Development and Preliminary Testing of a Computer-based Program for Training Stop Consonants

Mise au point et essai préliminaire d'un programme informatique pour la formation d'occlusives

Nancy Thomas-Stonell
Department of Speech-Language Pathology & Audiology
The Hugh MacMillan Rehabilitation Centre

Leslie Dolman
Designing Aids for Disabled Adults
Toronto, ON

Michael McClean
Graduate Department of Speech Pathology
The University of Toronto

Bruce Oddson
Department of Rehabilitation Engineering
The Hugh MacMillan Rehabilitation Centre

Key Words: computer, treatment, articulation, stop consonants

Abstract

Consonant imprecision is a major factor in reducing intelligibility in speech production disorders. This article describes the development of a computer-based speech training (CBST) program for training stop consonant productions. Called the "Bow and Arrow Game," this program provides visual feedback on both manner and place of articulation. Results of recognition accuracy reveal that the program is 91% accurate at discriminating the /p/ phoneme from the /t/ and /k/ sounds but less accurate at discriminating the /t/ and /k/ sounds (75% and 71%, respectively).

Résumé

L'imprécision consonantique constitue un facteur important dans la diminution de l'intelligibilité présente dans les troubles de production de la parole. Ce texte décrit la mise au point d'un programme informatique de formation de la parole (PIFP) en vue de la formation de productions d'occlusives. Ce programme, appelé le «Jeu de l'arc et des flèches,» fournit une réaction visuelle sur le mode et de lieu d'articulation. Les résultats de l'exactitude de la reconnaissance révèlent que le programme parvient dans une proportion de 91% à distinguer le phonème /p/ des sons /t/ et /k/ (75% et 71% respectivement.)

It is estimated that nearly 75% of all communication deficits are articulatory disorders (Newman, Creaghead, & Secord, 1985). More than ten percent of grade one children who are six years of age have some impairment of speech, language, or hearing. The majority of these difficulties are in the speech production area. Speech disorders are also prevalent in children with acquired neurological deficits, such as cerebral palsy, hearing impairment, and cleft lip/palate.

A major factor causing reduced speech intelligibility is consonant imprecision. Difficulties with tongue and lip placement result in the production of plosive sounds that are distorted or confused with other sounds (i.e., /t/ for /k/). Inability to completely occlude the vocal tract results in the production of fricative-like rather than plosive sounds. Lack of respiratory support decreases intraoral breath pressure. Both of these problems result in sound productions with a reduced or absent plosive burst.

Our research has found only one commercially available articulation training program that uses consonant recognition: Computer-Aided Speech Production and Training (CAPST). An evaluation of this program by Fitch (1989) revealed the following percent accuracy scores for recognition of plosive productions: /p/ = 22.7%; /t/ = 36.3%; and /k/ = 34.4%. Fitch concluded that the program had extremely limited application for articulation treatment.

Computer-based speech training (CBST) has the potential to improve the speech training process (Watson & Kewley-Port, 1989). Bernstein, Goldstein, and Mahshie (1988) reviewed several new CBST systems including: the Johns Hopkins Speech Training Aid (Ferguson, Bernstein, & Mahshie, 1988); the Indiana Speech Training Aid (ISTRA) (Kewley-Port, Watson, & Cromer, 1987); the Vowel Corrector (Povel & Wansink, 1986), and the SpeechViewer (IBM-France Speech Training Project) (International Business Machines, 1988). The Video Voice Speech Training System (MicroVideo Corp, 1985) and the Visible Speech Aid (VSA) (Dickson, Ingram, & Snell, 1984) also have been developed recently. Most of these systems provide feedback on the suprasegmental aspects of speech, such as pitch, loudness, and voicing. Several of these systems also provide feedback on the accuracy of vowel productions. Recent advances in technology now permit the development of computer-based consonant recognition systems.

This article describes the design and testing of a CBST software program for modifying stop consonant production called the "Bow and Arrow Game." The program was designed to be compatible with the IBM SpeechViewer system and meet the following requirements: (1) present visual and auditory feedback on the manner (strength and degree of occlusion) and/or place (accuracy of tongue/lip placement) of stop consonant (plosive) articulation; (2) be useful and appealing to young non-reading children as well as

school-aged children and adolescents/adults; and (3) provide the clinical flexibility to gradually shape correct sound productions.

