Automatic measurement of aspects of speech reflecting motor coordination

MARK VOLLRATH

University of Würzburg, Würzburg, Germany

An experimental setting and a method of measurement are described in which five parameters are recorded to reflect speech motor activity. From a psychological or clinical point of view, these parameters are a valuable supplement to traditional motor parameters. In the first experiment, it is shown that this method allows for an objective measure of these parameters. The second experiment proves their high short-term stability. This method requires an IBM-compatible PC, an AD-converter card, software, and microphones. Since a computer is used to evaluate the parameters, a simple registration of these parameters is possible. Applications in the area of speech disorders and short-term changes of speech are discussed.

In both psychological research and clinical applications, the measurement of parameters reflecting motor processes is of great importance. These parameters range from very simple motor reactions (e.g., tapping) to quite complex parameters (e.g., electro-oculogram, EOG). Measurement is especially difficult in one of the most complex motoric behaviors—namely, speech. Acoustic analysis requires substantial temporal and computational effort. However, from a clinical point of view, this effort may be worthwhile, since speech disorders are quite frequent and an objective assessment of these disorders is highly desirable (see, e.g., Catts & Jensen, 1983), and from a psychological point of view, this measurement is especially promising since, as Scherer (1989), for example, points out: “vocalization can be considered to be one of the most direct correlates, and consequently one of the most powerful indicators, of motivational and emotional processes” (p. 167). Thus, by using parameters reflecting motor coordination in speech, activational and emotional processes may be measured indirectly as they influence these motor processes.

The effort needed for the analysis can be reduced when one concentrates on very special aspects of speech, such as the duration of certain kinds of pauses that reflect the time needed to coordinate laryngeal and supralaryngeal structures during the articulation of plosives or stop consonants (e.g., /p/, /t/, and /k/).

To produce a stop consonant, the oral cavity is closed for a very short time in order to build up air pressure. When the stop consonant follows a vowel, the result is a short gap or closure; this is called the “closure interval” (CI). When the air pressure is released, the result is a short burst of noise. After this burst, a certain amount of time is needed before the next vowel can be produced.

This second pause (from the beginning of the burst to the beginning of the vowel) was termed voice-onset time (VOT) by Lisker and Abramson (1964).

Both of the pauses have been investigated in psycholinguistic research because their duration has proved to be a very strong cue for the perception of stop consonants and the distinction between different plosives (for the CI, see, e.g., Dorman, Raphael, & Liberman, 1979; Liberman, Harris, Eimas, Lisker, & Bastian, 1961; for VOT, see, e.g., Lisker & Abramson, 1964).

Consequently, because of their great diagnostic value, CI and VOT have been used in the study of different kinds of speech disorders (for the CI, see Borden, Kim, & Spiegel, 1987; Catts & Jensen, 1983; for the VOT, see Blumstein, Cooper, Goodglass, Statlender, & Gottlieb, 1980; Perkins, 1981; Shenker & Finn, 1985). However, even those well-defined pauses are not easily measured.

Usually, spectrograms are used. Although they produce a lot of data, they require much time to analyze.

Moreover, CI and VOT are very sensitive to the context in which they are articulated (see, e.g., Crystal & House, 1988a, 1988b; Klatt, 1973; Summerfield, 1981), which means that if a certain stop consonant is produced after different vowels, either in different positions within the syllable or word or with different speech rates, the duration of these parameters changes. Thus, when CI and VOT are measured in fluent speech, a great degree of variation in their duration results, making it difficult to find differences between subgroups of subjects.

The first problem can be overcome when only the speech energy is used (which roughly corresponds to loudness) and not the frequency information. This reduces the amount of data that has to be sampled and facilitates the measurement of CI and VOT since the signal is much easier to handle.

Concerning the second problem, a restriction of the context in which the stop consonants are produced provides a simple solution, and a much more stable duration of the CI and VOT is gained.

Correspondence should be sent to M. Vollrath, Psychologisches Institut III, Röntgenring 11, 97070 Würzburg, Germany (FAX: 49-931-16220; e-mail: mvollrath@vax.rz.uni-wuerzburg.d400.de).

