure prefixed females. The feature types used were those that achieved the highest recognition results for each group and are shown in the 'Feature Types Used' column of Table 4-4.

For closure prefixed stops the data set was split in the following ways; even and odd, male and female, male even and male odd. Two sets of feature types were

Feature Types Used	closure even (476)		closure odd (476)	
	train	test	train	test
(base) mean(x,y), skew(x,y), var(x,y)	90.7	57.5	89.7	55.1
(ext) mean(x,y), mad(x,y), skew(x,y), var(x,y)	96.5	65.6	96.1 g	---
	males (539)		females (413)	
	train	test	train	test
(base) mean(x,y), skew(x,y), var(x,y)	90.0	50.3	95.2	46.7
(ext) mean(x,y), mad(x,y), skew(x,y), var(x,y)	97.0	53.1 g	100 g	54.0
	males even		males odd	
	train	test	train	test
(base) mean(x,y), skew(x,y), vary	90.6 g	53.0	90.4	55.7 g
(ext) mean(x,y), var(x,y), skewy	---	55.2	91.1	---

Table 4-4 Recognition results for testing across data sets for closure prefixed stop consonants. The training data included 4 frames plus left context using a 15 msec (192-point) data window and 7.5 msec shift.

applied to all data groups, the base set and extended set (ext). The features used for training and testing are listed on the left side of Table 4-4. For each data group, testing was done using the training statistics from the adjacent group. For example, closure prefixed even training were used to test the closure prefixed odd data group; 90.7 train, 55.1 test. All results are shown in Table 4-4. In general, the recognition rates were low.

Chapter 5
Summary and Conclusion

Differences between the database for this work versus prior work were that databases of prior work consisted of simple syllables, discrete words, or discrete words extracted from a fixed carrier phrase. The database for this work consisted of phonetically marked stop consonants in a multi-speaker continuous speech database. This database represents the definitive problem of stop consonant recognition in conversational English. See Appendix B for a complete discription of the database.

The data format for this work differs from prior work by context. The surrounding phonetic environment, particularly left context, is usually measured relative to burst release. In some cases left context is taken from burst release backward some number of milliseconds, e.g., 5 or 10 msec. Others may include data prior to burst release as part of the first frame. The main problem with those approaches were that they attempt to classify information from the stop closure. With high quality sampled data (low levels of background noise in the closure region), the most one could expect to achieve would be to classify voicing mode.

For this work, left context was taken to be the last 10 msec of the phonetic segment preceding the closure (Figure 3-2, p 3-5). It was believed that this context would carry statistical information relating phonetic groups. No discrete tests were done to evaluate this issue.

The preliminary testing to ascertain signal processing characteristics concluded that there were minimal classification differences between parameters using energy versus frequency power spectra. Another result of this testing showed that the

running cosine windowing filter (modified Hamming) outperformed the more traditional windowing filters, including no filter at all, consistently by 2 to 3 percent. It is theorized that there is a high degree of sensitivity with the amount of leakage into neighboring frequency bins. The original goal was to use the Cosine windoing filter for the burst release and the standard Hamming filter after the release. Although it was not tested, this may improve the separability fo latter frames.

In discrete feature performance, results indicated that the frequency features were generally better than those based on energy. However, with feature optimization using dynamic programming, features did not always perform in sets the way they did discretely. As an illustration of the contradiction between the two analytical methods (discrete versus grouped features), Table 5-1 shows a rank ordering of performance for discrete and optimization feature analysis. Although

Rank order	Discrete feature	Optimized Subsets
1	MADy	vary
2	skewx	skewy,meanx
3	meanx, vary	meany
4	varx	skewx
5	meany	varx
6	MADx	MADx
7	median	MADy,kurt (x,y),medi
8	skewy	
9	kurtx	
10	kurty	

Table 5-1. Rank ordering of Discrete features and optimized subsets. Both techniques consider 1,2,3,4 and left context frames to represent the feature type. The feature analysis used a 15 msec data window and 7.5 msec shift.

most features did correlate within the top 6, the feature MADy went from the highest ranking for discrete feature analysis to the lowest for optimized feature

subsets. Also, skewy went from eighth for discrete feature analysis to second for optimized feature subsets. The subset optimization technique shows that discrete feature performance does not indicate how well the feature will perform with other feature types. This notion was also eluded to by Edwards [EDWA81]. Edwards concluded in his discrete feature analysis that "...feature interdependence would be required for accurate recognition".

