Journal of Experimental Psychology:
Human Perception and Performance
2008, Vol. 34, No. 5, 1317–1326

0096-1523/08/$12.00 DOI: 10.1037/0096-1523.34.5.1317

Improving Accuracy in Detecting Acoustic Onsets

Wouter Duyck, Frederik Anseel, Arnaud Szmalec, Pascal Mestdagh, Antoine Tavernier, and Robert J. Hartsuiker
Ghent University

In current cognitive psychology, naming latencies are commonly measured by electronic voice keys that detect when sound exceeds a certain amplitude threshold. However, recent research (e.g., K. Rastle & M. H. Davis, 2002) has shown that these devices are particularly inaccurate in precisely detecting acoustic onsets. In this article, the authors discuss the various problems and solutions that have been put forward with respect to this issue and show that classical voice keys may trigger several tens of milliseconds later than acoustic onset. The authors argue that a solution to this problem may come from voice keys that use a combination of analogue and digital noise (nonspeech sound) detection. It is shown that the acoustic onsets detected by such a device are only a few milliseconds delayed and correlate highly (up to .99) with reaction time values obtained by visual waveform inspection.

Keywords: voice key, measurement accuracy, naming, threshold, acoustic bias

In the late 19th century, Donders laid the foundations of mental chronometry with his famous article on the reaction time (RT) subtraction method (Donders, 1868). Ever since, researchers have used RT measurements to learn more about cognitive processes that may not be revealed by simple observation or introspection. Whereas manual (motor) responses are often recorded by millisecond-accurate response boxes or by less accurate keyboard presses or mouse clicks, the speed of vocal responses is commonly assessed by voice keys. This is true not only for psycholinguistic studies, but also for other domains of research that use spoken responses. For instance, Rastle and Davis (2002) calculated that 95% of all articles that reported naming latencies, published between 1995 and 1999 in the *Journal of Experimental Psychology: Human Perception and Performance,* used voice keys. As noted by Kessler, Treiman, and Mullennix (2002), very few published articles contain information about the specific type of voice key used or about the parameters that may affect its functioning (e.g., its trigger level, see further). This is probably due to the widespread belief that most voice keys are based on the same principle and should therefore be comparable across studies. Indeed, it is the case that most current electronic voice keys are based on the same amplitude threshold principle (e.g., see also the incorporated voice key routines in the popular experiment-generating DMDX software; Forster & Forster, 2003). The speech of the participants, which contains a certain amount of acoustical energy, is transduced into electrical energy by means of a microphone. The signal derived from the microphone is an electrical voltage that has to exceed a certain voltage threshold in order to discriminate between background noise and speaker utterances. As soon as the electrical voltage exceeds this threshold, a logical signal is sent to a connected computer. Then, in a prototypical RT experiment, the amount of time that has passed since the presentation of a stimulus is calculated. Most voice keys of this type allow the experimenter to manually adjust the amplitude threshold to account for differences in the level of background noise and the loudness of the speaker. Also, many of these voice keys have an operator-adjustable amplification switch that determines how much the input signal is amplified before it is evaluated against the threshold. Notwithstanding the widespread use of the voice key, there is surprisingly little attention for the apparent problems that follow from this amplitude threshold starting principle. However, recently a lively debate has originated from a few high-impact studies that have reported very disturbing findings about the accuracy of classical voice keys as measurement devices (e.g., Kessler et al., 2002; Rastle, Croot, Harrington, & Coltheart, 2005; Rastle & Davis, 2002). Because of its possible far-reaching consequences, this debate should and probably will receive more attention in the near future. In the next section, we discuss the major problems that may arise in precisely determining the acoustic onset of speech with classical voice keys. Later, we discuss the various possible solutions for some of these issues.

Voice Key Problems

The reason why most voice keys are based on the amplitude threshold principle relates to background noise. Even in a typical, quiet experiment room, there is always some continuous or intermittent noise due to computer equipment, ventilation, irrelevant movement, participant breathing, lip clicks, or other interfering factors. It would be highly undesirable if a voice key often triggered due to these nonspeech sounds. In order to limit these false alarms to an acceptable minimum, an experimenter typically adjusts the threshold of the voice key before the experiment so that

Wouter Duyck, Frederik Anseel, Arnaud Szmalec, and Robert J. Hartsuiker, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium; Pascal Mestdagh and Antoine Tavernier, Audiovisual Support Lab, Ghent University.

This research was made possible by the Research Foundation Flanders of which Wouter Duyck is a postdoctoral research fellow. We thank Sarah Bernolet for help with the phonetic categorization of our stimuli. Technical inquiries concerning the NEVK's electronic circuitry may be directed to Pascal Mestdgh, pascal.mestdagh@ugent.be.

Correspondence concerning this article should be addressed to Wouter Duyck, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000, Ghent, Belgium. E-mail: wouter.duyck@ugent.be

the device is not triggered by the background noise. However, this threshold may not be set too high, because this increases chances of not detecting a speech sound (misses) or detecting a sound too late (see below).

There are a number of problems with the prototypical approach sketched above. A first group of problems concerns comparability across participants and studies. Operator-adjustable threshold levels and amplification settings imply different detection times for the same sounds. Because it is not easy to measure or quantify these threshold and amplification levels (often specified by an inaccurate turning switch), they are virtually never reported in published papers, which limits comparability across studies. Even if they were reported, identical parameters could not always yield equivalent experimental settings as they may also be influenced by the specific equipment used, such as the type of microphone or its distance from the speaker. Also, the threshold and amplification levels may interact with speaker characteristics such as gender, voice loudness, or reading skills, so that identical threshold and amplification levels may not even guarantee comparability across participants. From this, it may be advisable that the ideal voice key should not allow operator-adjustable settings, even if these settings could be implemented and quantified more accurately (digitally). So, there should be another solution to deal with the variability in the experimental setting.

A second group of problems is perhaps even more serious. Because the rising time of acoustic energy is not zero, some of the true speech signal just below the threshold level will be considered as noise, and the voice key will only be triggered some time after the initial acoustic onset. This delay is not a matter of milliseconds but may be surprisingly large. For instance, Pechmann, Reetz, and Zerbst (1989) reported that voice key triggers may occur only 100 ms after the acoustic onset is visible in the waveform. Similar delay values of 139 ms and 87 ms with two types of threshold voice keys were recently obtained by Rastle and Davis (2002; see also Sakuma, Fushimi, & Tatsumi, 1997; Yamada & Tamaoka, 2003). If the delay was constant across stimuli and phonemes, this would not be a real problem for experiments. In that case, RTs in different experimental conditions would just be artificially inflated by the same constant value. However, this is not the case. Phonemes differ not only in the acoustic energy that is released when producing them, but also in the time that it takes to build up this energy. Hence, different phonemes will exceed the voice key threshold at a different time even if they are initiated at exactly the same moment. This delay problem is not restricted to a few particular phonemes, and there is great variability in the voice key delays for different phonemes. Generally, nasals are easily detected, whereas more problems arise for fricatives and plosives (Pechmann et al., 1989; Rastle & Davis, 2002; Sakuma et al., 1997). So, a threshold voice key may be triggered 10 ms after the true acoustic onset of speech for some phonemes but only after 90 ms for others so that the latter stimuli yield artificially increased RTs. This phenomenon is illustrated in Figure 1. Note that the size of the biases may also vary across participants.