METHOD AND APPARATUS

The subject's speech is recorded by means of a small condenser microphone taped to the throat. It would also be possible to use a normal air microphone. This, however, is not recommended since background noise (e.g., the speaking of the experimentator) may contaminate the signal. To decrease background noise even further, a low-pass filtering of 10 kHz is required (this value was chosen because most of the speech energy needed for identifying the CI and VOT lies below 10 kHz, resulting in a very good reduction of noise). This analog signal is then converted to digital values (0 to 2,048) by means of an AD card (ME-30-2, Meilhaus Electronic GmbH) that is inserted into an IBM-compatible computer. The stimulus presentation, speech recording, and evaluation of the signal are accomplished by means of a program written in Turbo Pascal 6.0. The timing is done with a software timer (Bovens & Brysbaert, 1990). An IBM-compatible personal computer is required; no extended memory is necessary.

The speech energy is recorded as fast as the computer will allow (with a maximum of 30 kHz), rectified and averaged over 4 msec, yielding a rate of measurement of 250 Hz. Because of the rectification and averaging, only the information about the speech energy remains.

Figure 1 gives the speech energy of a German "ak" (pronounced as in the English word *luck*) and "ka" (pronounced as in the English word *car*). The CI is measured from the end of the high speech energy of the "a" to the beginning of the burst (the small spike to the right). The VOT is determined from the beginning of the burst to the beginning of the high energy of the "a" following the "k." Although not all signals produced are as good as those displayed, in most cases it is very easy to measure the duration of the two parameters.

Figure 1. Speech energy of a German "ak" (left) and "ka" (right) of the picture. At the abscissa, time is displayed (each picture giving about 600 msec); at the ordinate, speech energy is displayed. The horizontal dotted lines divide the ordinate by 200 (the unit is the digital value of the speech energy, with a maximum of 2,048 units).

Since the quality of the signal depends on the location of the microphone at the subject's throat, an on-line display of the speech energy is included in the program. The use of this option provides both the best location for the microphone and a clear signal for most of the subjects. But even in the "worst cases," a measurement of CI and VOT is possible for at least most of the subject's articulations.

A simple algorithm is used to measure the duration of the CIs and VOTs (see Figure 2 for a flow chart). For this purpose, a digital threshold is introduced which transforms the energy signal to a pattern of *on* (above the threshold) and *off* (below the threshold). For the detection of the CI, a long period of *ons* (corresponding to the vowel) is looked for first. The algorithm for locating this period proceeds as follows (see left part of Figure 2): From the beginning of the appearance of the signal on the screen, any value lying above the threshold is searched. Whenever an *on* is found, the next 200 msec are evaluated; if more than 80% of these 200 msec are above the threshold, this *on* corresponds to the beginning of the vowel (200 msec were chosen because this value is a good approximation of the usual duration of a vowel).

In a similar way, the beginning of the CI is found (see second flow chart from the left in Figure 2): From the beginning of the vowel, an *off* is searched for, followed by at least 32 msec of other *offs* (32 msec were chosen because this is certainly a lower limit for the CI in this context). To obtain the end of the CI, the first value above the threshold after the beginning of the CI is looked for.

An analogous search algorithm is used for the measurement of the VOT (see right part of Figure 2): In order to find the beginning of the burst, an *on* is looked for. If, within the next 200 msec, more than 60% *ons* are included, the beginning of the burst will have been found (60% is chosen because besides the following vowel, the pause belonging to the VOT is also included). To find the end of the VOT, from the start of the VOT onward, the beginning of the following vowel is searched for in the same way as above (to qualify as the beginning of the vowel, an *on* has to be followed by at least 80% of *ons* in the next 200 msec).

This algorithm yields a very good detection of the CI and VOT. However, the threshold has to be chosen manually so that background noise and the pauses of the CI and VOT lie below the threshold and the energy of the burst and the vowel are above the threshold. To avoid measurement errors, the speech energy of each syllable is displayed on the screen, and the duration of the CI or VOT as found by the algorithm is indicated by vertical lines. The user may change these values manually as well as change the parameters (such as percentages and durations of periods to search within) of the search algorithm. However, in most of the cases, this is not necessary if a sensible threshold has been chosen.

