Effect of Voice Onset Type on Vocal Attack Time

*Ben C. Watson, *R. J. Baken, and †Rick M. Roark, *†Valhalla, New York

Summary: Vocal attack time (VAT) is the time lag between the growth of sound pressure (SP) and electroglottographic (EGG) signals at vocal initiation. The characteristics of voice initiation are associated with issues of vocal hygiene, efficiency, and quality. Vocal onsets have commonly been qualitatively characterized into three types: hard, simultaneous, and breathy. This study examines the effect of voice onset type on VAT values in normal speakers. SP and EGG recordings were obtained for 55 female and 57 male subjects while producing multiple tokens of three tasks (sustained /ɑ/ and "always" as unaspirated onsets, and "hallways" as an aspirated onset). Results revealed a significant effect of onset type on VAT, with the mean VAT for the "hallways" (aspirated) task greater than the mean VAT for the sustained /ɑ/ and "always" (unaspirated) tasks. There was no significant VAT difference between the sustained /ɑ/ and "always" tasks. Findings confirm the sensitivity of the VAT measure to vocal onset type and suggest its potential application as an objective and quantitative clinical measure of the type of vocal onset.

Key Words: Voice onset–Vocal attack time–Phonation–Vocal fold physiology.

INTRODUCTION

Voice researchers, teachers, and clinicians have long identified voice initiation as an important aspect of vocal performance and have tied it to issues of vocal hygiene,[1,2] efficiency,[3,4] and quality.[5–7] Indeed, the speed with which the vocal folds adduct to the midline is considered an important variable in the etiology of some voice disorders and a meaningful indicator of neural dysfunction.[8–14] It is generally accepted that the initiation of phonation includes two somewhat distinct phases: (1) the prephonatory adjustment phase that is associated with setting the appropriate tension, gross adduction, and aerodynamic forces[15]; and (2) the attack phase that is associated with the onset of vocal fold oscillation and sound generation. In one of the first voice studies to use ultrahigh-speed cinematography, Moore[16] addressed the issue of "the movements of the larynx leading to and including the initiation of tone." Moore described the observed differences between what he termed the "glottal stroke" or "glottal shock" and the "breathed" and "simultaneous" vocal attacks. Based largely on an auditory-perceptual categorization, voice onset continues to be popularly divided into abrupt or hard, normal or simultaneous, and breathy, aspirate, or soft glottal attack.[17,18]

Baken and Orlikoff[19,20] proposed that the time delay between the rise of the sound pressure (SP) and electroglottographic (EGG) signals is related to characteristics of the attack gesture. Specifically, a lag in the rise of the SP signal relative to the EGG signal may correspond to a hard glottal onset while a lead in the rise time of the SP signal may correspond to an aspirate onset. The inter-signal lag, termed vocal attack time (VAT), can be automatically and objectively estimated using a cross-correlation method and is, therefore, free of operator/investigator bias and imprecision. Orlikoff et al[21] validated the VAT measure against high-speed videoendoscopy, from which a digital kymogram was generated, and showed that "breathy" onsets have positive VAT values, whereas "hard" onsets have negative VAT values. Roark et al[22] and Ma et al[23] reported normative VAT values for English and Cantonese speakers, respectively, in tasks for which voice onset was not constrained.

The present study builds on the work of Roark et al[22] to examine effects of aspirate voice onsets on VAT values in normal speakers. Specifically, we compare VAT values for tasks that can be produced with either hard or simultaneous onsets with a task that requires an aspirate initiation.

SUBJECTS

This study was approved by the Committee for the Protection of Human Subjects (Institutional Review Board) of New York Medical College, and all subjects provided informed consent before data collection began. The subject sample for the present study is the same sample previously described in Roark et al[22] and included 55 females (mean age: 28 years; range: 22–50 years) and 57 males (mean age: 29 years; range: 21–50 years). A health history that included chronic speech, voice, cardiac, respiratory, or neurologic disorder was cause for exclusion from the study. A current health status that included upper respiratory infection was cause for a delay in participation until the infection resolved.

METHODS

Simultaneous acquisition of SP and EGG signals was achieved by use of a headphone-mounted microphone (model 33–3012; RadioShack, Ft. Worth, TX) positioned approximately 5 cm in front of the participant's lips at approximately 15° off midline and an EGG (Glottal Enterprises EG2; Syracuse, NY), respectively. The SP and EGG signals were routed through an amplifier and digitizer (M-Audio Fast Track; Irwindale, CA) and digitized (44.1 ksamples/second; 16-bit resolution) to a laptop computer for storage and subsequent analysis. Signal acquisition used Audacity software (v.3, General Public License, http://audacity.sourceforge.net). Subjects sustained a vowel at comfortable pitch and loudness while gain was adjusted to

Accepted for publication December 8, 2014.

From the *Department of Speech-Language Pathology, New York Medical College, Valhalla, New York; and the †Department of Otolaryngology, New York Medical College, Valhalla, New York.

