INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automatic Discrimination of Soft Voice Onset Using Acoustic Features of Breathy Voicing

Keiko Ochi[12]*, Koichi Mori*[2]*, Naomi Sakai*[2]*, Nobutaka Ono*[13]

[1]National Institute of Informatics, Tokyo, Japan
[2]Research Institute, National Rehabilitation Center for Persons with Disabilities, Tokorozawa, Japan
[3]SOKENDAI (The Graduate University for Advanced Studies), Hayama, Japan
ochi@nii.ac.jp, mori-koichi@rehab.go.jp, sakai-naomi@rehab.go.jp, onono@nii.ac.jp

Abstract

Soft onset vocalization is used in certain speech therapies. However, it is not easy to practice it at home because the acoustical evaluation itself needs training. It would be helpful for speech patients to get objective feedback during training. In this paper, new parameters for identifying soft onset with high accuracy are described. One of the parameters measures an aspect of the soft voice onset, in which the vocal folds start to oscillate periodically before coming in contact with each other at the beginning of vocalization. Combined with an onset time exceeding a threshold, the proposed parameters gave about 99% accuracy in identifying soft onset vocalization.
Index Terms: Acoustic analysis, soft voice onset, speech therapy, stuttering

1. Introduction

Voice onset is classified into soft and hard depending on the coordination of the laryngeal and respiratory muscles. Soft voice onset is generally used in speech therapy for stuttering, a fluency disorder characterized by the repetition and prolongation of sound and silent blocking [1]. The abnormal laryngeal activity of people who stutter during phonation has been reported [2] and soft onset is useful for countering speech blocking. Soft voice onset is also used in therapy for vocal disorders [3] and for the prevention of vocal cord nodules [4] to eliminate hard attacks of the vocal folds.

It is necessary to practice regularly to acquire a new speech pattern such as soft voice onset. However, patients have difficulty in evaluating their own production. Home-based training would be easier if the objective rating of the onset was fed back to patients. Fig. 1 shows a conceptual diagram of speech therapy assisted by a computer. The patients practice soft voice onset using a computer at home after being trained by a therapist in a therapy room. The effectiveness of speech therapy assisted by a computer has been demonstrated in an extensive investigation [5].

Peters et al. suggested measuring the logarithm of the time needed for the amplitude envelope to increase from 10% to 90% of the maximum level to judge the onset of a single-vowel phonation [6]. However, it is difficult to accurately determine the point corresponding to 90% when a word or sentence is uttered. Moreover,such mesurement would lead to unnaturally elongated syllables because patients would be trained to raise their voice slowly if a long onset was evaluated as better. Orlikoff et al. used the provided vocal attack time (VAT) to objectively measure the onset of phonation from sound and electroglottograph (EGG) signals [8]. However it is impractical to use an EGG at home-based training because it would be too expensive for most patients to purchase.

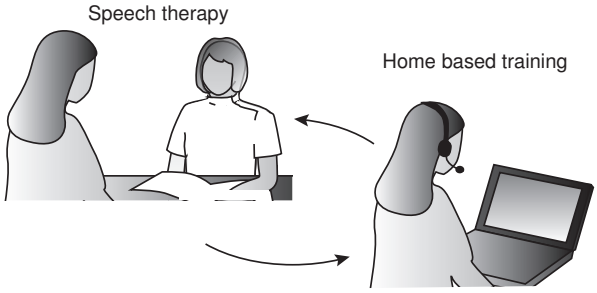


Figure 1: *Speech therapy and analogous home-based training assisted by a computer.*

In this paper, we propose new parameters to discriminate soft and hard onset obtained only from acoustic speech signals acquired from a microphone. Using some acoustic characteristics of soft voice onset, it is possible to measure the softness independently from the prolongation of a syllable.

2. Definition of soft onset

Observation using a digital kymograph revealed that the glottis starts to oscillate before the vocal folds come in contact at the beginning of a "breathy" phonation [8]. The soft onset defined in this paper is based on this observation.