Finally, Rastle and Davis (2002) showed that even the second phoneme of a vocal response influences the variability in the delay with which the voice key is triggered (see also Kessler et al., 2002). Comparing with visual waveform inspection, they found that complex /s/ onsets (e.g., *spat*) are generally detected later by a voice key than are simple onsets (e.g., *sat*), because voicing occurs earlier in the latter case, even though both conditions shared the initial phoneme /s/. From the above, it should be clear that the ideal voice key should detect acoustic onsets as soon as they are visible in the waveform, even for sounds that build up slowly or reach only low amplitude.

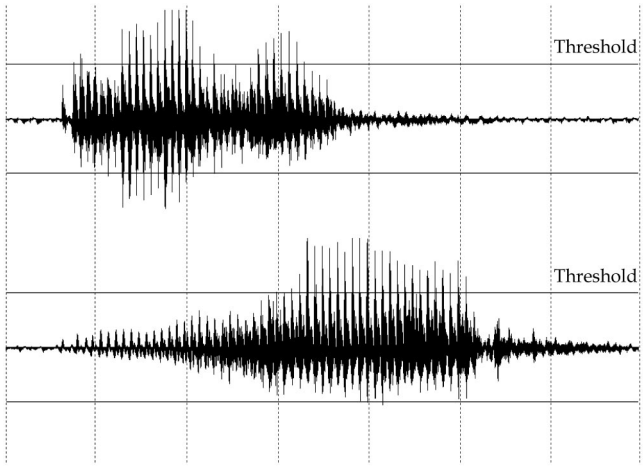

Figure 1. Recorded waveforms for Dutch words *paar* [pair] (upper part) and *jaar* [year] (lower part), matched for acoustic onset (63 ms after initial recording, determined by visual analysis with adequate zooming). With the displayed threshold level (full line), *paar* triggers the voice key after 75 ms (12-ms delay). *Jaar* triggers the voice key after 289 ms (226-ms delay). Dotted vertical lines represent 100-ms intervals.

How serious are these problems? Of course, researchers may argue that these variable delays in voice key triggering (even if they are around 100 ms) only make up a small part of the total time between stimulus presentation and response (typically between 500 and 700 ms). However, it is important to realize that the theoretical effects of interest in many studies are not hundreds but rather tens of milliseconds or even fewer. If these variable onset detection inaccuracies are not randomly spread across experimental conditions, this may even change the direction of the effect. Whereas this may seem more like a theoretical possibility than an actual danger for research practice, Rastle and Davis (2002; see above) have proved otherwise. They identified an apparent contradiction in the psycholinguistic literature, with some authors (e.g., Frederiksen & Kroll, 1976) reporting slower naming of complex onsets (e.g., *spat*) relative to simple onsets (e.g., *sat*) and others claiming the opposite (e.g., Kawamoto & Kello, 1999). Replicating these earlier studies using visual analysis, a classical amplitude threshold voice key, and an integrator voice key (for technical details, see Rastle & Davis, 2002), they surprisingly obtained the three logically possible statistical results for the two onset conditions. The more accurate visual analysis showed significantly faster naming (9 ms) for complex onsets (replicating Kawamoto & Kello, 1999), whereas the two voice keys, respectively, showed significantly slower naming (−11 ms) and a statistical null effect (2 ms). These findings clearly demonstrate that threshold voice keys may be responsible for some controversies or contradictory results in the literature.

We have now discussed some major issues that are detrimental for the accuracy of current voice keys in determining the acoustic onset of vocal responses. In the next section, we present the technical description for a new, easy-to-make voice key designed by Pascal Mestdagh (an electrical engineer). We argue that this

device offers a hardware-based solution for some of the problems mentioned above, and we show that the acoustic onsets detected by such a device correspond very closely to the onsets determined by visual waveform analysis. In the general discussion section, we discuss performance of this new voice key and compare it against other approaches of a nontechnical nature that have been put forward to deal with voice key inaccuracies.

Toward a Possible Solution

In the previous sections, we argued that the ideal voice key should not allow operator-adjustable parameters (threshold level, amplification setting) because different parameters imply different detection times for the same sounds across experiments and participants. In the voice key presented below, all features and settings were implemented in the hardware. Hence, it did not yield RT variations induced by the operator's settings. Second, we argued that a good voice key should trigger on low-energy and slowly rising sounds as soon as acoustic onset is visible in the waveform, just as for more easily detectable sounds. In our concept, which we call the noise elimination voice key (NEVK) from now on, this is achieved through massive amplification of the voice key's input signal. This results in fast triggering after acoustic onset, even for low amplitude signals. To avoid the false triggers that this amplification mechanism may cause, the resulting amplified sound is processed through a digital hardware-based noise detection circuit, which rejects most of the amplified nonspeech sounds. In the following experiments, we show that this combination results in a system with accurate acoustic onset detection without a high rate of false alarms. In the next section, we describe the NEVK in more technical detail (see http://cornea.ugent.be/nevk/nevk.html for the NEVK's electronical scheme).

Technical Specifications

As described above, the NEVK consists of two parts, one of which is analogue (as are classical voice keys) and one of which is digital.

The NEVK's Analogue Part

In our concept, a headset microphone (Sennheiser PC 130) signal is fed into a low voltage audio amplifier (LM386), which amplifies the speech signal about 20 times. This signal passes through two filter stages (340 Hz high-pass and 2850 Hz low-pass). The filtered audio signal is then amplified by a noninverting amplifier (LM324) with a gain of about 2000. At this point, the resulting signal is no longer an audio signal but rather a series of analog spikes representing all the oscillations of the original acoustical signal: The NEVK tends to "over-amplify" (saturate) the incoming acoustical signal in order to achieve the fastest possible detection of any incoming signal. This means that the amplified acoustical signal in the circuitry is no longer an exact copy of the input but merely serves as a derivative detector signal that "gets the [triggering] job done." This amplified signal is forwarded to the second digital noise-detection part of the NEVK.