To get a stable production of the syllables, only the two German syllables "ak" and "ka" are used, and the manner of their production is strongly limited. The stop consonant /k/ was chosen because here the energy of the burst

Figure 2. Flow charts for the algorithm used for the semi-automatic measurement of the duration of the closure interval (CI) and voice-onset time (VOT). The "t" corresponds to the units of measurement (4 msec) from the beginning of the appearance of the signal on the screen. For the description, see text.

is greatest (Vollrath, 1993). The measurement of the duration of CI and VOT in other voiceless plosives is also possible for most subjects, but, for some, the energy of the burst is too low to achieve a measurement (Vollrath, 1993). CI and VOT cannot be measured in voiced plosives by using speech energy, since, during the closure, the vocal cords keep vibrating, producing energy that is recorded. Thus, no pause is visible. The influence of the vowel on the clarity of the signal is not very strong, therefore other vowels could be used as well.

This procedure of presenting "ak" and "ka" yields a separate analysis of CI and VOT. These two syllables are presented on the screen of the computer in random order. The subject is instructed to produce the appropriate syllable as soon as it appears on the screen. The subject is instructed to trigger each presentation by pressing the "enter" key on the keyboard. After a variable foreperiod of 1, 2, or 3 sec the stimulus appears. Each of these foreperiods is given three times, resulting in nine reactions for each of the two syllables.

In addition to the CI and VOT, a speech motor choice reaction time can be measured, since the measurement of speech motor reaction times yields interesting material for a comparison of these reaction times with the two pauses; moreover, two simple reaction tasks were added in the second part of the test.

The subjects are told to begin producing an "a" (as in *car*) as soon as it appears on the screen, to hold it as long as it is displayed on the screen, and to stop as soon as it disappears. The foreperiod is again varied (1, 2, or 3 sec), as is the duration of the "a" (1, 2, or 3 sec). Both are combined, yielding nine stimuli, each presented twice

in random order. Thus, two simple reaction times, a start and a stop reaction time, can be measured.

These three reaction time tasks were chosen because they allow the differentiation of processes involved in reacting to a stimulus. For each of these reaction times, the time needed for the perception of the stimulus on the screen and for the triggering of the reaction is included. In the start reaction time, the time needed to activate the appropriate motor program is added, unlike with the stop reaction time. In the choice reaction time, an additional decision time is included.

Thus, by comparing these reaction times, the process of triggering a reaction can be analyzed quite thoroughly. By comparing CI and VOT with those reaction times, two parameters reflecting automatic motor processes in speech are contrasted with parameters reflecting both speech motor times and cognitive components.

This whole test is called the MSMT (measurement of speech motor times). In this test, five parameters are measured: CI, VOT, choice reaction time, start reaction time, and stop reaction time. A German and an English version of the Turbo Pascal program are available from the author (see below).

EXPERIMENT 1

Since the measurement of the duration of the choice reaction time, CI, and VOT is done manually, although with the aid of a computer algorithm, the objectivity—that is, the independence of the result of the measurement from the rater—has to be shown. In the case of the simple reaction times, this is not necessary, because the evalu-

Figure 3. Means and standard deviations for the two raters of the closure interval (CI), the voice-onset time (VOT), and the choice reaction time (Choice RT). The correlations between the two raters have been added.

ation is done automatically and the detection of the high energy of the vowel presents no problem.

Thirty American male subjects between 21 and 35 years of age performed one trial of the MSMT. These data were analyzed by two raters. The duration of every single CI, VOT, and choice reaction time was compared. Figure 3 shows the means and standard deviations of the two raters for the three parameters as well as the correlations between the raters.

The difference in the mean duration of the three parameters between the two raters is very small (0.8 msec for the CI, 0.2 msec for the VOT, and 0.4 msec for the choice reaction time). These differences are far smaller than the rate of measurement (4 msec). Moreover, the correlation between the raters is nearly perfect. Thus, the MSMT gives a measure of the parameters with a very high objectivity. This experiment also shows that the MSMT with the two syllables "ak" and "ka" can be used in American English without any problems in the measurement of IC and VOT.

EXPERIMENT 2

If differences between subgroups of subjects in a certain parameter are to be detected, it is necessary to show that the value of this parameter is stable for the subjects (at least for short periods of time). To examine this aspect, the MSMT was done twice with a short pause between the two trials. If the stability of the parameters is high, the mean duration for each subject must not change very much, and the correlation between the two trials has to be very high.