In this paper, we define the soft onset voice as the onset of phonation with the glottis open immediately before and at the moment of the initiation of phonation, and define the hard onset as an abrupt phonation onset with the glottis closed.

A small, almost sinusoidal sound corresponding to the oscillation of the glottis appears prior to the contact of the vocal folds [8]. An abrupt increase in volume is observed in a hard voice onset, on the other hand, corresponding to small openings in the glottis.

3. Soft voice onset in speech therapy

When a speech therapist discriminate soft and hard onset by auditory impression, they should consider the rate of increase in the volume. Peters et al. showed that the logarithm of the time for the amplitude envelope to increase from 10% to 90% of the eventual maximum level was highly correlated with the perceived abruptness of the onset [6].

To introduce soft voice onset to patients, therapists tend to

http://dx.doi.org/10.21437/Interspeech.2016-765

teach their patients to increase the volume slowly. Mallard et al. asked subjects to phonate a vowel at a rate slower than 52.3 dB/s [7] to produce soft onset voices.

Whereas prolonged soft onset is easy to discriminate by human hearing, patients should be trained to avoid prolonged phonation to ensure natural speech. Since no prolongation of onset is ideal even for soft onset, the boundary between soft and hard voice onsets is not easily defined if only the rate of increase in the volume is used to judge the softness of the onset.

4. Acoustic features for discriminating soft/hard onset voice

4.1. Conventional features: rise time

In the conventional methods, the softness of the onset is judged by the time required to increase the volume. We have examined two types of rise time measures: the rise time defined by Peters et al [6] (rise time 1 (RT1)) and the modified version of the rise time defined by Koike [10] (rise time 2 (RT2)).

RT1 is defined as the logarithm of the time in which the volume envelope increases from 10% to 90% of the maximum value. The amplitude envelope is the root mean square (RMS) of the sound signal calculated with a window length of 40 ms and a sliding window step of 2.5 ms. The RMS value of the kth frame from signal $x[i]$ is calculated using

$$m[k] = \sqrt{\frac{1}{N}\sum_{i=ks+1}^{ks+N} x^2[i]}, \tag{1}$$

$$\text{RT1} = (k_2 - k_1)s, \tag{2}$$

where k_1 and k_2 are the first frame indices that satisfy

$$m[k_1] \geq 0.1 \max_k m[k], \tag{3}$$

$$m[k_2] \geq 0.9 \max_k m[k], \tag{4}$$

respectively, s is the window step, and N is the length of the window.

RT2 is defined as the time when the amplitude envelope reaches its mean amplitude. Although Koike defined the rise time as the time required for the volume to rise to the mean level of the steady portion, we used the mean amplitude to calculate the parameter automatically for the interval while the RMS exceeded a threashold. For the calculation of the RT2, the RMS window length was 100 ms and the window step was 1 ms.

4.2. Proposed acoustic features of breathy voicing

Hillenbrand et al. [9] showed that the signal periodicity is highly correlated with the perceptual breathiness, which is caused by the insufficient closure of the glottis. Because the onset of soft voicing starts with oscillation without the vocal folds of the glottis coming in contact, it is expected that the periodicity will be correlated with the softness of the voice onset.

In this study, we define two acoustic features related to the periodicity at the onset of soft voicing and utilize them to discriminate soft and hard onset.

One feature is the time lag (dT), which is the delay period from the time when the signal periodicity is observed to the time when the volume becomes large. It is represented by

$$\text{dT} = T_l - T_p, \tag{5}$$

where T_l is the time when the RMS level reaches a threshold and T_p is the time when the signal periodicity exceeds a threshold. The periodicity can be measured from the peak value in the autocorrelation. Because the small oscillation starts prior to the contact of the vocal folds of the glottis, a small periodic signal is assumed to be observed before the volume rises above the set threshold, which makes the time lag between T_l and T_p.

