COMPUTER RECOGNITION OF PLOSIVES IN RUNNING SPEECH

B. Tang and C. Y. Suen
Department of Computer Science
Concordia University
1455 de Maisonneuve West
Montreal, Quebec
H3G 1M8

ABSTRACT

In this paper the authors propose a completely automatic speaker independent system to recognize all six stop consonants in continuous speech. This system makes use of three different distinctive features including formant transitions, silent interval and voice onset time. One hundred and twenty sentences from six English speakers were selected and tested by the system. The results confirm that no single feature can account for the distinction of voiced and unvoiced stop consonants. A comparison of the results of three different distinctive features has been made.

INTRODUCTION

In the past fifteen years, extensive study on phoneme recognition has been done by a number of researchers [1]-[27]. Typical phonemes that they have studied are English stop consonants /p, t, k, b, d, g/. The reason is that stop consonants occur very often in the English language [27]. At present, many models using acoustic features such as transitional cues, duration, silence and voicing, have been developed to recognize stops. But most of them require human assistance, and are only limited to isolated words. In this research, the authors propose an automatic speaker-independent system using three distinctive acoustic features to recognize English stops in running speech without human intervention: viz., Formant Transitions (F. T.), Silent Interval (S.I.), Voice-Onset-Time (V.O.T.).

Table 1 summarizes the results on the recognition of stop consonants and related experiments obtained by other researchers. Partial result of the present research is also included.

CSRS SCHEME

Figure 1 shows a diagram of the proposed continuous speech recognition system (CSRS). The function of this system is to detect and extract features from the input speech signals, and to use these features to recognize and classify the signals into different phoneme classes. The system is divided into four stages - digitization, preprocessing, feature extraction and classification. All signals must pass through these four stages before they can be identified.

The whole system, with the exception of the digitization system which is implemented in the INTEL 8085 micro computer, is programmed in the Fortran IV computer language and processed by a CDC Cyber-172 computer.

DATABASE

Data collected for this study include 160 unstressed and 160 stressed sentences uttered by eight untrained paid native speakers of English, four males and four females. They belong to the age group of twenty to forty. The recordings of sentences were made in two sessions in a 12 ft x 15 ft x 10 ft sound-proof room on two 1200-ft Scotch tapes at the speed of $7\frac{1}{2}$ ips. The recording system setup consists of one Sennheiser MD 421U dynamic cardiod microphone, one Tascam Model 10 mixer and one Ampex AG 440B tape recorder.

Speech Digitization

After the speech samples have been recorded, they are sent to the speech digitization system which consists of one Sony 2-track mono tape recorder, one band-pass filter, one 8-bit (256 levels) Analog/Digital (A/D) converter, one INTEL 8085 micro computer and one CDC Cyber-172 computer. The analog signal is converted into 256 digital levels at a sampling rate of 10 kHz to provide enough information for subsequent processing (Markel [2]). Once the speech signal has been digitized, it is transferred to the CDC Cyber-172 computer for storage and further treatment. All digitized signals are stored on two 2400-ft 1600 BPI magnetic tapes.