The NEVK's Digital Part

The noise eliminating circuitry operates according to the following principle: Impulsive noise signals (falling pencil, lip pops, clothing frictions, etc.) decay rapidly, and most of the time exponentially, toward zero. Therefore, they will produce very little or no (detectable) oscillations about 50 ms after signal start. In the voice key circuitry, these signals are eliminated by looking for oscillations in a time window of 75 ms that begins 50 ms after signal start. Because speech signals generally produce oscillations 50 ms after signal start, they will be validated. Of course, theoretically, there may still be noise signals that surpass this elimination circuitry, considering the simplicity of the logic used. Nevertheless, practical tests have shown that the circuitry works fine in most cases (see further). In technical terms, three monostable multivibrators (74123) produce the timing signals necessary for the speech/noise discriminating circuitry. As soon as an (amplified) acoustical signal is detected, one of these timers is started, after which discrimination occurs during the fixed time slot of 200 ms. A second timer is started after 50 ms (potentiometer 2), which again starts a third timing window of 75 ms (potentiometer 3), during which a counter (74393) counts the spikes that appear. If the number of counts is greater than eight, a signal is clocked into a data flip-flop (7474). This signal is presented to the outside world at the end of the timing interval of the first timer (set at 200 ms). If the amplified acoustic signal is not rejected as noise by the circuitry above, it will be validated and hence produce an output signal. Because of the massive signal amplification before evaluation against the threshold, the sound level that triggers the NEVK is many times lower than for classical voice keys. So, although the NEVK theoretically still has a threshold, its amplification mechanism yields a triggering threshold that is very close to zero amplitude.

Additional Features and Considerations

Because the NEVK is built around dedicated hardware circuitry, it operates independently of computer hardware, experimentation software, or operator switches. In order to give some kind of visual feedback to the operator about the functioning of the voice key, two light emitting diodes are added to the circuitry. One diode signals the start of presence of acoustical energy. It is emitting light during the complete signal validation phase that takes 200 ms. The other diode signals the validation of the speech signal that has passed the noise detection and the sending of a logical signal to the outside world. This signal is sent to a parallel port and a game port, which may be monitored by any experimental presentation and response registration software. Because the speech validation phase always takes 200 ms, this output signal is always delayed 200 ms relative to acoustic onset. This fixed delay should be accounted for when analyzing RTs. Finally, just as for any other voice key, in order to have a fully operational voice key setup, we recommend choosing an experimental room that is very well acoustically isolated and where there is not much ambient noise coming from the equipment used (PCs, screens, etc.). Furthermore, the electricity supply system should be free of electrical noise (due to drilling machines, refrigerators, motors, etc.) because this kind of noise may raise the noise of the electronic circuitry of the voice key above the fixed threshold setting and thus produce false triggers.

Experiment 1: Demonstration of Onset Detection Accuracy

Method

Stimuli. The 22 target words to be named were selected from the CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993) using the WordGen stimulus selection program (Duyck, Desmet, Verbeke, & Brysbaert, 2004). All targets were highly frequent, four-letter nouns. All 22 targets had a different initial phoneme, corresponding to a different beginning letter, in order to assess the NEVK's performance across onset types. Words beginning with C, Q, X, and Y were excluded. The resulting targets were as follows: *arts* [doctor], *blik* [can], *deur* [door], *eeuw* [century], *fles* [bottle], *glas* [glass], *hand* [hand], *idee* [idea], *jaar* [year], *kind* [child], *laag* [layer], *mate* [degree], *naam* [name], *orde* [order], *paar* [pair], *raam* [window], *stad* [city], *unie* [union], *voet* [foot], *wijn* [wine], and *zaak* [case]. Each target was named 25 times, resulting in 550 trials.

Apparatus. Stimulus presentation and timing was programmed in the C language using timing routines published by Bovens and Brysbaert (1990). RTs from two voice keys were recorded simultaneously through two different pins of the PC's gameport. The first voice key was the device described above. The second device was a classical threshold voice key (CTVK) with an operator-adjustable threshold and amplification setting, as described in the introduction, designed by Antoine Tavernier. This voice key was calibrated for the experimental setting (e.g., experimental room, speaker volume) by an experienced user (Wouter Duyck) so that the trigger and amplification levels were just high enough not to be triggered by the background noise. The test was performed in a regular, quiet (but not sound shielded) office room, in order to have a conservative setting for measuring voice key performance. Waveforms of spoken responses were recorded through a Sennheiser MD 421-U-4 microphone and forwarded through the right input channel of an Inter-M MX-642 mixing panel. Simultaneously with the presentation of the naming cue, the experiment program sent a pulse to the parallel port (LPT1). This signal was registered by an electronic circuit, which instantly generated a tone burst that was fed to the left channel of the same mixing panel. Both left and right channels were used as input of an external USB Sound Blaster MP3+ soundcard. This card's input was recorded by a second PC as 16-bit 44 kHz WAVE files. By evaluating the left channel against the right, the externally recorded waveforms could be synchronized with the stimulus/cue presentation that was handled by the first PC. Two trained judges (Frederik Anseel and Arnaud Szmalec) identified the acoustic onset of responses in these waveforms using a combination of visual and auditory inspection in the WaveLab software package (Version 5.01). They had no access to the recorded voice key RTs and did not consult with each other at any point. So, all of the following results were obtained without recoding of divergent values.

Procedure. A 28-year-old male speaker, naïve to the purposes of the experiment, was instructed to perform a delayed naming task. On each trial, a target word was presented centered on the screen during 1000 ms. The target was replaced by a blank screen for 750 ms, which was followed by the presentation of a question mark. This was the naming cue for the target word that had just been presented. In order to assess voice key performance independently of other potentially confounding factors, a delayed naming task was used in order to minimize variance in RTs due to cognitive (stimulus) processing. The intertrial interval was 1000 ms. Twenty practice trials were administered before the actual experiment. After each hundred trials, a short break was included. The entire experiment lasted about 40 min.

Results

Mispronounced tokens were excluded from the analysis. Following earlier research, a voice key RT was treated as a false alarm if it occurred more than 50 ms earlier than the visual inspection RT.[1] The NEVK yielded 2.55% false alarms, whereas the CTVK yielded none. This is consistent with earlier research (e.g., Rastle & Davis, 2002) that also reported very low false alarm rates for voice keys that yield large triggering delays. Table 1 shows that the NEVK combines excellent RT accuracy with an acceptable false alarm rate. Mean absolute deviation between the NEVK's RTs and visual inspection RTs was only 5.6 ms. This is much lower than the average delay of 72.33 ms obtained with a classical threshold voice key (on exactly the same spoken responses), which is comparable to the threshold voice key delays reported in the literature (e.g., Rastle & Davis, 2002).