The Bow and Arrow Game

This stop consonant recognition program is called the "Bow and Arrow Game." The game format and the bow and arrow fantasy were chosen to provide an intrinsically motivating learning environment (Malone, 1981). The bow and arrow fantasy reflects the sudden burst of energy produced during plosive articulation. The program was designed with two screens: a set-up screen to enter the target sound or word and a game screen. The screens parallel the layouts, function keys, and format used by the SpeechViewer.

To play the Bow and Arrow Game, children must match the target sound entered by the clinician (see Figure 1). The game screen consists of a figure of a person with a bow and arrow positioned in front of an archery target (see Figure 2). The goal of the game is to hit the bullseye of the target. When the child successfully matches the target sound, the arrow hits the bullseye, which flashes with accompanying music. A game feature, incorporated into the software, allows the clinician to define the number of times the bullseye must be hit (i.e., 8/10) in order to win a game and receive an additional reward.

Children must match two features of stop consonant production: manner and place. The rise time of burst energy is used to measure the manner of production. The spectral characteristics of acoustic energy associated with the burst and aspiration phases are used to determine the place of production. The game has been designed so that manner and place of production can be trained independently or in combination.

The manner of production (rise time variation) controls the distance the arrow travels. A very weak plosive or fricative-like production (resulting from an incomplete closure of the vocal tract) will cause the arrow to fall short of the target. The place of articulation (spectral feature) controls where the arrow lands on the target. Spectral features of both the target and the child's plosive sound productions are compared. A threshold number, set by the clinician, defines the extent to which the child's sound must match the target sound to be considered successful. A successful match causes the arrow to hit the bullseye of the target. If the child is not successful, the arrow will land on one of the outer rings of the target. The distance of the arrow from the bullseye is proportional to the distance variation of the two spectra.

Hardware Requirements

The stop consonant game was designed to be compatible with the IBM SpeechViewer system (Crepy, Denoix, Destombes, Rouquie, & Tubach, 1983). SpeechViewer requires an IBM PS/2 computer (either model 25 or 30) with 512K of memory and a colour monitor. The software is designed to execute on IBM DOS 4.0. The stop consonant program interfaces with the Analog to Digital (A-D) board used by the SpeechViewer system (SpeechViewer Hardware Option). This board processes speech at a rate of 9,600 samples per second with a 12 bit A-D converter. The speech is prefiltered with an effective roll-off of 28 dB per octave,

Figure 1. Screen #1 from the Bow and Arrow Game. The target sound is displayed. The slope (rise time calculation) is displayed at the top of the screen. The LPC spectrum is displayed in the upper right hand corner of the screen.

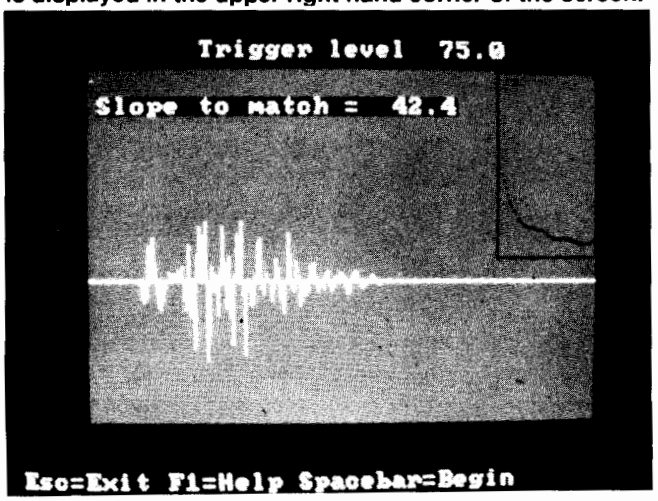
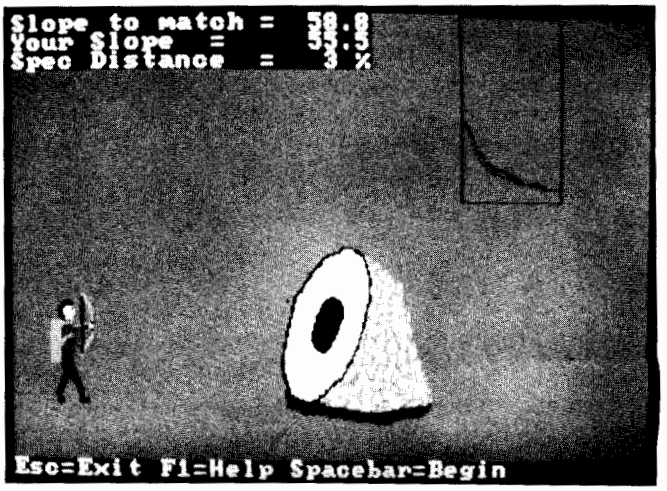