This was tested with two samples. The first sample consisted of 30 German subjects (mean age 23 years, 11 males, 19 females, students); the second sample consisted of 18 German subjects (mean age 24 years, 9 males, 9 females, students).

Figure 4 gives the means and standard deviations of the two trials and the correlations between them for the five parameters in the two samples.

The mean duration of the three parameters reflecting automatic motor processes (IC, VOT, and stop reaction time) was very stable in these two trials (differences in the first sample were 3.4 msec for the CI, 0.6 msec for the VOT, and 0.1 msec for the stop reaction time; differences in the second sample were 0.5 msec for the CI, 2.1 msec for the VOT, and 0.5 msec for the stop reaction time). Since these differences are substantially smaller than the rate of measurement and the correlations between the two trials are very large, the three parameters show a high short-term stability.

The mean duration of the choice and start reaction time changed between the first and the second trials (in the first sample—17.3 msec for the choice reaction time, 14.8 msec for the start reaction time; in the second sample—19.3 msec for the choice reaction time, 9 msec for the start reaction time). Although these changes are substantially greater

Figure 4. Samples 1 (top) and 2 (bottom). Means and standard deviations for the two trials of the closure interval (CI), the voice-onset time (VOT), the choice reaction time (Choice), the start reaction time (Start), and the stop reaction time (Stop). The correlations between the two trials have been added.

than the rate of measurement, the high correlations indicate that these two reaction times also show a high short-term stability. The larger differences may be due to the larger overall duration of these parameters.

The results of this experiment show that there is very high short-term stability in the mean duration as well as in the correlation of all the parameters measured. Although there are some differences in the mean duration of the parameters between the first and second trials, these differences are not very strong, especially for the CI, VOT, and stop reaction time, for which the mean difference is less than the rate of measurement. Thus, with the use of the MSMT, a very stable measurement of the parameters is possible.

SUMMARY AND CONCLUSION

An experimental setting is presented that provides a relatively easy and objective measurement of two parameters reflecting speech motor coordination, the closure interval, and the voice-onset time, in addition to three speech motor reaction times differing in complexity. This setting yields a very stable production of the syllables used to measure those parameters. Thus, the typical individual duration of these parameters can be measured. These durations can, of course, not be directly related to specific movements of muscles involved in speech production since only acoustical parameters are evaluated. However, they are the result of a very fast and highly coordinated muscular action, thus reflecting the state of motor coordination processes in speech.

The results also show a substantial variation between the subjects of the duration of the parameters, even for a homogenous sample of German speakers, which may be due to dialect or individual preferences of articulation. It remains to be shown if these parameters are then able to differentiate subgroups of subjects—for example, stutterers and nonstutterers.

However, due to the high short-term stability it can be expected that at least intra-individual variations can be detected—for example, during the course of speech therapy in subjects with speech disorders.

A strong limitation of the MSMT for this kind of clinical use is due to the fact that only one stop consonant (/k/) is produced in a very restricted manner. Generally, in the clinical field, a very wide range of vowel environments and stops with different places of articulation are used—sometimes by reading a text, sometimes even by analyzing fluent speech. In contrast to those designs, the MSMT yields only very restricted information, hence transferring the results may not be possible.

However, the parameters obtained with the MSMT may be especially apt to detect short-term variations of speech due to activational or emotional changes that are caused, for example, by psychoactive substances. This idea was investigated by Swartz (1992), who examined the effect of alcohol on the VOT variability. Although he did not find changes in VOT variability, this may be due to the

setting in which the speech was produced and the limited range of parameters that he used. Pisoni and Martin (1989) found longer CIs when subjects were intoxicated as opposed to sober. Since other highly automatic motor processes are very sensitive to the effects of alcohol, one would expect the same for the duration of the IC and VOT. For example, in his review on the effects of low dosages of alcohol, Krüger (1993) showed that the precise measurement of eye movements is the most sensitive psychophysical function that can detect impairment induced by alcohol, since this function is interfered with even at blood alcohol concentrations noticeably below 0.03%.