The other feature is the difference between the amplitudes of the first harmonic and second harmonic in the log domain as follows:

$$\text{H1-H2} = 20\log_{10}\frac{a(2F_0)}{a(F_0)}\ [\text{dB}], \tag{6}$$

where $a(f)$ is the amplitude spectrum of the waveform and F_0 is the fundamental frequency. We refer to this difference as H1-H2 hearafter [11][12]. This measure is essentially the same as the "first harmonic amplitude" described in [9].

5. Experimental Evaluations

5.1. Speech samples

The speech samples used in this study were collected in a sound-attenuated chamber at a digitization rate of 48 kHz with 16-bit quantization levels. Four Japanese speech therapists with more than one year of experience in stuttering therapy were asked to phonate the five Japanese vowels (/a/, /i/, /u/, /e/, and /o/) and 10 to 16 words that began with the five vowels two to three times with very soft, soft, hard, and normal onsets.

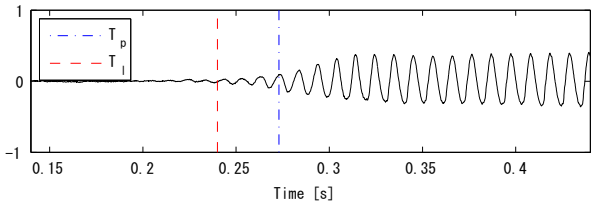
Very soft voice onset is phonated with the volume rising extremely slowly. It is used at the beginning of therapy because of its easiness to practice, although it is not suitable for daily use because of its unnaturalness. The difference between soft and very soft onset is the degree of the softness, which varied from person to person. Becasse there is an overlap of the distributions of soft and very soft onsets, we discriminate only soft and hard onsets in this study by regarding soft and very soft onset as being included in the same soft onset class.

Normal onset is phonated as the speaker produces without any conscious effort to achieve soft or hard onset and varies with the speaker. It was not included in the analysis.

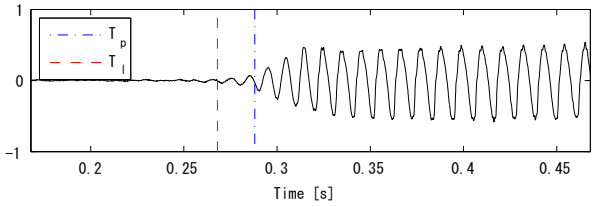
216 speech samples of soft and hard onsets and 175 samples of very soft onset were used for the discrimination, which were downsampled to 16 kHz and analyzed.

5.2. Methods

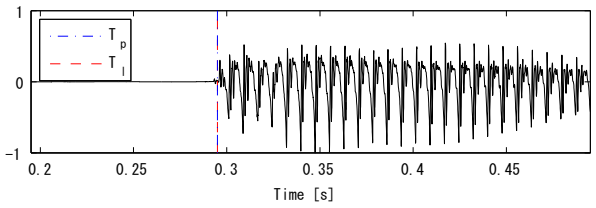
Acoustic analysis was undertaken to obtain the parameters to discriminate soft and hard onset voices. We examined the two types of rise time (RT1 and RT2) and the proposed features (H1-H2 and dT). Praat [13] [14] estimated the F_0 values and T_p, where T_p was the first time point where the F_0 value was extracted. The RMS window step for T_l was 1 ms and the window duration was 10 ms. At the earlier time of T_l and T_p, the amplitude spectrum was calculated after zero padding to 2048 points to increase the frequency resolusion using MATLAB version R2013b. Then, the harmonic amplitude of first and second harmonics was approximated by the amplitude at the nearest discrete frequeny to F_0 and $2F_0$, respectively, where continuous-valued F_0 was determined at time T_p. Finally, H1-H2 was computed by eq.(6).