Figure 2. Screen #2 from the Bow and Arrow Game. The slopes of the target and child's production are displayed at the top of the screen along with the spectral distance between the two sounds. The two spectra are displayed in the upper right hand corner of the screen. The threshold for an acceptable match is set by the clinician.



attenuating the frequencies above 3.8 kHz (International Business Machines Corporation, 1988). The microphone used in all testing was a Unidyne A Dynamic microphone (model 580SB). This microphone was an omnidirectional, medium impedance (150 Ohms) microphone, with a frequency response between 50 and 13,000 Hz.

Software Description

A rise time calculation and Linear Prediction Coding (LPC) spectra are computed on a brief time window, following the trigger point of the consonant burst (Markel & Gray, 1976). The trigger point is a threshold voltage that may be varied by the software user because some adjustment may be needed to compensate for variations in background noise levels. The rise time portion of the game responds to the rate of change of the rectified acoustic wave. The distance the arrow travels is a linear function of the average rate of change of the acoustic wave. During the game, two slope values are displayed in the upper left hand corner of the screen, one representing the target and the other the most recent performance.

The spectral characteristics of the acoustic energy immediately following the release of stop consonant articulation is known to provide important information on the place of articulation (Blumstein & Stevens, 1979; Kewley-Port, 1983). Blumstein and Stevens (1979) achieved a high rate of success in consonant identification based on visual recognition of a static spectral template taken from the first 26 ms following articulatory release. Kewley-Port (1983) argued that consonant identification could be further improved by evaluating the spectra of successive time segments.

Due to the difficulty of implementing a matching algorithm which is sensitive to multiple spectral time segments, this research project has focused on static spectral slices with durations and temporal onsets that are under the control of the game user. LPC is used to perform the spectral analysis (Markel & Gray, 1976). LPC spectra are calculated using an autocorrelation method over a time window which can be specified by the clinician or system user (Quackenbush, Barnwell, & Clements, 1988). The time interval over which the spectral match is performed may be varied by the system user. The proximity of the arrow to the centre of the bullseye is a function of spectral distance and the threshold number selected. The threshold level varies from 0-100, and the number selected determines the size of the bullseye in spectral distance units.

The spectral matching aspect of the game involves computation of a spectral distance measure (Quackenbush, Barnwell, & Clements, 1988). The computer compares similarities in spectral shape and amplitude between the target sound and attempted match. The distance is expressed as a percentage and displayed in the upper left corner of the screen along with the two slope measures. The closer the match between the two spectra, the smaller the spectral distance measure. A perfect match results in a spectral distance measure of 0%.

There are two algorithms used for calculation of the spectral distance measure. One algorithm divides the target spectrum into six segments. The algorithm calculates the average of each segment, as well as the point which is furthest from that average. These values reflect the shape of the spectrum within that segment. The same computations are performed on the matching spectrum. The point furthest from the average (i.e., the maximum or minimum value) must be within a specified frequency [+/- 350 Hz] and amplitude range [+/- 30%] from the average value of the target segment.

The second algorithm scans selected amplitude slices of the spectra. The scanning begins at an upper value arbitrarily set at 5 times the average of the total curve. The size of the areas of the target and matching spectra are compared. This algorithm discriminates the spectra on the basis of the size and location of their peaks, weighting higher peaks more than lower ones.

Preliminary testing of the algorithms with a variety of spectral shapes revealed that the first algorithm is most useful for determining the overall similarity and relative distance between two curves. The second algorithm is more reliable at discriminating between the plosive sounds. Values obtained from the two algorithms are averaged to provide a spectral distance measure. Extreme error values on either algorithm, however, result in an arbitrarily assigned spectral distance score of 100% (the maximum error value). The combination of the two algorithms perform better than either alone.

Evaluation of Recognition Accuracy

Subjects

The performance of the spectral matching aspect of the game was evaluated using five normal speaking adults. Subjects included three female and two male speakers, ranging in age from 20 to 35 years. All speakers had no perceived or recorded speech, language, or hearing problems.