More important, in the present study, we also wanted to calculate a more sensitive measure of voice key performance. As pointed out earlier, voice key delays may not be such a big problem if they are constant across phonemes. Therefore, we also calculated association measures between voice key RTs and the benchmarking visual inspection RTs to see how closely less time-consuming voice key generated RTs may approach visual inspection. To our knowledge, such association measures have never been reported before in the literature for any type of voice key. It is important that the NEVK's RTs correlated highly with visual inspection RTs, respectively, .98 and .97 for the two independent judges (which showed a .97 interrater reliability), leaving not much room for improvement. Consistent with the large mean deviation and with the notion that the acoustic bias varies greatly across phonemes, the CTVK's RTs correlated only .78 with visual inspection RTs. This implies that only 61% of the variance in the actual RTs may still be explained by the measured RTs. This is quite disturbing, certainly because our CTVK even deviates less (72 ms) from visual inspection RTs than do similar devices tested in earlier studies (e.g., 138 ms for a simple threshold voice key; Rastle & Davis, 2002) and was therefore probably performing better.

Experiment 2: A Female Speaker

In this second extended test, we wanted to test whether the NEVK's performance is similar for other speakers and tasks.

[1] Because classical voice keys are subject to triggering delays up to 100 ms, studies using these devices often use a 0-ms criterion (e.g., Rastle & Davis, 2002). Because more accurate voice keys yield RTs that are much closer to the actual acoustic onset, these studies typically use a more conservative false alarm criterion such as 50 ms (e.g., Tyler et al., 2005). Because amplitude may be very low for certain phonemes in visual waveforms, a voice key RT that precedes human judgment by 10 ms, for example, may not always be a false alarm. Acoustic onset detection through visual waveform analysis may also be subject to minimal errors, especially for low amplitude sounds.

Table 1
Performance of the Noise Elimination Voice Key (NEVK) and a Classical Threshold Voice Key (CTVK) in Experiments 1 to 5

	Experiment 1		Experiment 2		Experiment 3		Experiment 4		Experiment 5	
Measure	NEVK	CTVK	NEVK	CTVK	NEVK	CTVK	NEVK	CTVK	NEVK	CTVK
Deviation (ms)[a]	5.60	72.33	4.11	74.61	5.30	45.65	4.69	22.53	5.34	84.01
Accuracy[b]	0.97	0.78	0.98	0.51	0.95	0.53	0.99	0.88	0.98	0.54

Note. ms = milliseconds.
[a] Reported values are mean absolute value deviation scores between voice key reactions times (RTs) and visual inspection RTs across judges (Experiments 1, 2, 4, and 5) and across participants (Experiment 3). [b]Reported values are mean Pearson correlations across judges (Experiments 1, 2, 4, and 5) and across participants (Experiment 3) between voice key RTs and visual inspection RTs.

Because some authors claim that voice keys may be differentially accurate for men versus women (e.g., Tyler, Tyler, & Burnham, 2005), we wanted to generalize our previous results with respect to gender. Therefore, the speaker in this test was a woman. Also, we used a speeded naming task instead of delayed naming because this is more commonly used in psycholinguistic research practice.

Method

Stimuli and Apparatus. All stimuli were the same as in Experiment 1. The apparatus setup was identical to that in the previous experiments.

Procedure. A 22-year-old woman, naïve to the purposes of the experiment, was instructed to perform a speeded naming task. On each trial, a fixation point was presented centered on the screen during 500 ms. This was replaced by the target, which had to be named as soon as possible. As soon as the two voice keys were triggered, the target disappeared. The intertrial interval was 1000 ms. Twenty practice trials were administered before the actual experiment. The entire experiment lasted about 40 min. All other aspects of the experiment were identical to Experiment 1.

Results

Results were analyzed according to the criteria used in Experiment 1. The false alarm rate for the NEVK (1.91%) was somewhat lower than for Experiment 1, whereas the CTVK again yielded no false alarms. As can be seen in Table 1, RT performance of the NEVK in this speeded naming task with a female speaker was very similar to performance in Experiment 1 with a male speaker performing a delayed naming task. Mean absolute deviation between the NEVK's RTs and the visual inspection RTs was even less this time, 4.1 ms. It is important that the NEVK's RTs again correlated highly with visual inspection RTs, .99 and .96, respectively, for the two independent judges (which mutually showed a .97 interrater reliability). As expected, the NEVK's performance was less susceptible to speaker differences than the CTVK's. For this speaker, the CTVK mean deviation scores increased only slightly to 74.61 ms, but the correlation with visual inspection RTs dropped to an alarming .51. Again, whereas this correlation is quite problematic, we would like to emphasize that our CTVK deviation scores are in line with values reported in the literature (e.g., Kessler et al., 2002; Pechmann et al., 1989; Rastle & Davis, 2002). Because our CTVK was therefore performing within typical experimental limits, these first reports of association measures between visual inspection RTs and threshold voice key RTs are very worrisome.

Finally, it may be interesting to have a look at the NEVK's and the CTVK's performance for different phoneme types. We combined the data from Experiment 1 (male speaker) and Experiment 2 (female speaker), and the results are displayed in Figure 2. Replicating earlier research (e.g., Pechmann et al., 1989; Rastle & Davis, 2002; Sakuma et al., 1997; Tyler et al., 2005), we found that detecting acoustic onsets is most difficult for fricatives, both for the NEVK and for the CTVK, although the former still performs very well for this category. For the NEVK, accuracy for approximants, nasals, plosives, and vocals is comparable. This is not the case for the CTVK, which performs almost as poorly on nasals and approximants as on fricatives.

Experiment 3: Group Experiment

Because of the strong amplification mechanism in the NEVK, triggering delays may not be affected as much by speaker differences (loudness, gender, etc.) as is the case for classical voice keys. This was confirmed in the previous two extended tests, which yielded similar results. However, we still wanted to run a demonstration test with a small group of participants in order to be able to generalize our findings across participants with greater confidence. Because of the fact that benchmarking voice key RTs with visual inspection of waveforms for multiple participants (by two judges) is very time consuming, this test contained far fewer trials than did the experiment described above. It is important that the NEVK also performs well under such conditions, because a psycholinguistic experiment may contain fewer trials than the extended experiments above. We believe that good performance on a smaller number of trials may therefore only add further strength to the NEVK's performance. Also, because of the high interrater reliability obtained in the previous two experiments and similar to Rastle and Davis (2002), visual inspection of waveforms was only done by one judge.

Method

Eight speakers (4 men and 4 women, naïve to the purposes of the experiment) participated in the experiment. The stimuli were the same as in the previous experiments, except that each participant named each target only twice, resulting in 44 trials. The apparatus setup and procedure were identical to those in Experiment 2 (speeded naming). The classical threshold voice key was individually calibrated for each participant in order to account for speaker loudness and clarity differences.