(a)

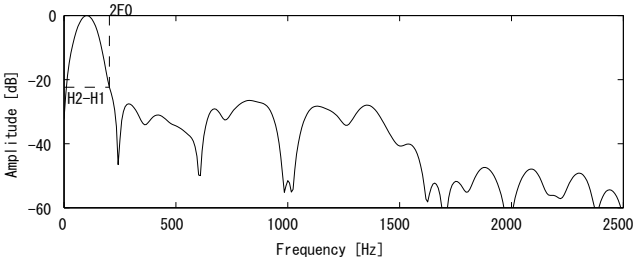


(b)

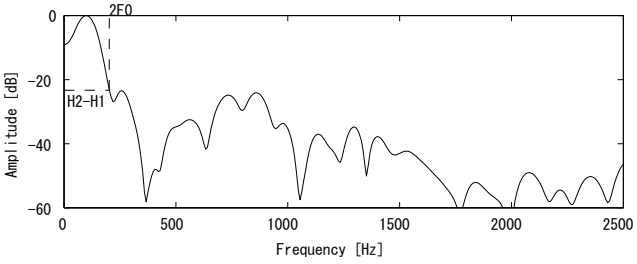


(c)

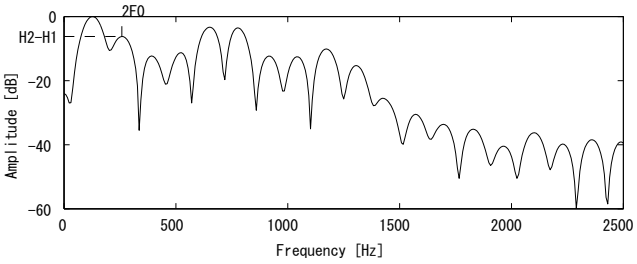
Figure 2: *Waveforms of (a) very soft, (b) soft and (c) hard voice onsets.*



(a)

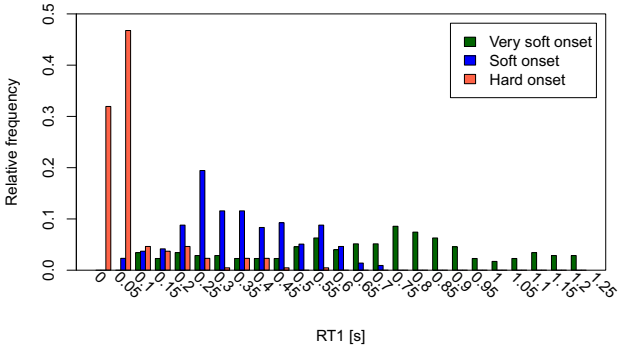


(b)

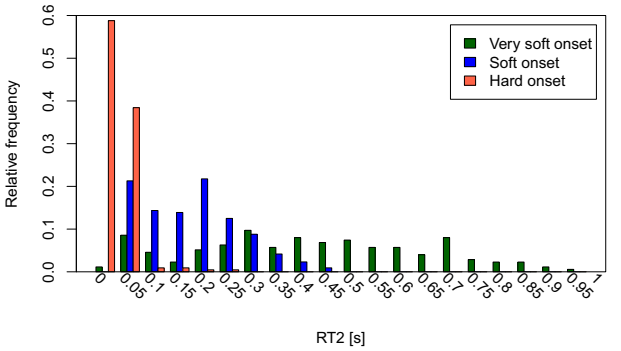


(c)

Figure 3: *Spectra of (a) very soft, (b) soft and (c) hard voice onsets.*



(a)



(b)

Figure 4: *Histgrams of (a) Rise time 1 and (b) Rise time 2.*

We applied support vector machine (SVM) to discriminate soft and hard onset using the kernlab package in the statistical environment R [15]. We used Gaussian radial basis kernel defined as

$$k(x, x') = \exp(-\sigma||x - x'||^2), \tag{7}$$

where x and x' are one or two-dimensional feature vectors and σ was experimentaly determined as 0.1. Leave-one-speaker-out cross-validation was applied to assess the accuracy of discrimination of the samples.