Procedures

Speakers were matched against their own speech productions. The speakers were asked to produce a random list of 30 words beginning with the voiceless plosives /p/, /t/, and /k/. These words were matched against two different models. The first model consisted of a voiceless plosive sound

production (e.g., /pʌ/). The second model consisted of a word beginning with the target voiceless plosive (e.g., pool). Speakers were also asked to match the word model a second time using minimal pair contrast words (e.g., tool and cool). To complete the total protocol, each speaker produced a total of 90 words per target phoneme. The accuracy of the program for recognizing normal speech (matching targets with the model) ideally should be 100%, or accurate recognition of each production. Of these 90 words, 30 words should be scored by the computer as correct matches for the target phoneme. Sixty words should be scored as incorrect.

Each speaker recorded all 90 words for each phoneme in one testing session. A constant lip to microphone distance of three inches was used. The time interval over which the spectral match was performed was set at 12 ms because preliminary testing suggested that the program was most reliable at this setting across a variety of vowel contexts.

Results

Results of the recognition accuracy testing are presented in Table 1. The results revealed that the program was 91% accurate (sensitive) at discriminating the /p/ phoneme from /t/ and /k/. The program was able to distinguish /t/ from /p/ and /k/ with a mean accuracy of 75%. The program's ability to distinguish /k/ from /p/ and /t/ was found to be, on average, 71% accurate. Discriminations between the /t/ and /k/ phonemes were the least reliable.

Discussion

In general, the Bow and Arrow Game was found to be more reliable at stop consonant identification than the CAPST program, which achieved recognition accuracy levels of 27-36%. The 90% accuracy level for bilabial voiceless plosive productions is encouraging. This suggests that the program has the potential to be of clinical benefit for teaching bilabial plosive productions.

Accuracy for the identification of the alveolar and velar plosives was found to be 71-75%. Further research and development is needed to increase performance on /t/ and /k/. Future work will focus on the use of matching algorithms specific to the identity of the target phoneme. Preliminary testing has suggested that variations in the weighting factor applied to different portions of the spectra may increase the performance accuracy for /t/ and /k/. Variations in the length of the time window used for the spectral matching also may improve recognition accuracy. The optimal time window for spectral matching in normal speakers has not yet been established. Blumstein and Stevens (1979) used a 26 ms time window. Forrest, Weismer, Milenkovic, and Dougall (1988) evaluated voiceless stop consonants using information (linear moments analysis) derived from the first 40 ms of the voice onset time interval. In this study, a 12 ms window was chosen for reliability testing because preliminary testing with the algorithms suggested that the program was most accurate at this setting across a variety of vowel contexts.

Table 1. Summary of recognition accuracy results for /p/, /t/, and /k/ from five normal subjects expressed as percent correct.

Subjects	Correct Acceptance (sensitivity)(%)	Correct Rejection (specificity)(%)		Overall Accuracy (%)
	/p/	/t/	/k/	
1	93	97	87	92
2	87	90	93	90
3	93	100	87	94
4	97	100	90	96
5	87	73	87	82
Mean	91	92	89	91

Subjects	Correct Acceptance (sensitivity)(%)	Correct Rejection (specificity)(%)		Overall Accuracy (%)
	/t/	/p/	/k/	
1	63	100	63	75
2	80	97	63	80
3	57	93	53	68
4	83	93	63	80
5	70	93	50	71
Mean	71	95	58	75

Subjects	Correct Acceptance (sensitivity)(%)	Correct Rejection (specificity)(%)		Overall Accuracy (%)
	/k/	/p/	/t/	
1	57	90	73	73
2	50	100	53	68
3	53	87	77	72
4	70	93	57	73
5	60	80	60	67
Mean	58	90	64	71

Note. Sensitivity is the ability of an instrument to detect true cases.

Specificity is the ability of an instrument to correctly reject false cases.

The Bow and Arrow Game allows the clinician to vary the time window over which the spectral matching is performed. This feature may be particularly important for application to disordered speakers for whom the optimal time window may differ from that of normal speakers. The optimal time window for matching stop consonant productions may also vary over the course of speech therapy. The variable time window allows acoustic spectra at different time segments to be emphasized (e.g., the initial phase of the burst versus the transition period). This provides the potential for improving articulation skills by focusing on the spectral characteristics of the vowel transitions, if patients do not have the physical capabilities of occluding the vocal tract to produce a plosive burst. Further studies on the clinical application of this feature are warranted.