Also derived from the optimization tests was Table 5-2. This table shows the totals for the number of times a feature type was combined as the optimum into an optimized subset. The data groups are All, closure prefixed, closure prefixed females, and closure prefixed males. The training, discussed earlier, included 4 frames plus left context using a 15 msec (192-point) data window. To highlight certain features, a threshold was set to 4 optimum subset combines. If a feature type optimally combined 3 or less times it was not documented in Table 5-2. Considering

threshold = 4	meany	meanx	mady	madx	skewy	skewx	kurty	kurtx	vary	varx	medi
all		**9**	**7**		**5**				**9**		**8**
closure prefixed	**8**	**9**			**9**	**7**			**10**	**5**	
female	**9**	**6**			**7**	**7**			**8**		
male		**8**			**7**	**4**			**8**	**9**	**4**

Table 5-2 Summary of the number of times a feature type was combined as the optimum in an optimized subset. The training included 4 frames plus left context using a 15 msec (192-point) data window and 7.5 msec shift.

that MAD and variance are both measures of diffuseness and that median is another measure of central frequency, a pattern emerges from Table 5-2. This pattern

(shaded areas) shows that spectral centers, diffuseness, and tilts for both frequency and energy are key parameters for recognition of stop consonants.

The subset optimization for the full set of mixed features (11 types) clearly indicated robust classification for stop consonants knowing gender (pg 4-9, fig 4-2). The results for 30 features (6 feature types) were 95.6% and 93.1% for females and males respectively. For 40 features (8 feature types) the results for females and males was 100% less /g/ and 97% respectively. In the 40 features result for females, there were only 35 /g/ tokens for classification. Consequently, the classifier deleted that class. However, for the remaining tokens, 378 of /b,d,p,t,k/, 100% recognition was achieved.

Features which did not correlate with the findings of Forrest et al. [FORR87] were variance and kurtosis in frequency. Their analysis showed that variance did not add to stop discrimination when observed discretely. For this work, discrete analysis of variance in frequency ranked second and for optimized subsets ranked fifth. Forrest found that kurtosis was distinct for velar stops. Yet this work found that kurtosis ranked last for both discrete and optimized subset evaluations. There is good reason for both these differences. The first is that the database used for Forrest's work consisted of discrete words extracted from a fixed carrier phrase. It is well established that characteristics for isolated words are different than for the same words in continuous speech. The carrier phrase fixes the phonetic context for words used in their study. The second point to make is that Forrest et al. only evaluated voiceless stop consonants. This study included both voiced and unvoiced stops.

The final issue concerns the discrete evaluation of features. The present work has shown that discrete feature evaluation does not always correlate with feature subset evaluations.

Training and Testing

The results using half the data base for training and half for testing produced lower than expected performance. Particularly when splitting the males data set; 90% training and 53% testing (268 tokens per group).

Because of the increased performance earlier for knowing gender, it was not surprising to see low scores for the cross testing of closure prefixed tokens and males versus females. What was surprising was the low scores for the split group of males. This set was expected to achieve higher results. It is believed that training inadequately represented classes. For example, because of the difficulty in distinguishing /k/'s for varying phonetic context, a possible approach to increase training and testing results would be to subdivide the class structure. Class types may include a category for /k/ preceding front vowels versus back vowels. Finer detail may also be warranted for other stop consonant classes as well.

Thesis Bibliographies:

[AHD81] American Heritage Dictionary: New College Edition, Houghton Mifflin Company, **1981**.

[BLUM79] Blumstein S.E., Stevens K.N., "Acoustic Invariance in Speech Production: Evidence From Measurements of the Spectral Characteristics of Stop Consonants", J. Acoust Soc of America, Vol 66 #4, Jan **1979**.

[BLUM82] Blumstein S.E., Isaacs, E., Mertus, J., "The Role of The Gross Spectral Shape as a Perceptual Cue to Place of Articulation in Initial Stop Consonants", J. Acoust Soc of America, Vol 72 #1, July **1982**, pg43-50.

[BUSH83] Bush, M.A., Kopec, G.E., Zue, V.W, "Selecting Acoustic Features for Stop Consonant Identification", ICASSP, **1983**, pg742-745.

[COLE74] Cole, Scott, "The Phantom in the Phoneme: Invariant Cues for Stop Consonant", Perception & Psychophysics, vol15, #1, **1974**, 101-107.

[DEMI83] Demichelis, P., DeMori, R.,Laface, P., O'Kane, M., "Computer Recognition of Plosive Sounds Using Contextual Information", IEEE Trans on Acoustics, Speech, and Signal Processing, Vol ASSP-31 #2, April **1983**, pg359-377.

[EDWA81] Edwards, T.J., "Multiple features analysis of intervocalic English plosives[(a)] ", J. Acoust Soc of America, Vol 69 #2, **Feb 1981**, pg535-547.

[FANT70] Fant, G., "Acoustic Theory of Speech Production", Mouton, The Hague, 1970, 2nd printing.