For these applications, the above-mentioned seeming disadvantage of the MSMT (restricting the speech produced in a very strong manner) turns out to be a strong factor in favor of the test: Since there is a very high individual short-term stability of the parameters obtained, small changes in their value can be detected. The sensitivity of the MSMT in this field is currently being tested by using alcohol (with a strong sedative component) and caffeine (with an activating component).

Availability

A copy of the MSMT program on an MS-DOS compatible disk may be obtained from the author by sending \$10 (covering the cost of mailing and reproduction). The software is also available via electronic mail from the author at mvollrath@vax.rz.uni-wuerzburg.d400.de. Because this program is adapted to the Meilhaus AD-card used, it will only work with this card. This card can be acquired from Meilhaus Electronic GmbH, Fischerstr. 2, 82178 Puchheim, Germany (Phone: 49-89-807081, FAX: 049-89-808316) at a price of approximately \$1,200 plus shipping. This card will also be available in the near future from dealers in the US when they are selected.

For use with other AD cards, the source code of the TURBO-PASCAL program is included so that the program can be changed accordingly. Because the main program uses a unit in which all functions regarding the AD card are stored, only this unit has to be changed and is included in the main program.

The author will also include a sketch of the filter and the connection between the microphone and the AD card.

REFERENCES

- BLUMSTEIN, S. E., COOPER, W. E., GOODGLASS, H., STATLENDER, S., & GOTTLIEB, J. (1980). Production deficits in aphasia: A voice-onset time analysis. *Brain & Language*, **9**, 153-170.
- BORDEN, G. J., KIM, D. H., & SPIEGLER, K. (1987). Acoustics of stop consonant-vowel relationships during fluent and stuttered utterances. *Journal of Fluency Disorders*, **12**, 175-184.
- BOVENS, N., & BRYSSAERT, M. (1990). IBM PC/XT/AT and PS/2 Turbo Pascal timing with extended resolution. *Behavior Research Methods, Instruments, & Computers*, **22**, 332-334.
- CATTS, H. W., & JENSEN, P. J. (1983). Speech timing of phonologically disordered children: Voicing contrast of initial and final stop consonants. *Journal of Speech & Hearing Research*, **26**, 501-510.
- CRYSTAL, T. H., & HOUSE, A. S. (1988a). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, **83**, 1553-1573.

- CRYSTAL, T. H., & HOUSE, A. S. (1988b). Segmental durations in connected-speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, **83**, 1574-1585.
- DORMAN, M. F., RAPHAEL, L. J., & LIBERMAN, A. M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, **65**, 1518-1532.
- KLATT, D. H. (1973). *Voice-onset time, frication and aspiration in word-initial consonant clusters* (Quarterly Progress Report No. 109, pp. 124-136). Cambridge: Massachusetts Institute of Technology, Research Laboratory of Electronics.
- KRÜGER, H.-P. (1993). Effects of low alcohol dosages. In H.-D. Utzelmann, G. Berghaus, & G. Kroj (Eds.), *Alcohol, drugs and traffic safety—T92* (pp. 763-778). Cologne, Germany: Verlag TÜV Rheinland.
- LIBERMAN, A., HARRIS, K. S., EIMAS, P., LISKER, L., & BASTIAN, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language & Speech*, **4**, 175-195.
- LISKER, L., & ABRAMSON, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.
- PERKINS, W. H. (1981). Implications of scientific research for treatment of stuttering—a lecture. *Journal of Fluency Disorders*, **6**, 155-162.
- PISONI, D. B., & MARTIN, C. S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical & Experimental Research*, **13**, 577-587.
- SCHERER, K. R. (1989). Vocal correlates of emotional arousal and affective disturbances. In H. Wagner & A. Manstead (Eds.), *Handbook of social psychophysiology* (pp. 165-197). Chichester, U.K.: Wiley.
- SHENKER, R. C., & FINN, P. (1985). An evaluation of the effects of supplemental "fluency" training during maintenance. *Journal of Fluency Disorders*, **10**, 257-267.
- SUMMERFIELD, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1074-1095.
- SWARTZ, B. L. (1992). Resistance of voice onset time variability to intoxication. *Perceptual & Motor Skills*, **75**, 415-424.
- VOLLRATH, M. (1993). *Mikropausen im Sprechen*. Frankfurt, Germany: Peter Lang.

(Manuscript received January 8, 1993;
revision accepted for publication August 19, 1993.)