The threshold which defines acceptable productions is under clinician control. This allows speech skills to be improved gradually until the patient's optimal production is achieved. This optimal production may not be a normal sound but may still improve intelligibility. For example, treatment for individuals with dysarthria is usually aimed at maximizing intelligibility because "normal" speech is often not a realistic goal (Yorkston, Beukelman, & Bell, 1988). An advantage of the Bow and Arrow Game is that it provides the flexibility to train speech sounds using a variety of approaches while leaving the determination of "acceptability" to the clinician. This area, too, warrants more investigation.

This current research is preliminary. The program has been evaluated only with normal speakers. It has yet to be evaluated with individuals who have impaired speech to determine if similar recognition accuracy levels can be achieved. Even if accurate speech recognition thresholds are achieved, this does not guarantee the program's value as a speech training tool. Thus, it is crucial that the program be evaluated in actual speech training situations to determine its effectiveness for training stop consonant productions. This work is currently in progress.

Acknowledgements

This research was supported by The Ontario Ministry of Colleges and Universities (Grant # TO8-010) and IBM Canada Ltd. The authors wish to acknowledge the support of Mr. R. Mighton (IBM Canada Ltd.) and Mr. John McTyre, Ms. L. Boyer, Mr. John Roberts, and Mr. Walt Nawrocki (IBM; Boca Raton) and the administrative support of Ms. R. Gannon, Dr. S. Naumann, and Dr. M. Milner (HMRC).

Address all correspondence to: N. Thomas-Stonell, The Hugh MacMillan Rehabilitation Centre, 350 Rumsey Rd., Toronto, ON, M4G 1R8.

References

Bernstein, L., Goldstein, M.H., & Mahshie, J.J. (1988, Fall). Speech training aids for hearing impaired individuals: Overview and aims. *Journal of Rehabilitation Research and Development, 25*(4), 53-62.

Blumstein, S., & Stevens, K. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66,* 1001-1017.

Crepy, H., Denoix, B., Destombes, F., Rouquie, G., & Tubach, J.P. (1983). Speech processing on a personal computer to help deaf children. In R. Mason (Ed). *Information Processing 83: Proceedings of the IFIP Ninth World Computer Congress* (pp. 19-23). Paris, France: North-Holland.

Dickson, B. C., Ingram J. C. L., & Snell, R. C. (1984). Development of microcomputer-based visible speech aids for the hearing impaired. *Proceedings of the Second International Conference on Rehabilitation Engineering* (pp. 275-276). Ottawa, Ontario, Canada

Ferguson, J. B., Bernstein, L.E., & Goldstein, M.H. (1988, Fall). Speech training aids for hearing-impaired individuals: II Configuration of the Johns Hopkins aids. *Journal of Rehabilitation Research and Development, 25*(4), 63-68.

Fitch, J. L. (1989). Computer recognition of correct sound productions in articulation treatment. *Journal for Computer Users in Speech and Hearing, 5*(1), 8-18.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America, 84* (1), 115-123.

International Business Machines Corporation (1989). *Personal System/2 Independence Series SpeechViewer Application Software* (1st ed.) [Computer program manual]. Atlanta, GA.

Kewley-Port, D., (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 322-335.

Kewley-Port, D., Watson C. S., & Cromer, P. A. (1987) *The Indiana Speech Training Aid (ISTRA): A microcomputer-based aid using speaker-dependent speech recognition.* Paper presented at the American Speech-Hearing-Language Foundation Computer Conference, Houston, Tx.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 4,* 333-369.

Markel, J., & Gray, A. (1976). *Linear prediction coding of speech.* New York: Springer-Verlag.

MicroVideo Corporation (1985). *Video Voice Speech Training System* [Computer program]. Ann Arbor, MI.

Newman, P. W., Creaghead, N. A., & Secord, w. (1985). *Assessment and remediation of articulatory and phonological disorders.* Columbus, Ohio: Charles E. Merrill Publishing Co.

Povel, D. J. & Wansink, M. (1986). A computer controlled vowel corrector for the hearing impaired. *Journal of Speech and Hearing Research. 29,* 99-105.

Quackenbush, S. R., Barnwell III, T. P., & Clements, M. A. (1988). *Objective measures of speech quality.* New Jersey: Prentice Hall Inc.

Watson, C. S., & Kewley-Port, D. (1989). Advances in computer-based speech training (CBST): Aids for the profoundly hearing impaired. *Volta Review, 91*(5), 29-45.

Yorkston, K., Beukelman, D., & Bell, K. (1988). *Clinical management of dysarthric speakers.* Boston: College-Hill Press, Inc.