[FORR87] Forrest, K., Weismer, G., Milenkovic, P.M., and Dougall, R.N., "Statistical Analysis of word-initial voiceless obstruents: preliminary data", Submitted to J. Acoust Soc of America, **1987**.

[GLEA61] Gleason, H, "An Introduction to Descriptive Linguistics", Holt, Rinehart, & Winstopn, New York, **1961**.

[HALE57] Halle M., Hughes G.W., & Radley J., "Acoustic Properties of Stop Consonants", J. Acoust Soc of America, Vol 29 #1, **Jan 1957**.

[KEWL81] Kewley-Port D., "Representations of Spectral Change as Cues to Place of Articulation in Stop Consonants", PhD Thesis City U of N.Y., **1981**.

[KEWL83[a]] Kewley-Port D., "Time-varing features as correlates of place of articulation in stop consonants", J. Acoust Soc of America, Vol 73, #1, Jan **1983**, pg322-335.

[KEWL83[b]] Kewley-Port D., Pisoni, D.B., Studdert-Kennedy, M., "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants", J. Acoust Soc of America, Vol 73, #5, **May 1983**, pg1779-1793.

[KLAT76] Klatt D.H., "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence", J. Acoust Soc of America, Vol 59 #5, Jan 1976.

[KLAT80] Klatt, D.,"Software for a Cascade/Paralled Formant Synthesizer", J. Acoust. Soc. Am. 67, 1980, pg 971-995.

[KOPE84] Kopec, G.E., "Voiceless Stop Consonant Identification Using LPC Spectra", ICASSP, 42.1.1, 1984.

[LEUN88] Leung, H.C., Zue, V.W., "Recognition fo Vowels Using Artificial Neural Networks",Meeting of the Acoust Soc of America, 1988.

[LISK64] Lisker, L and Abramson, A.S., "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements", Word, 20, 1964.

[MINE78] Mines, M.A., Hansen, B.F., Shoup, J.E., "Frequency of occurrence of phonemes in conversational English", Lang. Speech, vol21, 1978, pg221-241.

[MINI73] Minifie, F.D., "Speech Acoustics", in Minifie et al., (eds.), "Normal Aspects of Speech, Hearing and Language", Englewood Cliffs, Prentice Hall, 1973, pg235-284.

[PARS87] Parsons T.W., "Voice and Speech Processing", MacGraw Hill 1987, pg119-123.

[PRES88] Press, W.H., Flannery, B.P., Tevkolsky, S.A., Vetterling, W.T.," Numerical Recipes in 'C'; The art of Scientific Computing", Cambridge University Press, 1988.

[ROSN88] Rosen, M.L., Niles, L.T., Tajchman, G.N., Bush, M.A., Anderson, J.A., Blumstein, S.E., "A Connectionist Model for CV Syllable Recognition", ICASSP, vol 1, April 1988, pg59-62.

[SEAR79] Searle, C.L., Jacobson, J.Z., and Rayment, S.G., "Stop Consonant Discrimination Based on Human Audition", J. Acoust Soc of America, Vol65, #3, Mar 1979, pg799-809.

[STEV73] Stevens K.N., "The Potential Role of Property Detectors in the Perception of Consonants", Res. Lab of Elect. Eng., MIT., approx 1973.

[STEV78] Stevens K.N., Blumstein S.E., "Invariant Cues for Place of Articulation in Stop Consonants", J. Acoust Soc of America, Vol 64, #5, Nov 1978, pg1358-1368.

[STEV81] Stevens K.N., "Perspectives on the Study of Speech", Erlbaum, Hillsdale, NJ, 1981.

[SYRD86] Syrdal,A.K., and Gopal, H.S.,"A Perceptual Model of Vowel Recognition Based on Auditory Representation of American English Vowels", J. Acoust Soc of America, Vol 79, 1986 , pg1086-1100.

[WALL83] Walley, A., and Carrell, T., "Onset spectra and formant transitions in the adult's perception

of place of articulation in stop consonants", J. Acoust Soc of America, Vol 73, #3, Mar **1983**, pg1011-1022.

[YODE87] Yoder, S.K., and Jamieson, L.H., "Speaker-Independent Recognition of Stop Consonants", ICASSP Proceedings, **1987**.

[YODE86] Yoder, S.K., and Jamieson, L.H., "Accurate Recognition of Stop Consonants", Purdue Univ. Tech. Rep., TR-EE 86-42, **1986**.

[ZUE85] Zue, V., "Speech Spectrogram Reading Course", MIT, July **1985**.

Software tools and Stop Analysis Program (SAP)

1.) Manual Pages for SAP Program

2.) Supporting Software Tools

3.) Example Execution of SAP

Manual Pages for SAP

(Stop Analysis Program)

The parameter options associated with SAP are described below. The 'on/off' options respond in a toggle mode.