DATA PREPROCESSING

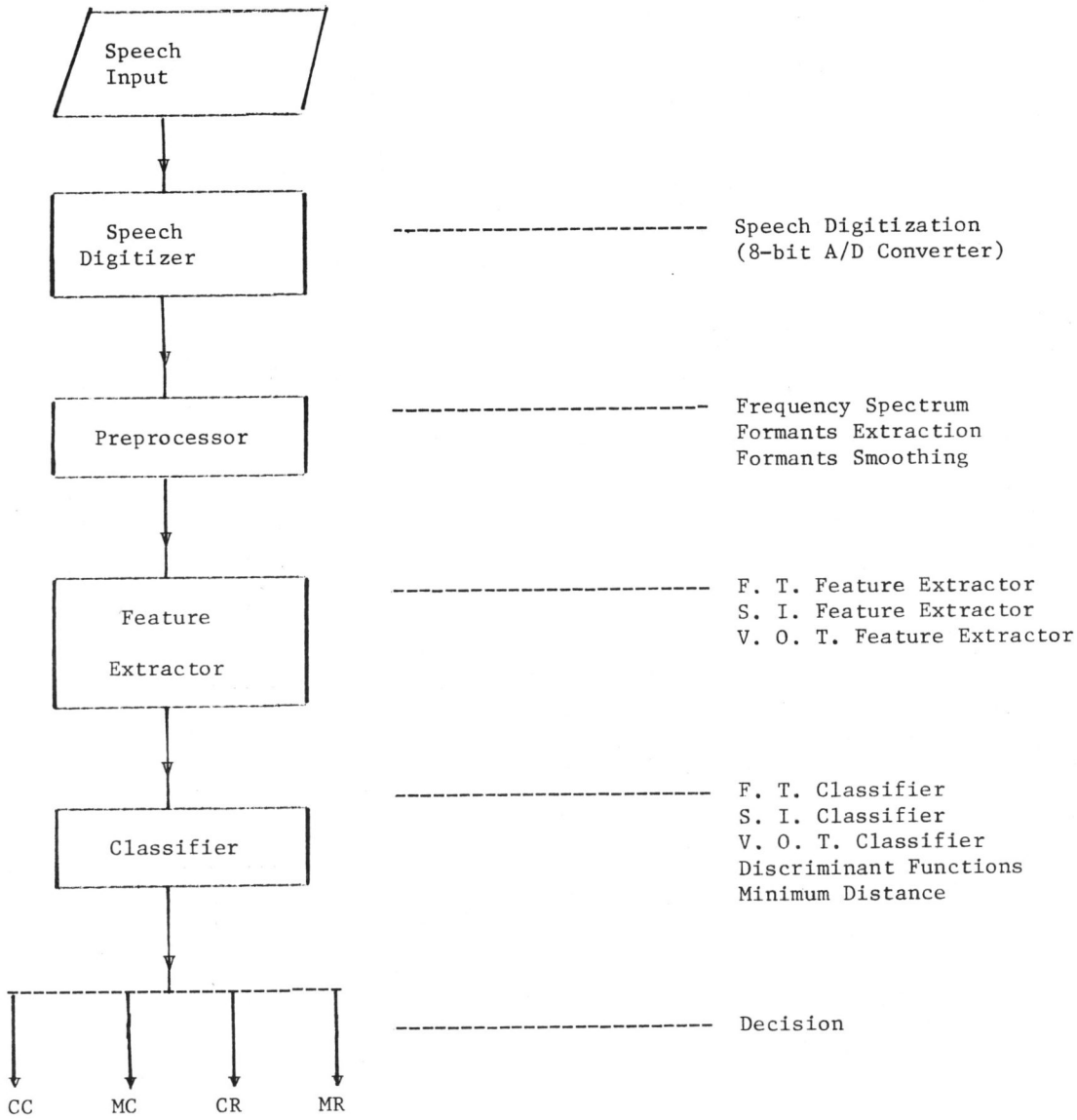
In this system, the preprocessing stage contains two phases, speech spectrum and formant extraction. Speech waveform is converted from the time domain into the frequency-domain using Discrete Fast Fourier Transform (DEFT). It identifies the frequency components at each 12.8 ms segment of the waveform. An automatic formant extraction