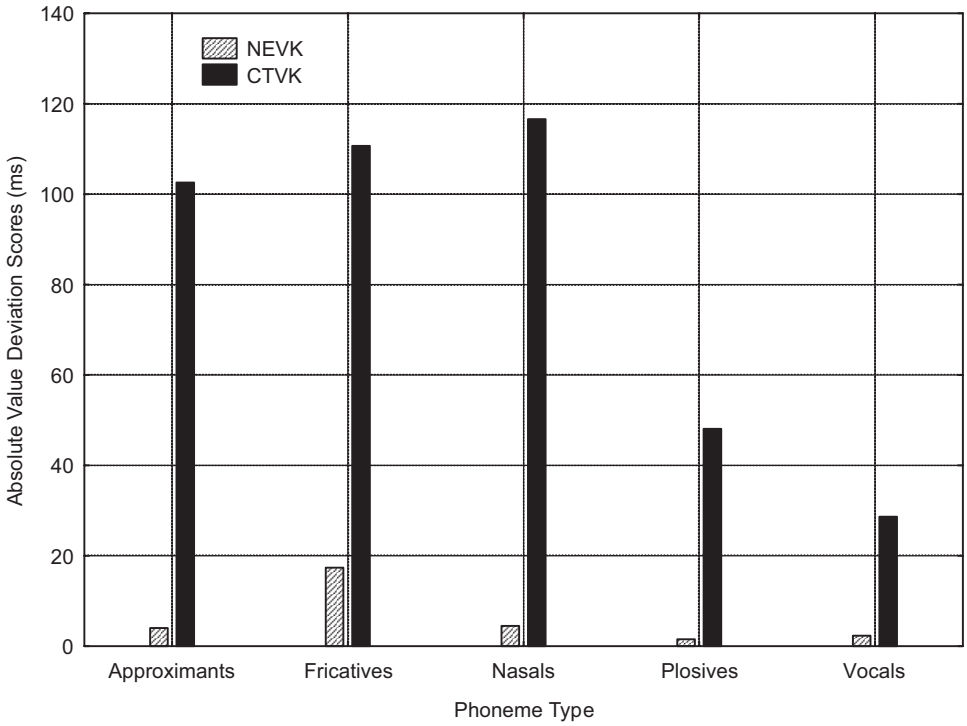

Figure 2. Mean absolute difference scores (Experiments 1 and 2) between voice key and visual inspection reaction times by voice-key type and phoneme type (approximants: *jaar*, *laag*, *raam*, *wijn*; fricatives: *fles, glas, hand, stad, voet, zaak*; nasals: *mate*, *naam*; plosives: *blik, deur, paar, tijd*; vocals: *arts, eeuw, idee, orde, unie*). NEVK = noise elimination voice key; CTVK = classical threshold voice key.

Results

Results were analyzed according to the criteria used in the previous experiments. False alarm rates for the NEVK and the CTVK were 3.98% and 0.28%, respectively. As can be seen in Table 1, RT performance of the NEVK in this experiment with very few trials and with 8 different speakers was very similar to performance in the previous two demonstrations. Mean absolute deviation between the NEVK's RTs and visual inspection RTs was 5.3 ms. More important, the NEVK's RTs correlated highly with visual inspection RTs. Mean Pearson correlation across participants was .95, with a small standard deviation ($SD = 0.02$). These results confirmed that the NEVK's performance is not very susceptible to speaker differences. Indeed, Pearson correlations only ranged from .99 to .91. Results for the CTVK were a bit mixed: Mean deviation was smaller than in the previous experiments, but mean correlation between visual inspection and the CTVK's RTs was only .53 ($SD = 0.27$), with a maximum of .89 and a minimum of only .17. Interestingly, the NEVK's performance for this worst CTVK participant was still .95.

Experiment 4: The Stimuli of Rastle and Davis (2002)

In the previous experiments, we used test stimuli that contained a wide range of onset phonemes so that the NEVK's performance in these tests would be representative for a typical naming experiment, which also often contains a lot of different onsets. However, as mentioned in the introduction, Rastle and Davis (2002) showed that acoustic biases may be especially large for complex onsets beginning with the /s/ phoneme. In the present test, we wanted to investigate the NEVK's performance in such a worst-case scenario by using exactly the same stimuli as Rastle and Davis (2002).

Method

The 40 test stimuli were 40 monosyllabic words beginning with the phoneme /s/, half of which contained two-phoneme onsets in which the second phoneme was /p/ or /t/ (e.g., *spat*) and half of which contained one phoneme onsets (e.g., *sat*). These stimuli were taken from Rastle and Davis (2002; see also Kawamoto & Kello, 1999). Each target was named 10 times, resulting in 400 trials. The apparatus setup and procedure were identical to those in Experiments 2 and 3 (speeded naming). Acoustic onset detection through visual inspection of waveforms was performed by the same two independent judges as in Experiments 1 and 2. The participant was a 29-year-old male speaker (Wouter Duyck).

Results

Results were analyzed according to the criteria used in the previous experiments. The false alarm rate for the NEVK (2.49%) was similar to that in the previous experiments. Again, the CTVK yielded almost no false alarms (0.75%). As can be seen in Table 1, RT performance of the NEVK in this naming task with /s/ onset stimuli was also very similar to performance in the previous experiments. Mean absolute deviation between the NEVK's RTs

and visual inspection RTs was 4.69 ms.[2] It is important that the NEVK's RTs again correlated highly with visual inspection RTs, .99 for two independent judges (which mutually showed a .996 interrater reliability). For this speaker, the CTVK mean deviation score (22.53 ms) and correlation (.88) were better than in the previous studies (note that the CTVK also yielded a .89 correlation for a single participant in Experiment 3). This proves that our CTVK, used as a benchmark for the NEVK, was not an atypically inaccurate or unrealistic device. It is also an illustration of the unstable and varying performance of such devices (see also the poor performance of the CTVK in the next experiment).

Experiment 5: Bisyllabic Stimuli With Noninitial Stress

In the experiments above, test stimuli were almost exclusively monosyllabic words. By definition, these words (but also the few tested polysyllabic words) always had stress on the onset syllable. In the naming of polysyllabic words, however, words that do not have stress on the initial syllable have lower onset amplitudes,[3] which may deter acoustic onset detection. In this experiment, the NEVK's performance was assessed for bisyllabic words with stress on the second syllable.

Method

Stimuli. The 22 Dutch target words to be named were selected from the CELEX lexical database (Baayen et al., 1993) using the WordGen stimulus selection program (Duyck et al., 2004). All targets were bisyllabic words with stress on the second syllable and were four or five letters long. Similar to Experiments 1 and 2, targets varied with respect to the initial phoneme, corresponding to a different beginning letter in order to assess the NEVK's performance across onset types. The resulting targets were as follows: *amok* [amuck], *beton* [concrete], *dieet* [diet], *enorm* [tremendous] (long /e/), *erna* [after] (short /ə/), *fobie* [phobia], *gazet* [newspaper], *hallo* [hello], *idee* [idea], *japon* [dress], *kopie* [copy], *libel* [dragonfly], *menu* [menu], *nabij* [close], *olijf* [olive], *pion* [pawn], *raket* [rocket], *salon* [drawing room], *tenue* [outfit], *uniek* [unique], *vanaf* [from], and *zopas* [just]. Each target was named 20 times, resulting in 440 trials.