Command line: ***sap*** ***infile*** [arguments]

'***sap***' is the name of the executable program.

'***infile***' is the token name to be analyzed. The token name consists of the speaker identification, utterance identification, stop type, and the indexed suffix for the stop number of that utterance, e.g., fdmv4-1.t-h.3. In the example, fdm is a female speaker with initials 'dm', v4-1 is the utterance, t-h is the stop type, and 3 means it is the third stop token in that utterance (see the following section, Supporting Software Tools, for further discussion).

'arguments' are described below;

-#[c] The '#' is the integer number of frames to be analyzed. The 'c' option defines whether or not to include the context frames in the analysis data. The default takes the left and right context frames and the entire 'stop'

Default [all]

-c If the token includes a **c**losure segment, process the rms energy and the zero crossings per unit time. Default [off]

-d Data preemphasis by differencing. The differencing used here is $\mathbf{x}_{i+1} - x_i$.

Default [off]

-d[#] The number option following 'd' is input data ***d***ivision by the integer '#'. This has the effect of scaling the input data. Default [1]

-f# FFT array size (should be a power of 2). If a power of 2 is not entered, SAP will take the next higher power of 2 above the entered value. Default [512]

-fc# Feature compression type (0-6). Compression schemes are discussed at the end of this section. Zero '0' is no compression. Default [0]

-h Print *h*eader information. The header information consists of explicitly stating those parameters chosen. Default [on]

-k[fs] Store (keep) features and/or *s*pectral data. Default [infile.[fs]]

-l Take log base 10 of the power spectrum. Default [on]

-l[c] If context analysis is requested, take the left *c*ontext frame only and ignore the token stop time. In this case, if no value is specified for the number of frames, SAP will continue to process frames until it reaches the end of file. Default [off]

-np# Number of poles for linear prediction (LP) algorithm. Default [14]

-os[str] The output spectral data will be saved in file name 'str'. Default [infile.s]

-**of[str]** The output feature data will be saved in file name 'str'. Default [infile.f]

-p Take the power spectrum of the data. Default [on]

-p[f] Print the feature data to standard output. Default [off]

-p[n] Perform *p*ower spectrum *n*ormalization. When normalization is chosen, the feature 'mean' is eliminated. Default [off]

-s[#] Shift size in data points. (64 points = 5msec @ 12800sps) Default [64]

-s[fl] Spectral analysis technique used, Fourier or LP. Default [Nothing]

-sw# SAP includes an algorithm for smoothing the power spectrum. The smoothing program requires a data window in which to smooth over. This window does not have to be an integer value. '0' performs no smoothing. Default [0]

-w[#] The number of elements in the data window. The default FFT window size is 512 points. If the data window is less than the FFT window, the remaining points are padded with zeros. If the data window size exceeds the FFT window, SAP will take the larger of the two as the FFT window size. If the maximum value is not an integer value of 2, SAP will index the value to the next power of 2.

(128 points = 10msec @ 12800sps) Default [128]

-w[chw] The data ***w***indow filtering type, ***C***osine, ***H***amming, ***W***elch, or Square. Default [Square]

-w[n] Data ***w***indow normalization. This option normalizes the area for all window filtering types. Default [off]

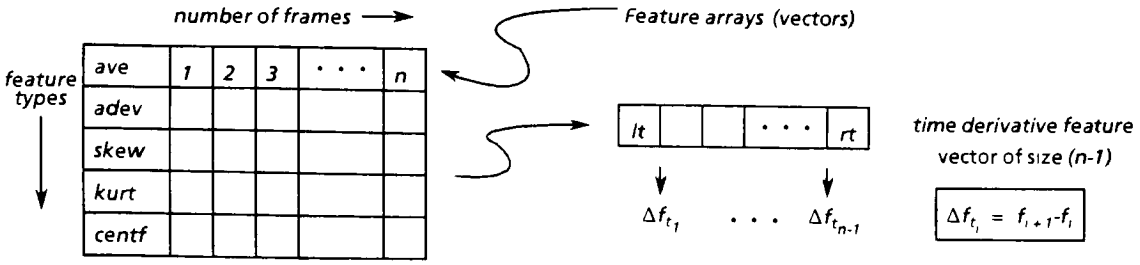
-x Existence of a label (time) file. If SAP is to analyze a binary data file at its starting point and it does not have a label file, this option must be included. If a label file does not exist, SAP will exit with a usage message suggesting this option. The default assumes a label file. Default [yes]

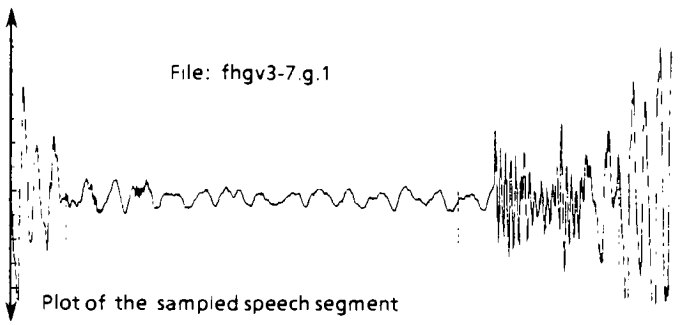
A condensed version of the 'option' information is contained in a usage file and is printed to standard output if the SAP program is executed incorrectly, e.g., without an 'infile' name.