Table I-1. Summary of Stop Consonants Recognition and Related Experiments

Author	Date	Features Used	Features Extr. By	Decision Method	Materials Used	Recognition
Tang	1981	FT,SI & VOT	Automatic	Min. Dist.	Unconstrainted continuous speech spoken by 3 male 3 female speakers (120 sentences)	79.58% for stressed data (48 sentences) & 70.54% for unstressed data (72 sentences)
Santerre & Suen	1981	FT,SI,VD & VOT	Manual	NA	Isolated words spoken by 3 male & 3 female speakers	NA
Blumstein & Stevens	1980	VD	Manual	Human Perception	Synthesized initial stops in isolated syllables	Over 80%
Datta et al	1980	FT & VOT combined	Manual	Statistical techniques with parametric repesentation	Isolated words by 3 male speakers (600 CV words)	Max. 74.9%
Demichelis et al (Proposal)	1979	FT & VOT combined	Automatic	Fuzzy algorithm	Pseudo-syllables for voiced & CV syllables for unvoiced stops in continuous speech	NA
Port	1979	SI	Manual	Human Perception	Syllables in running speech	NA
Pal & Majumder	1978	VOT & FT	Manual	"	PB CVC context spoken by 5 male speakers (600 samples)	60% for dentals & 85% for bilabials
Searle et al	1978	VOT spectral peaks,shapes	Filter bank	Discrimin. analysis	Initial stops in isolated words (148 words)	77%
Wolf	1978	FT,SI,VD & VOT	Manual	Human Perception	Six repetitions syllables spoken by two male speakers	~68.72% combined
Lisker et al	1977	VOT	Manual	Manual	Synthesized CV syllables	NA
Molho	1976	Autocorrelation coef.	Automatic	Fuzzy algorithm	21 sentences spoken by 4 speakers	~50% for "p,t,k" & 66.7% for "d"
Itakura	1975	"	"	Minimum prediction residual	Designated male speaker via telephone input in isolated words (200 Japanese words)	97.3%
Weinstein et al	1975	FT & VOT	Manual	Ebst & Fbst analysis	Segmented phonemes in connected speech	76% for d,g,t,k

Table I-1. Summary of Stop Consonants Recognition and Related Experiments

Author	Date	Features Used	Features Extr. By	Decision Method	Materials Used	Recognition
Winitiz et al	1975	VOT	Manual	NA	CV monosyllables spoken by one male speaker	NA
Cole & Scott	1974	FT & VOT	Manual	NA	Tape spliced initial stop syllables	Average 93.83% & 75.83% with target vowels i,u respec.
Suen & Beddoes	1974	SI	Manual	NA	Word pairs spoken by 3 male & 3 female speakers	NA
Stevens & Klatt	1974	FT & VOT	Manual	Human Perception	Synthetic CV syllables	NA
Eimas & Corbit	1973	VOT	Manual	Human Perception	Synthetic speech	NA



Symbols: CC - Correctly Classified;
MC - Misclassified;
CR - Correctly Rejected;
MR - Misrejected.

Figure 1. The Structure of the Proposed CSR System

algorithm has been developed to extract the first two formants from the speech spectrum. The algorithm is based on peak picking. An interpolation technique has been applied to obtain smooth formants.

RECOGNITION ALGORITHMS AND FEATURES USED

Three different methods have been developed to recognize /p, t, k/ and /b, d, g/. They are 1) Formant Transitions (F.T.), 2) Silent Interval (S.I.) and 3) Voice-Onset-Time (V.O.T.). This system, in fact, contains three subsystems. Each of them works independently and has been designed to process one sentence at a time.

FORMANT TRANSITIONS

As pointed out by Cole and Scott [8], Datta, Ganguli and Ray [9], Menon, Rao and Thosar [10], Pal and Majumder [11], Santerre and Suen [12], Sharf and Hemyer [13], and Wolf [14] there is a rapid change in the shape of the vocal tract which makes the transition from one place of articulation to another when a stop consonant is uttered with a preceding or following vowel. They also conclude that the change in formant frequency (transitional cue) of the vowel associated with the stop consonant(s) may provide information general enough to distinguish voiced and unvoiced stops in most cases. This method simply computes the percentage change of the first two formants from the plosive release to the steady state of the associated vowel for preceding stops (including initial and medial stops), and from the steady state to the closure for final stops.