Address correspondence and reprint requests to Ben C. Watson, School of Health Sciences and Practice, New York Medical College, 30 Plaza West, Valhalla, NY 10595. E-mail: Watson@nymc.edu

Journal of Voice, Vol. 30, No. 1, pp. 11-14
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.12.004

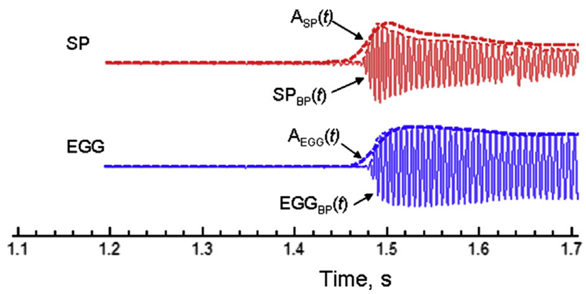

FIGURE 1. Smoothed amplitude functions of the bandpass-filtered sound pressure (SP) and electroglottographic (EGG) signals, yielding A_{SP} (t) and A_{EGG} (t) respectively.

achieve adequate amplification without clipping while signals were monitored on a two-channel oscilloscope.

The stimuli "always" and "ahh" (to represent the sustained vowel, /ɑ/) were used to elicit unconstrained, albeit typically unaspirated, voice onsets while the task "hallways" was used to elicit a specifically aspirate onset. The stimuli were printed on individual 3 × 5 inch index cards (five cards per task). The experimenter shuffled the cards and placed them face-down in front of the subject. The subject was instructed to turn over each card, silently read the task item, and then to say the word at a comfortable pitch, loudness, and rate or to sustain the vowel at a comfortable pitch and loudness for approximately 2 seconds. Subjects were instructed that this was not a reaction time task and to pause for several seconds between cards. EGG and SP signals were recorded continuously until the subject completed all 15 cards. The cards were then reshuffled, and one or two more trials were recorded. VAT data for the vowel task were originally reported by Roark et al.[22]

SP and EGG signals were stored as stereo wave (.wav) files and analyzed using custom algorithms implemented in *MATLAB* (MathWorks, Natick, MA). The signal analysis process is described in Roark et al[22] in sufficient detail to replicate this method and is summarized briefly here.

The data analysis process consists of four steps, each of which includes graphical display and optional acoustic playback. Analysis was allowed to progress to the next stage only if graphic display and/or acoustic playback met criteria for acceptable data quality (eg, production of the correct task):

Step 1, signal verification: signals could be zoomed and panned and/or played acoustically through external speakers to verify accuracy of task target and fidelity of signal quality;

Step 2, signal segmentation: automated identification of a 600 milliseconds segment of the SP and EGG signals centered at the approximate time of vocal offset;

Step 3, F_0-based frequency filtering and signal modeling: automated extraction of a representative value of fundamental frequency (F_0) of the raw EGG signal and band-pass filtering the 600 milliseconds segment at ±40% of the F_0 value. The analytic signal models of the band-pass–filtered SP signal and EGG signal are as follows:

$$X_{BP}(t) = A_X(t)\sin\left(2\pi f_x(t)t\right), \tag{1}$$

where X is SP or EGG, and $A_X(t)$ is the instantaneous amplitude and $f_X(t)$ is the instantaneous frequency of the modeled band-pass–filtered signals (Figure 1);

Step 4, extraction of measures: the instantaneous amplitudes [$A_X(t)$, where X is SP or EGG] were smoothed and difference functions [$A_X\prime(t) = A_X(t) - A_X(t - \Delta t)$, where X is SP or EGG and $\Delta t = 10$ milliseconds] were derived and normalized to represent local variation in the SP and EGG amplitude models. The normalized cross-correlation of the smoothed difference functions is shown in Figure 2. Here, the peak of the correlation occurs at 9.98 milliseconds, which represents the VAT for this token of "hallways."

VAT measures were grouped by sex and task ("always", /ɑ/ , "hallways") and analyzed using a mixed model with sex and task entered as fixed effects. A compound symmetric covariate structure was used for the within subjects repeated measure (token).

RESULTS

Following Roark et al,[24] the linear correlation coefficient, Pearson *r*, computed for the smoothed SP and EGG difference functions was used as the figure of merit (FOM) for measures of VAT. An *a priori* criterion FOM of greater than 0.75 applied to the database resulted in the rejection of 32 (2.8%) of the 1133 tokens for /ɑ/, 41 of the 1086 tokens (3.7%) for "always," and 95 of the tokens 971 (9.7%) for "hallways." Increased variability in the production of aspirate onsets likely led to more frequent violation of assumptions underlying our FOM criterion. There was no discernable pattern to the rejection of these

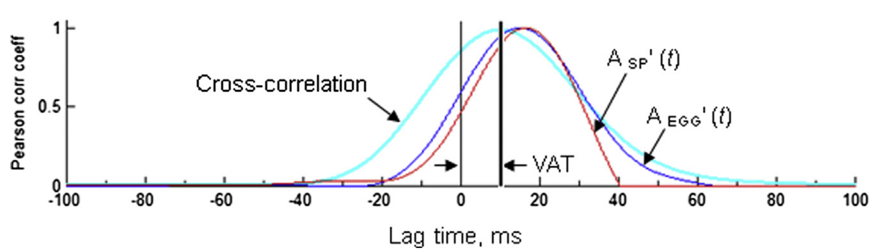

FIGURE 2. Difference functions, A_{SP}' (t) and A_{EGG}' (t) and the cross-correlation function. The time lag at the peak of cross-correlation provided the vocal attack time, VAT.