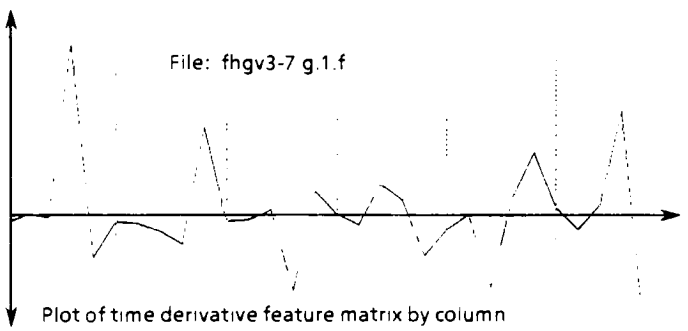
Feature Compression Schemes:

- **0** No feature compression applied.
- **1** Applies the features to the feature vectors, excluding the context frames. The reasoning for 1and 2 is that each feature vector could be analyzed in the same fashion as the spectra (similar in concept to cepstral analysis).
- **2** Applies the features to the feature vectors, including the context frames.

- **3** Applies the features to the concantenation the feature vectors, excluding the context frames. Here, the total number of features for classification is simply feature set.
- **4** Applies the features to the concantenation the feature vectors, including the context frames.
- **5** Sequential difference output. This technique assumes the information content of the spectral signature to be in the difference of feature values and not so much in the discrete feature values. By taking the difference between sequential frames, we are essentially looking at the time derivative feature matrix. The ultimate reduction of features is simple one less than the number of frames times the number of feature types. Classification can be viewed by either the row or column stream of features. An illustration of sequential differencing of the feature matrix is shown on the following page.







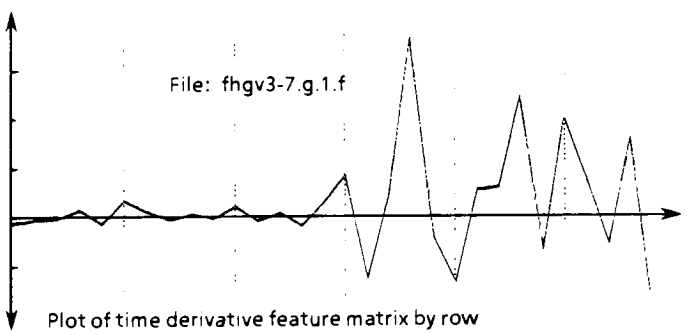


Illustration of sequential differencing of the feature matrix.

Supporting Software Tools

When executing SAP, the first argument is a token descriptor. A descriptor consists of the following; gender type, individuals initials, primary utterance, stop consonant type, and index. For example, a female with initials "md" talking vowel dense utterance category "v4" number "5" generates a /t/ phoneme that is the 6th stop consonant in the utterance. The token descriptor would be described as fmdv4-5.t.6. The descriptors are generated from the CMU utterance label files using additional software tools.

The CMU label files (one label file per utterance) contain the start and stop times and a type description for each phonetic segment in the utterance. These files are screened to generate new "stop" marker file sets. The new marker file contains the directory path to the data file (binary data) on the first line followed by a list of phonetically labeled stop consonants. In cases where the stop consonant is prefixed by its closure segment, the two are combined and the stop type labeled according to the latter segment. For example, if two consecutive segments in the original marker file were marked as;

Start time	Stop time	Type
560	640	t-cl
640	711	t-h

the new marker file would condense these two into one phonetic segment as follows:

Start time	Stop time	Segmentation time	Type
560	711	640	t-h

The segmentation time is zero if no closure is preceding.

The conversion file for this process is "stop_labels" (the source file is stop_labels.c). "stop_labels" should be executed in the directory where the new marker files will reside. To handle the volume of marker files in the database, a set of scripts were set up hierarchically to make all new marker files. These are called "run" files and would generate a new marker file directory; those for vowel dense, stop dense and individual talker. Additional scripts (shell and awk filters) were written to group classes, i.e., b, d, g, p, t, k, separate closure prefixed from non closure prefixed, and generate statistics for each class, e.g, min, max and average stop consonant duration time. These files and there function are listed below:

runv - Generates a new stop marker file set for the *vowel* database,

runs - Generates a new stop marker file set for the *stop* database,

run[talker initials] Generates a new stop marker file for the talker.