SILENT INTERVAL

As suggested by Cole and Scott [8], Liberman et al. [15], Lisker [16], Port [17], Santerre and Suen [12], Slis et al. [18][19], Suen et al [20], and Wolf [14], the S.I. is another important cue to distinguish voiced and unvoiced stop consonants. The S.I. is defined as the duration between closure and plosive release. The above authors conclude that voiced stop consonants usually have shorter duration of S.I. than unvoiced ones.

As noted in the previous method, there is usually a pause - silent period before the burst of a stop consonant. In this study, the S.I. is the second cue suggested to recognize medial and final stops.

VOICE-ONSET-TIME

V.O.T. is the third method used in this research. It is regarded as the primary cue to distinguish voiced and unvoiced stop consonants. Most researchers like Blumstein and Stevens [21], Eimas and Corbit [22], Lisker and Abramson [23], Lisker, Liberman and Erickson [24], Santerre and Suen [12], Stevens and Klatt [25], Winitz, LaRiviere and Herriman [26], and Wolf [14] conclude that the V.O.T. is usually shorter in voiced stop consonants than in unvoiced stop consonants. V.O.T. is defined as the time interval between the burst that marks the release of the stop closure and the onset of quasi-periodicity which reflects laryngeal vibration (Lisker and Abramson [27]).

This method measures the burst period, i.e. the time difference in the first two formants between 1) the starting point of plosive release and the starting point of the steady state for the preceding stops, 2) the starting point of plosive release and the end point of the burst for the final stops.

EXPERIMENTAL RESULTS

Two different sets of data (half of them were from training set) were prepared for testing the system. They were selected from the six speakers with highest recognition scores. The data were chosen equally in number from three male and three female speakers. One set contains seventy-two sentences and the other one contains forty-eight sentences. As mentioned earlier, the difference between these two sets of data is: in the first set, the sentences were uttered by the speakers in their usual way; in the second set, the sentences were uttered by the same speakers but they were requested to emphasize the stop consonants. The respective total number of preceding stop consonants and final stop consonants are 354 (including 138 medial stop consonants) and 228 in unstressed data, and 236 (including 92 medial stop consonants) and 152 in stressed data. Each set of data was tested individually by the system. The outcomes of the experiments consist of four types (Figure 1). The first type called correct classification (CC), which counts the phonemes correctly classified in the corresponding classes. The second type is called misclassification (MC), which is a count of the phonemes classified in the wrong classes. The third type called correct rejection giving a count of the phonemes not belonging to any classes of stop consonants. The fourth one called misrejection indicating the number of instances the system incorrectly rejects the phonemes which are actually stop consonants.

Based on different features used in the CSR system, six classification experiments have been conducted. The results of each experiment show the performance of each CSR subsystem, or each feature used, for each set of data. Tables 2a and 2b summarize the scores of correct rejection, misrejection and correct classification of the three distinctive features for all speakers

Spkr	F.T.			S.I.			V.O.T.		
	CR	MR	CC	CR	MR	CC	CR	MR	CC
1	75.71	19.48	74.15	43.64	15.91	52.53	45.61	5.48	50
2	65.52	15.49	70.54	29.79	4.17	46.48	41.86	4.48	46.67
3	75	17.19	70.97	15.22	3.85	33.33	42.59	7.58	50
4	52.70	2.94	65.49	31.48	11.11	38.27	46.03	4.76	50
5	61.67	20.31	68.55	50.98	21.74	55.41	41.18	7.94	50
6	75.47	23.61	73.60	60.47	17.24	58.33	41.86	1.64	46.15

Symbols: CC - Correctly Classified; MR - Misrejected; CR - Correctly Rejected

Table 2a. Summary of Recognition Scores (%) for All Features (Unstressed)

Spkr	F.T.			S.I.			V.O.T.		
	CR	MR	CC	CR	MR	CC	CR	MR	CC
1	88.76	22.73	79.35	37.50	7.69	43.97	46.97	1.14	52.55
3	85.19	28.77	77.92	58.97	12.28	55.56	46.15	1.33	52.14
5	95.77	32.81	80.71	52.54	6.52	56.19	48.44	2.99	56.49
6	91.03	25.71	80.41	56.92	7.89	58.25	45	1.30	54.74