Apparatus and Procedure. The apparatus setup and procedure (speeded naming) was identical to those used in Experiments 2, 3, and 4. The participant was the same speaker as in Experiment 4.

Results

Results were analyzed according to the criteria used in the previous experiments. The false alarm rate for the NEVK (6.12%) was somewhat higher than in the previous experiments. Again, the CTVK yielded no false alarms. As can be seen in Table 1, RT performance of the NEVK in this naming task with polysyllabic stimuli was very similar to performance in the previous experiments. Mean absolute deviation between the NEVK's RTs and visual inspection RTs was 5.34 ms.[4] It is important that the NEVK's RTs again correlated highly with visual inspection RTs, respectively, .98 and .99 for two independent judges (which mutually showed a .99 interrater reliability). The CTVK mean deviation score was very large (83.53 ms), similar to those in the first three experiments. The correlation between CTVK's RTs and visual inspection RTs was again alarmingly low (.54), which confirms that noninitial stress may indeed cause problems for classical voice keys (but not for the NEVK). Figure 3 shows CTVK and NEVK deviation scores by phoneme type. From this figure, it is clear that the NEVK's performance was not worse for polysyllabic words. NEVK's accuracy was similar to that in the previous experiments for all phoneme categories. For fricatives, for example, performance was even slightly better compared with that in Experiments 1 and 2 (Figure 2). For the CTVK, trigger delays were again especially large (around 100 ms) for fricatives, nasals, and approximants (see also Figure 2; Pechmann et al., 1989; Rastle & Davis, 2002; Sakuma et al., 1997; Tyler et al., 2005).

General Discussion

In the introduction, we discussed several problems that may arise when using a classical threshold voice key for the detection of acoustic onsets. The few studies that have been conducted on this matter have shown that voice keys may only trigger several tens of milliseconds after the speech onset is visible in recorded waveforms. Also, this delay varies greatly across onset phonemes (Kessler et al., 2002; Pechmann et al., 1989; Rastle et al., 2005; Rastle & Davis, 2002; see also Figures 2 and 3). Rastle and Davis (2002) have convincingly shown that this acoustic bias may be responsible for some contradictory results in the literature. Comparing naming speed of simple versus complex onsets, they found that the former may be significantly slower, faster, or equally fast compared with the latter, depending on the measuring technique (visual analysis or two different types of voice keys). On the basis of these issues, we have concluded that, if the use of alternatives for voice keys is not possible (see below), the ideal voice key should not yield large delays when detecting acoustic onsets, even for low-amplitude or slowly rising sounds. Second, the ideal voice key should not have operator-adjustable settings in order to achieve maximal comparability across participants and studies.

We believe that the voice key presented above meets these two requirements and may therefore constitute an important contribution to the recently unfolding voice key debate. First, because of its combination of an amplification mechanism with noise detection, the NEVK has an extremely low, built-in threshold, which does not need to be, and cannot be, tuned to the experimental setting. All parameters affecting the NEVK's functioning are hardware-implemented. That way, voice key performance is maximally comparable across studies and participants. Second, the experiments above have shown that the NEVK is very accurately detecting acoustic onsets, as soon as they are visible in recorded waveforms. Whereas the absolute deviation between visual inspection and classical voice key RTs was as high as 75 ms in some of the tests above (and even over 100 ms for specific phoneme categories), the NEVK's mean absolute deviation was between 4 and 6

[2] Note that performance on /s/ onsets was somewhat better than for the general fricatives category in Experiments 1 and 2. Of course, these experiments had different stimuli and speakers.

[3] We thank Kathy Rastle for this suggestion.

[4] English native speakers may be interested in a few English analogues of our polysyllabic Dutch stimuli. For instance, short NEVK tests (20 repetitions each) yielded average deviation scores of 4.6 ms and 3.8 ms for *ahead* and *aloud*, which is similar to the results obtained in the more extended Dutch tests.

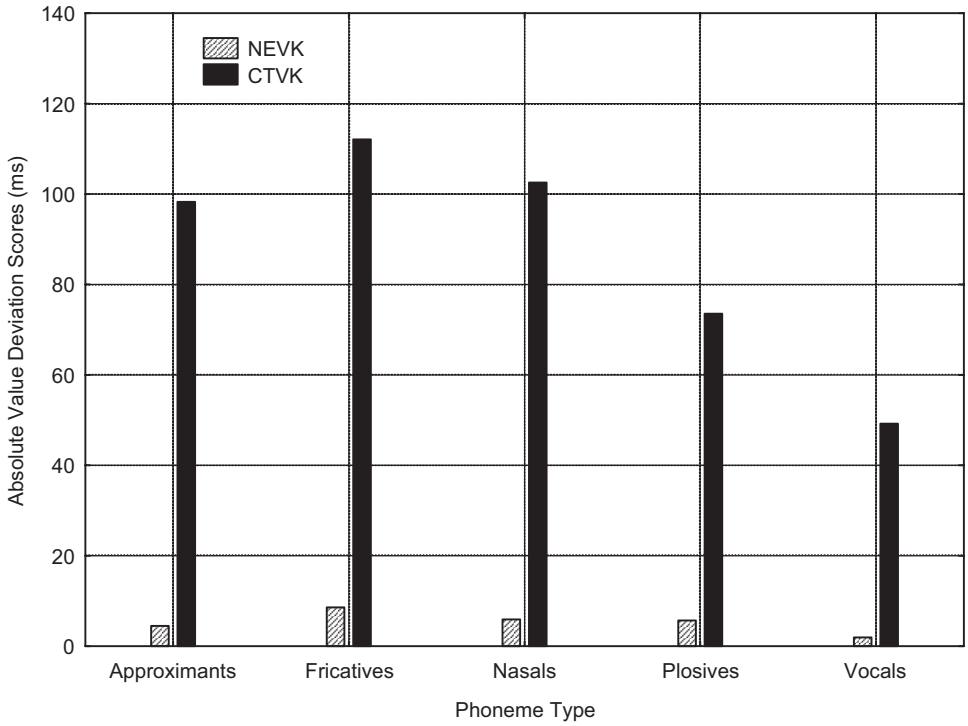

Figure 3. Mean absolute difference scores (Experiment 3—noninitial stress) between voice key and visual inspection reaction times by voice-key type and phoneme type (approximants: *japon*, *libel, raket*; fricatives: *fobie, gazet, hallo, salon, vanaf, zopas*; nasals: *menu, nabij*; plosives: *beton, dieet, pion, tenue*; vocals: *amok, enorm, erna, idee, olijf, uniek*). NEVK = noise elimination voice key; CTVK = classical threshold voice key.