Additional awk filter scripts were written to:

Group classes, i.e., b,d,g,p,t,k, given the marker files,

- Separate closure prefixed from non closure prefixed,
- Class Statistics;
  - min stop duration,
  - max stop duration,
  - average stop duration,
  - min non closure prefixed stop duration,
  - max non closure prefixed stop duration,
  - average non closure prefixed stop duration,

  number of stop with a duration greater than a set threshold,

  number of stop with a duration less than a set threshold.

A final set of script files were written to automate batch execution of SAP. These scripts would prompt the user for input and generate the appropriate experimental

executable files. The script was structured to take in the class_[bdgptk] files in the directory where the script was executed. The output files were interactively defined with the user. The executable script is called bat.xset (batch experimental set). It was not unusual to generate a batch file to execute hundreds of runs of SAP

Example Execution of SAP

A sample execution of SAP would have a command line with the following;

sap flss5-4.t.4 -4c -pf -sf -wc -w192 -s96 -sw0 -lc -pn

SAP first searches the directory with the abbreviated marker files for flss5-4.t.4. If the file exists, the first line contains the path where the binary utterance resides. It then reads down the list of stop tokens until it reaches the sixth one. It gets the start, stop and segmentation times. If the binary file is found, the above arguments specify sap to :

- analyze 4 frames plus context frames,
- print feature data to stdout,
- spectral analysis type is FFT, default 512 point,
- data window filter is Cosine,
- data window is 192 points (15 msec for RIT data),
- shift size is 96 points (7.5 msec for RIT data),
- no power spectrum smoothing,
- analyze left context frame and ignore stop consonant ending time,
- power spectrum normalization.

The output is a header and a set of feature values for each frame. Below is an actual execution.

Input file name = flss5-4.t.4 1147.189
FFT window size = 512
Data window size = 192 data points (15.00 msec)
Shift size = 96 data points (7.50 msec)
Class type = 6(t)
Cosine window
Power spectrum normalization
Left context only

#	centf	adev	skew	kurt	var	median
lt	2224.005	1555.16	0.66	-0.78	3335931.25	3.12
1	5475.070	790.99	-1.65	2.27	1068661.00	1.33
2	5061.588	1001.59	-1.11	0.59	1525030.38	2.33
3	4936.837	1013.73	-0.93	0.12	1534661.88	2.62
4	4924.055	1034.64	-0.85	0.00	1557623.00	2.76

Database Information

Speech signals were taken from the Carnegie Mellon University (CMU) continuous speech database. The vowel database (approximately 300 tokens) was generated by four male and three female speakers uttering ten sentences each. The stop dense section of the database was generated using ten male and eight female speakers uttering 10 sentences each. The waveforms were originally sampled at 16 kHz with 12 bits of amplitude resolution, and subsequently lowpass filtered at 6 kHz and downsampled to 12.8 kHz. The complete list of utterances from both databases is listed below.

V1 Vowel Sentences
1. They toiled in the fields all day long.
2. The angry crowd pushed open the doors.
3. He left them with a reason to believe in themselves.
4. The auctioneer accepted the bid.
5. The yellow rose is the most beautiful of all flowers.
6. The child lured the rabbit into the cage.
7. Put the damp towel over your head for protection.
8. Always look before you leap.
9. While you were away, we opened the package.
10. The acrobat walked the tightrope.

V2 Vowel Sentences
1. He bought a new clock at the Tick-Tock Shop.
2. She has a twenty percent hearing loss.
3. She allowed the boy to eat the cookie.
4. The old hound was unenthused at the sight of the cat.
5. The owl swooped down upon the mouse.
6. The photograph proved he was guilty.
7. A youth has many lessons to learn.
8. If it never rained, we'd never grow.
9. Where in the world is the Fountain of Youth?
10. The handsome wool jacket was an oxford gray.

V3 Vowel Sentences
1. A cooked yam is a tasty sweet potato.
2. He recorded a new album with his younger partner.
3. Take Cloey to the show.
4. We fell for it hook, line, and sinker.
5. She won a blue ribbon at the county fair.
6. He was covered with soot from head to foot.
7. Are you aware of the good things in life?
8. Outside, the nights are only colder.
9. The old woman rocked away the hours.
10. He had a deep gouge over his left eye.

V4 Vowel Sentences
1. No one aroused his curiosity like Eunice.
2. Annoying a wild boar is insane.
3. Get out before it's too late.
4. The girl had a collection of wooden dolls.
5. One is the loniest number.
6. The bull chased the clown from the arena.
7. The thirsty girl rehearsed her lines.
8. The little pooch wagged his tail.
9. The hoodlum was full of malice.
10. Why not make a white oak chair?