Symbols: CC - Correctly Classified; MR - Misrejected; CR - Correctly Rejected

Table 2b. Summary of Recognition Scores (%) for All Features (Stressed)

Method	F.T.	S.I.	V.O.T.
CR	252/375 = 67.20%	114/296 = 38.51%	135/311 = 43.41%
MR	68/417 = 16.31%	22/173 = 12.72%	21/388 = 5.41%
CC	559/792 = 70.58%	223/469 = 47.55%	341/699 = 48.78%
MC	233/792 = 29.42%	246/469 = 52.45%	358/699 = 51.22%

Symbols: CC - Correctly Classified; MC - Misclassified; CR - Correctly Rejected; MR - Misrejected

Table 3a. Comparative Recognition Results of Three Distinctive Features (Unstressed)

Method	F.T.	S.I.	V.O.T.
CR	287/319 = 89.97%	138/266 = 51.88%	119/255 = 46.67%
MR	75/278 = 26.98%	17/193 = 8.81%	5/260 = 1.92%
CC	475/597 = 79.56%	245/459 = 53.38%	291/515 = 56.50%
MC	122/597 = 20.44%	214/459 = 46.62%	224/515 = 43.50%

Symbols: CC - Correctly Classified; MC - Misclassified; CR - Correctly Rejected; MR - Misrejected

Table 3b. Comparative Recognition Results of Three Distinctive Features (Stressed)

including both unstressed and stressed data.

Tables 3a and 3b show the overall performance of each recognition method. In unstressed data, the best recognition method is Formant Transitions which obtained 70.58% compared with the other two methods which obtained below 50%. In stressed data, the best method is also Formant Transitions which obtained 79.56% compared with 53.38% by the Silent Interval method and 56.50% by the Voice-Onset-Time method.

CONCLUSIONS

The performance of each method used in the present system is tabulated. The results show that no single feature alone can perfectly account for the distinction of a voiced or unvoiced stop consonant from the others. The formant transition cue produced the highest recognition score only with a priori knowledge of the target vowels. The other two cues are mainly time considerations. As noted in the previous chapter, low recognition rates are due to considerable variations of burst or silent period when the position of the word is different. On the other hand, some speakers ignored the pronunciation of some phonemes especially the final stops. As a result, some final stops cannot be detected. However, this can be compensated by using the transitional cues.

Conclusion can also be drawn that timing would not be an effective cue to measure the phonemes in continuous speech unless certain constraints can be imposed, such as the speed and intonation of the same word spoken in different positions, and the attention of speakers paid to the pronunciation of the final stops.

REFERENCES

[1]. Rabiner, L. R., Schafer, R. W., "Digital Processing of Speech Signals", Prentice-Hall Inc., p. 43, 1978.

[2]. Markel, J. D., "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation", IEEE Trans. Audio Electro Accoust., Vol. AU-20, pp. 129-137, June, 1977.

[3]. Demichelis, P., De Mori R., Laface, P. and O'Kane, M., "Computer Recognition of Stop Consonants", IEEE International Conference on Acoust., Speech and Signal Processing, pp. 85-88, 1979.

[4]. Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-23, pp. 67-72, Feb. 1975.

[5]. Molho, L., "Automatic Acoustic-Phonetic Analysis of Fricatives and Plosives", IEE Acoust. Speech Signal Process. Rec., pp. 182-185, April 1976.

[6]. Searle, C. L., Jacobson, J. Z. and Rayment, S. G., "Stop Consonant Discrimination Based on Human Audition", J. Acoust. Soc. Am., Vol. 65, pp. 799-809, March 1979.

[7]. Weinstein, C. J., McCandless, S. S., Mondshein, L. F. and Zue, V. W., "A System for Acoustic-Phonetic Analysis of Continuous Speech", IEEE Trans. Acoust., Speech and Signal Process., Vol. ASSP-23, pp. 314-327, Feb. 1975.