ms. As pointed out earlier, large delays need not necessarily be a problem if they are constant across phonemes. Therefore, we also calculated correlations between visual inspection and voice RTs to see how well the latter may explain variance in the actual RTs. Correlations between the NEVK's RTs and visual inspection RTs were very high and ranged from .95 (Experiment 3) to .99 (Experiment 4). Moreover, this high level of performance was also observed for notoriously difficult stimuli, such as the /s/ onset stimuli of Rastle and Davis (2002; Experiment 4) or polysyllabic words with noninitial stress (Experiment 5). In contrast, the classical threshold voice key yielded correlations as low as .51 under typical experimental circumstances (measurements explaining less than 30% of variance in actual RTs). These correlations are quite alarming, given the fact that our threshold voice key's performance was not worse but rather comparable to reported values in the literature. For instance, our threshold voice key delays were often around 75 ms, which is less than the 135-ms and 81-ms delays for two types of voice keys reported by Rastle and Davis (2002). Specifically for /s/ onset stimuli (Experiment 4), our CTVK was even better (22-ms delay) than these other devices. This proves that the CTVK used in this study was not an atypical, unusually inaccurate voice key, but just as with all voice keys based on the threshold principle, its performance is highly unstable. For some stimuli/speakers, correlations with visual inspection RTs may drop from .88 (Experiment 4) to .54 (Experiment 5) or .51 (Experiment 2). Surprisingly, association measures of visual inspection and voice key RTs have never been reported before in the literature. We believe this alarming finding should encourage researchers to test their voice key's performance and consider an alternative if necessary.

In the following sections, we discuss the various alternative solutions that have been put forward in the voice key debate. First, there is a large consensus in the literature that visual wave form analysis of recorded spoken responses is probably the best way to avoid using a voice key. However, this approach is time consuming, especially when running experiments with a large number of trials and participants. As a work-around, some authors have used software algorithms that are designed to mimic human visual analysis (e.g., James, 1996). However, this has never been popular in research practice, probably because of the technical problems that may follow from this approach. For instance, both automated and manual visual analyses require high-quality waveform recordings with good equipment. Also, only a minority of stimulus presentation software packages allow precisely timed synchronization between stimulus presentation and wave recordings. Additionally, detecting acoustic onsets in a waveform by both humans and algorithms may not always be that straightforward, as some utterances yield minimal amplitude fluctuations that are hard to detect within the noise. Onsets that are difficult to detect for voice keys are often also more difficult to code through visual analysis. Note, however, that for some research questions, visual (and auditory) analyses may be the only alternative. This is true when one wants to examine the time course or onsets of speech that is preceded by other speech. For instance, if one wants to examine processes of speech error repair, one will need to manually identify the onset of the repair speech in recordings of the continuous

sound stream (e.g., Hartsuiker, Pickering, & De Jong, 2005). In such cases, a voice key will probably only trigger on the initial acoustic onset and will not be able to detect the onset of the second part of the utterance.

A second appealing technique would be to determine acoustic biases for all possible onsets, as Kessler et al. (2002) did, and to correct all RTs with these constants. The problem with this technique is, of course, that acoustic biases may vary greatly across participants and studies, depending on peripheral factors such as background noise, speaker loudness, or type of equipment used (see the comparability issue discussed above).

A third popular technique to deal with voice key inaccuracy related to the above is the use of delayed naming tasks (e.g., Duyck & Brysbaert, 2004). The reasoning behind this approach is that subtracting delayed naming RTs from speeded naming RTs yields residuals that only reflect cognitive processing, with acoustic biases in both tasks outweighing each other. The advantage of this technique is that the acoustic biases are not conceived as universal but are determined within studies and participants. However, as Kessler et al. (2002) noted, the core assumption that acoustic biases are equivalent across naming conditions may be wrong. Participants may respond more loudly, softly, quickly, or slowly in speeded versus delayed naming, such that one set of biases is traded for another. Also, Goldinger, Azuma, Abramson, and Jain (1997) showed that RTs in naming tasks that are delayed up to 400 ms are influenced, for example, by word frequency. So, delayed naming subtraction not only may erase acoustic biases but may also wipe out RT variance due to cognitive processing. For these reasons, as noted by Rastle and Davis (2002), the use of delayed naming has largely been abandoned.

A fourth solution that has been put forward is initial phoneme matching. The idea behind this approach is to distribute the acoustic bias evenly across conditions so that the bias adds noise to the data but doesn't interact with the experimental manipulation. Whereas this may seem an efficient way to circumvent the voice key delay problem, the survey of Kessler et al. (2002) showed that only half (56%) of a large sample of selected studies matched the initial phoneme between experimental conditions, even in the absence of other remedies. This is probably because applying initial phoneme restrictions may sometimes be virtually impossible in combination with matching on many other variables. Such complicated matching procedures often lead to small and exhaustive sets of stimuli on their own. Also, because the second phoneme affects the voice key's triggering delay (see above; e.g., Kessler et al., 2002), Rastle and Davis (2002) rightly argued that it is not sufficient to match initial phonemes across experimental conditions. Instead, the entire syllabic onset should be matched (i.e., "all phonological segments preceding the vowel"; Rastle & Davis, 2002, p. 313). More recently, Rastle et al. (2005) suggested that onset matching should be even more extended (e.g., including also the vowel). However, in the same survey of Kessler et al. (2002), only 45% of respondents were actually aware that noninitial phonemes may affect voice key accuracy. Finally, even though adequate and extensive phoneme matching would prevent confounding between the experimental manipulation and acoustic bias and therefore minimize chances of obtaining statistical Type I errors, it would still add noise to the data that is not accounted for by experimental factors in the design. Therefore, even with perfect matching, this approach is likely to yield an inflated number of statistical Type II errors.

From the above, it may be concluded that there is currently no fully satisfactory solution for the acoustic bias problem. Therefore, we believe that a good solution may be provided by a more technical approach. Indeed, there have been a few promising precursors in the literature that point to this direction. First, in the study of Rastle and Davis (2002), an integrator voice key was proposed. This voice key takes into account both the amplitude and duration of sounds and outperformed their benchmarking simple threshold voice key. However, its RTs were still delayed 81 ms (across onset types) relative to visual inspection. Also, an additional analysis of their data showed that RTs generated by the integrator voice key correlated .70 (averages across participants) with visual inspection RTs.[5] This was much higher than the .51 correlation for the DMDX voice key (Forster & Forster, 2003) but considerably lower than the correlations (up to .99) obtained in this study, even with exactly the same stimuli (Experiment 4). A second elegant technical solution has been suggested by Tyler et al. (2005). Following the software algorithm of James (1996) mentioned above, and similar to the integrator voice key, they recently proposed an improved analogue voice key that is basically a standard threshold voice key with minimum signal and silence duration settings. Because this voice key also uses a certain timeframe to decide whether sound is speech or not, they labeled this device as a delayed trigger voice key (DTVK). The powerful DTVK outperforms classical voice keys, but triggering is still (significantly) 11.8 ms delayed relative (Tyler et al., 2005, Experiment 1; excluding false alarms, same high-gain condition) to visual inspection RTs.[6] This is about twice as much as the mean absolute value deviations from 4 to 6 ms obtained in this study (without highly inflated false alarm rates, see earlier). Also, Tyler et al. (2005) obtained this level of performance with high-quality sound files recorded by very high-quality equipment on DAT tape, played by a PC and fed directly into the DTVK's circuitry. Hence it is likely that a real-life test (like the ones we conducted) in which the spoken responses are actually produced in a regular lab room and acoustic energy has to be recorded by a voice key microphone (instead of electrical signal input) would yield worse results. Finally, the DTVK's minimum silence duration, signal duration, and gain levels still need to be determined by the operator. As noted before, this yields results that are hard to compare across studies and participants. This is actually illustrated in the article of Tyler et al. (2005) itself. In their first experiment, the DTVK's performance was tested with a minimum signal duration setting of 100 ms (yielding the 11.8-ms delay mentioned above). However, in their second experiment, this same parameter setting yielded a