V5 Vowel Sentences
1. You rang?
2. A loud alarm can be an eye-opener.
3. We arranged to look at the young animal.
4. They stashed the loot in the pumpkin patch.
5. Toast and jam tastes good for breakfast.
6. The robot was programmed to clean house.
7. The sauerkraut boiled till it burned.
8. I am amused at the cowboy's style.
9. Awhile ago, we knew very little.
10. Try to remember the joyous occasions.

V6 Vowel Sentences
1. Is the piano easier than the violin?
2. Pay a fee to see a play on broadway.
3. The amoeba either either eats or procreates.
4. Let's fry oodles of Thai eggrolls every hour.
5. Julia used to enjoy a coy admiral she met while on a voyage.
6. The IOU was due yesterday at eight.
7. Why undo all my output?
8. How aimlessly you eyed the toy engine.
9. Go iceskating and build a bon fire below all the swaying trees.
10. A new antenna actually offers better reception.

V7 Vowel Sentences
1. I ought to show only the area of the slow oil leak.
2. She grew ochre over yonder.
3. Do we throw out the ointment now or later?
4. No aching back ever kept a compulsive buyer from enjoying a sale.
5. However embarrassed you are, bow out gracefully.
6. They were the worst raw oysters they ever ate.
7. Roy ought to outlaw odious characters.
8. Mia is making a fondue in her new Amana oven.
9. Practice the minuet during the piano hour.
10. Paramesia oozed by our microscope lens.

V8 Vowel Sentences
1. Go away after the orchestra gets through Aida.
2. Raw eggs in the afternoon make Dottie ill.
3. My only employee is Eddie Edwards.
4. Harriet winked a coy eye at Theo.
5. My auto is OK in town.
6. My extremely oily hair now appalls me.
7. Throw algebra out or reevaluate your curriculum.
8. May I ask where the show-off put my only extra umbrella.
9. They allow only aliens to enter on occasion.
10. Anna offered to auction off the olive afghan.

V9 Vowel Sentences
1. Let's draw our audience into our laps.
2. How easy our essay ought to be again!
3. The crying boy ought to allow all the plowing to be done by Roy Amos.
4. Say 'ego' again.
5. Try iodine on the boy over there.
6. How any agent can stay idle is beyond me.
7. Put the dough out to rise before the pizza orders flow in.
8. Use the gray oil can to fix the axle.
9. Throw any old socks away over there.
10. How about an allowance every other week?

V10 Vowel Sentences
1. Clay oozed from the eighty eyelets.
2. Everyone prefers geometry over algebra any old day.
3. Being dieters, they ate their ochra often.
4. Plow Anderson's field only after Asa offers you all the aid you need.
5. The youngsters enjoy every hour of drama
6. Kalua and cream made Lisa and Maria awfully awkward.
7. Hiawatha was so pious, it nauseated him.
8. Do you know how I saw Italy in three hours?
9. He ought to let the pup gnaw on the raw apple.
10. I employ only eighty umpires throughout the year.

S1 Stop Sentences
1. Bob's date caught the cake before the gull did.
2. Put the dew drops on the green grape or bleed to death.
3. The trick is to twist the tip till it comes off cleanly.
4. Mark Twain drew quaint pictures of Colorado bluffs.
5. The big creep crawled past the ill dog.
6. Dots and globs trot by in dreams of gooey pizzas.
7. Get the paint out of the truck and place it on the tray.
8. The klutz grew bored of breaking clay bricks on the ground.
9. Please press the tucks of the ten gauze dresses.
10. The complicated plot brought grunts from my brother's geese.
11. Grip the tweed blazer and kiss the dumb dancing queen.
12. Tea dripped on the drum and impaired its deep quality.
13. The key is to take the train in quest of better opportunity.
14. Twas the dove that had glossy plumes.
15. Talk of truth and prove the clue is but a glib gift.
16. The glum clods were prodded to print practice blocks.
17. Pop's cruddy boot is plugging the cool pool.

S2 Stop Sentences
1. A Goodyear atlas is the best type of book to help with homework.
2. This giant skull was discovered near that jawbone.
3. German chocolate and spun sugar are Steve's favorites.
4. Jim tried to go lunchless before his stomach surgery.
5. Bob babysat the pesky child and became incredibly perturbed.
6. Midge's deed led her into a web of intrigue.
7. Beechwood and Edgewood are both suburbs of Detroit.
8. Each jazz singer sang ten pop tunes and then sat down.
9. His spouse toiled over the doilies for days.
10. Please don't drop the plaque because it could break.