[8]. Cole, R. A. and Scott, B., "Toward a Theory of Speech Perception", Psychologic Review, Vol. 81, No. 4, pp. 348-374, 1974.

[9]. Datta, A. K., Ganguli, N. R. and Ray, S., "Recognition of Unaspirated Plosives - A Statistical Approach", IEEE Trans. on Acoust., Speech and Signal Process., Vol. ASSP-28, No. 1, pp. 85-91, Feb. 1980.

[10]. Menon, K. M. N., Rao, P. V. S. and Thosar, R. B., "Formant Transitions and Stop Consonant Perception in Syllables", Language and Speech 17, pp. 27-46, 1974.

[11]. Pal, S. K. and Majumder, D. D., "Effect of Fuzzification on the Plosive Cognition System", Int. J. Systems Sci., Vol. 9, No. 8, pp. 873-886, 1978.

[12]. Santerre, L. and Suen, C. Y., "Why Look for a Single Feature to Distinguish Stop Cognates?", Journal of Phonetics 9, pp. 163-174, 1981.

[13]. Sharf, D. J. and Hemeyer, T., "Identification of Place of Consonant Articulation from Vowel Formant Transition", J. Acoust. Soc. Am. Vol. 51, No. 2 (Part 2) pp. 652-658, 1972.

[14]. Wolf, C. G., "Voicing Cues in English Final Stops", J. of Phonetics, Vol. 6, 19, pp. 299-309, 1978.

[15]. Liberman, A., Harris, K. S., Eimas, P., Lisker, L. and Bastian, J., "An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance", Language and Speech, 4, p. 175, 1961.

[16]. Lisker, L., "Closure Duration and the Intervocalic Voiced-Unvoiceless Distinction in English", Language 33, pp. 42-49, 1957.

[17]. Port, R. F., "The Influence of Tempo on Stop Closure Duration as a Cue for Voicing and Place", J. of Phonetics, Vol. 7, pp. 45-56, 1979.

[18]. Slis, I. H. and Cohen, A., "On the Complex Regulating the Voiced-voiceless Distinction I", Language and Speech 12, p. 80, 1969.

[19]. Slis, I. H. and Cohen, A., "On the Complex Regulating the Voiced-voiceless Distinction II", Language and Speech 12, p. 137, 1969.

[20]. Suen, C. Y. and Beddoes, M. P., "The Silent Interval of Stop Consonants", Language and Speech 17, pp. 126-134, April-June, 1974.

[21]. Blumstein, S. E. and Stevens, K. N., "Perceptual Invariance and Onset Spectra for Stop Consonants in Different Vowel Environments", J. Acoust. Soc. Am., Vol.67, pp. 648-662, Feb. 1980.

[22]. Eimas, P. D. and Corbit, J. D., "Selective Adaptation of Linguistic Feature Detectors," Cognitive Psychology 4, pp. 99-109, 1973.

[23]. Lisker, L. and Abramson, A. S., "Some Effects of Context on Voice Onset Time in English Stops", Language and Speech 10, pp. 1-28, 1967.

[24]. Lisker, L., Liberman, A. M. and Erickson, D. M., "On Pushing the Voice-Onset-Time (VOT) Boundary About", Language and Speech 20, pp. 209-216, 1977.

[25]. Stevens, K. N. and Klatt, D. H., "Role of Formant Transitions in Voiced-Voiceless Distinction for Stops", J. Acoust. Soc. of Am., Vol. 55, No. 3, pp. 653-659, March 1974.

[26]. Winitz, H., LaRiviere, C. and Herriman, E., "Variations in VOT for English Initial Stops", J. of Phonetics, Vol. 3, pp. 41-52, 1975.

[27]. Suen, C. Y., "n-Gram Statistics for Natural Language Understanding and Text Processing" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAM-1, No. 2, pp. 164-172, April 1979.