[5] We would like to thank Kathy Rastle for providing us with the data necessary to calculate these correlations.

[6] In Tyler et al. (2005), performance of the DTVK is expressed as mean hand-coding advantage (HCAdv) scores, defined as the difference between the DTVK RTs and visual coding RTs. However, because their false alarm criterion was set to 50 ms, these HCAdv scores may also be negative values (if the DTVK triggered before [but no more than 50 ms] the onset was visible in the waveform). Consequently, a mean HCAdv score of zero does not necessarily mean perfect performance, as positive deviations (late triggers) may be compensated by negative deviations (early triggers). Therefore, we believe it is more appropriate to use mean absolute value deviation scores instead. These values are reported here. We thank Michael Tyler for providing us with the data necessary to calculate these values for the DTVK.

23-ms simple onset advantage when replicating the study of Rastle and Davis (2002) described above. This 23-ms facilitation effect is inconsistent with the significant simple onset disadvantage that was reported by Rastle and Davis (2002) using visual analysis (9-ms inhibition effect). It is also inconsistent with the onset disadvantage reported in the Tyler et al. study using the same DVTK device with a different signal duration setting of 50 ms and using visual analysis (both 8-ms inhibition effects). It is hard to see how the recommended 50-ms parameter setting for their second experiment (and not 100 ms as used in their first study) may be convincingly motivated a priori without actually doing the visual waveform analysis (which, of course, makes the use of a voice key redundant). To conclude, we believe the voice key presented in this article constitutes a significant step forward in reaching a technical solution for the acoustic bias problem.

It may be important to point out one remaining issue with respect to the voice key solution presented above. A problem that still applies to the improved voice key presented in this article concerns the distinction between acoustic onset and the onset of articulation (for a more detailed discussion of this issue, see Rastle et al., 2005). For some sounds (e.g., fricatives), acoustic and articulation onset occur virtually simultaneously. In contrast, for other sounds (e.g., plosives and affricatives) pressure may need to build up gradually before it is released, so that the acoustic onset is by definition systematically delayed relative to the articulatory onset. So, if one condition in an experiment contains more plosives than another condition with more fricatives, the RTs generated by a voice key (even if it were 100% accurate) in that condition may be systematically biased (delayed) relative to the cognitive processes behind the articulation (the processing of the stimulus). This issue is not only problematic for voice keys but also for the other RT measurement approaches discussed above. For instance, because the acoustic biases for different onsets reported in the study of Kessler et al. (2002) were obtained by regressing the initial phoneme on naming latencies, these biases do not only reflect voice key inaccuracy. They also include the varying time window between articulation onset and acoustic onset. Finally, note that this problem also cannot be solved by visual waveform analysis, as this technique also only depends on actual acoustic energy visible in recordings.

To conclude, we hope that this article increases researchers' awareness about the problems that using voice keys may involve. In the literature, there is a consensus that there is currently no work-around (e.g., delayed naming, matching) that provides a fully satisfactory solution to the acoustic bias problem, which leaves the field with an open question. We have argued that such a solution might be of a technical nature, such as the amplification voice key presented in this article, which yields RTs that approach visual inspection RTs very well.

References

Baayen, R., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical database* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium.

Bovens, N., & Brysbaert, M. (1990). Ibm-Pc/Xt/at and Ps/2 Turbo Pascal timing with extended resolution. *Behavior Research Methods Instruments & Computers, 22,* 332–334.

Donders, F. C. (1868). Over de snelheid van psychische processen. Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtse Hoogeschool [About the speed of mental processes. Investigations carried out in the Physiologisch Laboratorium der Utrechtse Hoogeschool]. *Tweede Reeks, 2,* 92–110.

Duyck, W., & Brysbaert, M. (2004). Forward and backward number translation requires conceptual mediation in both balanced and unbalanced bilinguals. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 889–906.

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). Wordgen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods Instruments & Computers, 36,* 488–499.

Forster, K. I., & Forster, J. C. (2003). A windows display program with millisecond accuracy. *Behavior Research Methods Instruments & Computers, 35,* 116–124.

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 361–379.

Goldinger, S. D., Azuma, T., Abramson, M., & Jain, P. (1997). Open wide and say "blah!" Attentional dynamics of delayed naming. *Journal of Memory and Language, 37,* 190–216.

Hartsuiker, R. J., Pickering, M. J., & De Jong, N. H. (2005). Semantic and phonological context effects in speech error repair. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31,* 921–932.

James, C. (1996). A vocal response time system for use with sentence verification tasks. *Behavior Research Methods Instruments & Computers, 28,* 67–75.

Kawamoto, A. H., & Kello, C. T. (1999). Effect of onset cluster complexity in speeded naming: A test of rule-based approaches. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 361–375.

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language, 47,* 145–171.

Pechmann, T., Reetz, H., &. Zerbst, D. (1989). Kritik einer Messmethode: Zur Ungenauigkeit von Voice-key Messungen [Critique on a measurement method: About the inaccuracy of voice-key measurements]. *Sprache & Kognition, 8,* 65–71.

Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1083–1095.

Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 307–314.

Sakuma, N., Fushimi, T., & Tatsumi, I. (1997). Measurement of naming latency of Kana characters and words based on speech analysis: Manner of articulation of a word-initial phoneme considerably affects naming latency. *Japanese Journal of Neuropsychology, 13,* 126–136.

Tyler, M. D., Tyler, L., & Burnham, D. K. (2005). The Delayed Trigger Voice Key: An improved analogue voice key for psycholinguistic research. *Behavior Research Methods, 37,* 139–147.

Yamada, J., & Tamaoka, K. (2003). Measurement errors in voice-key naming latency for Hiragana. *Perceptual and Motor Skills, 97,* 1100–1106.

Received September 27, 2006
Revision received February 1, 2007
Accepted February 4, 2007 ■