S3 Stop Sentences
1. Curt was gentle as he extracted one of my bicuspids.
2. Dirk choked on the dry T-bone steak.
3. The Dodgers beat the Tigers by a wide margin.
4. The bureau of statistics provides accurate figures.
5. Dee bought a pound of cheese for the kids' lunches.
6. Hogs and geese turned the old adobe church into a pig sty.
7. A shop in town sells all kinds of copper coins.
8. Joyce didn't want Boyd to know about her worst foibles.
9. Brutus plowed the field while Gabe planted the squash and potatoes.
10. Olga spilt milk on her special purple gown.

S4 Stop Sentences
1. The spotlight was on the display of spring fashions.
2. Scot spoke glibbly about high technology.
3. You have a choice between sukiyaki, chop suey, or chow mein.
4. The baby was a chubby, pudgy little butterball.
5. Gilbert, Tom, and Bill are all good at skating and diving.
6. The puny ape scratched his toe on the edge of his cage.
7. When that goop is gone, go to the stockroom for more.
8. Both Ken and Gwen go to ski school.
9. Is Peg capable of putting the bike back together?
10. Dr. Stenson will speak on standard cures for goiter.

S5 Stop Sentences
1. Spanish is the language spoken in Spain.
2. Stick a pin in the balloon and it will burst.
3. Edwin and Maud Cowper are identical twins.
4. The bad boy tipped over the cookie jar in the kitchen.
5. The etching was ineptly done on badly chosen paper.
6. Julie keeps tadpoles and guppies in a jar.
7. Did the supply package to Zimbabwe include scalpels?
8. Some tall, jowley guy stole the pie with the poison in it.
9. This cream will give you ageless skin at low cost.
10. The scout watched the sky for smoke signals.

S6 Stop Sentences
1. This poodle chases cats and chews on our best furniture.

2. The football coach always wore goggles.
3. A gust of wind made the wood chips tumble upwards.
4. Chester daubed the couch with cool-colored paint.
5. Everyone proved to be a spy, a spook, or a stool pigeon.
6. Pete became the grooviest piccolo player in Princeton.
7. James just doodles awkward sketches in art class.
8. Mr. Berg scolded the girl for dabbling in the paint.
9. Paul bought cowboy boots and chaps which cost a lot.
10. Joe gagged on the stout gravy.

S7 Stop Sentences

1. The top candidate was Bill Taylor.
2. Put the jam in the cupboard.
3. The dog chased the children into the cave.
4. Paul chose the juiciest peach from the bushel.
5. She coyly kissed her boyfriend.
6. A boa will not bite unless you bait him.
7. Don't touch the china in that case.
8. Her poise and charm caused comment.
9. Talcum powder can keep you cool in tropical climates.
10. Jane chugged her beer before passing out.

S8 Stop Sentences

1. The geyser in the desert gushed water.
2. Joyce's goiter took a turn for the worse in June.
3. Apple pie with cheddar cheese is delicious to eat.
4. He should choose his duties carefully.
5. Her bountiful goodness inspired us all.
6. The journalists from the Times interviewed Guy.
7. The gooey candy broke Glenn's tooth.
8. Tobey caught a tadpole and a toad down by the creek.
9. Bobby burned his elbow on the copper kettle of porridge.
10. Park the cab in the carport and pick up Jan's baggage.

S9 Stop Sentences

1. Joe stuffed the crepes with juicy berry filling.
2. Teach Peg how to cope with drug addicts.
3. Trouble awaited the boys in the deep, dark tunnel.
4. Pick and choose the best king crabs for dinner tonight.
5. Please don't beat the child for kicking the cat.
6. The plumber unclogged the drain for the innkeeper.
7. Clean the tub with a brush and some scouring powder.
8. The gardener kept the grounds looking superb.
9. Keep the tweed coat in Kirk's garment bag.
10. The trap door popped open knocking Betty down.

S10 Stop Sentences

1. Dip the gold chain in the jewelry cleaner.
2. The girl put grape jelly on her piece of pumpernickel rye.
3. Go down to the train station and get Charles.
4. Dave's pup was killed by the speeding car.
5. Please cook up a big batch of corn for Grandad.
6. John tried to rescue the drowning children.
7. Japanese teachers are better trained than their Chinese counterparts.
8. Jumbo burgers with lettuce and tomatoes are quite popular these days.
9. The cop gave Joe a ticket for jay-walking.
10. Gophers are tearing up Bob's garden again.

S11 Stop Sentences

1. This cold spell will destroy the barley that was just planted.
2. The toy jeep rolled under the bushes.
3. The geek bit the head off the duck.
4. It doesn't pay to buy cheap clothes.
5. Pooh Bear chatted with Tootsie about books and other such inconsequential things.
6. Do those guys standing against the bar look like trouble?
7. They bumped into that Buick and then tore down the street.
8. Jim had a good time downtown yesterday.
9. Take that chest back upstairs.

10. They enjoyed the hot clam chowder.