6. Results

Fig. 2 shows examples of waveforms of very soft, soft and hard onset voices of vowel /a/. Because a small near sinusoidal oscillation was observed before the volume rises at the beginning of the soft and very soft onset, T_p is less than T_l. The interval between T_p and T_l of very soft onset is larger than that of the soft onset because the small sinusoidal wavelike oscillation continue longer. On the other hand, T_p and T_l for the hard onset voice coincided with the abrupt rise in the volume.

Fig. 3 illustrates examples of the amplitude spectra of the soft and hard onset voices shown in Fig. 2. The H1-H2 values of the soft and very soft onset samples are lower than that of the hard onset sample because the first harmonic amplitude is relatively large in the former. This is characteristically observed in a breathy voice. The H1-H2 values of the soft and very soft onset samples are almost the same. They have the same sinusoidal wavelike oscillation at the beginning of the phonation, although the duration is longer for the latter.

Fig. 4 shows the distributions of the two types of rise time measures. Although the majority of the hard onset samples had

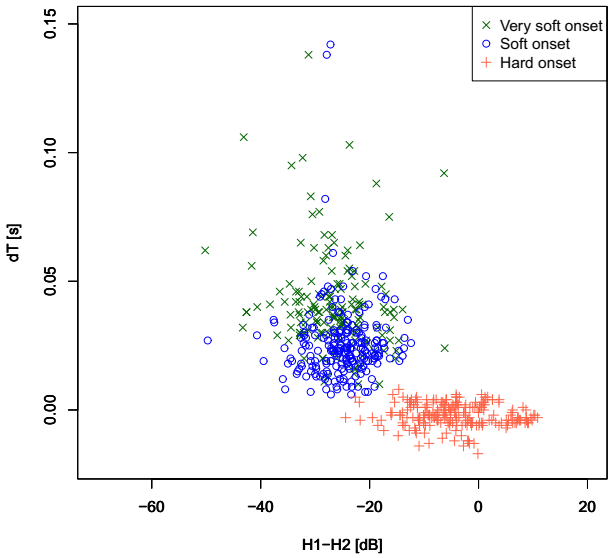


Figure 5: *Scatterplot of H1-H2 and dT.*

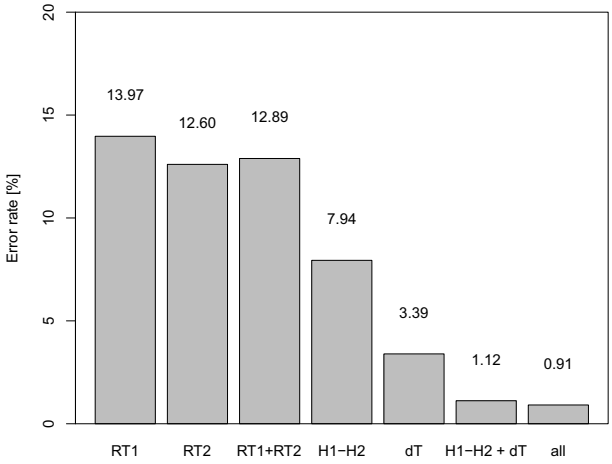


Figure 6: *Discrimination error rates*

smaller values than the soft onset samples, there were overlap regions for the different rise-time evaluation methods. This indicates that it is impossible to divide the two groups of voice onsets clearly with either of these parameters. Fig. 5 shows a scatterplot of the values of H1-H2 and dT. The soft and hard onset samples had separate distributions, whereas the distributions for the soft and very soft onset overlapped. We have experimentally confirmed that the distributions were not affected by the type of vowel.

Fig. 6 shows the discrimination error rates (%) for RT1, RT2, and the proposed parameters. The error rate was improved when the proposed parameters were used compared with the discrimination using the previously reported rise times. Using both H1-H2 and dT, the error rate decreased to about 1%. Although it has been reported that the acoustic parameters that correlate with breathiness partly overlap but differ by gender [16], H1-H2 can be utilized to discriminate soft and hard onset regardless of gender.

7. Conclusions and future work

We have provided a new discrimination measure for soft and hard voice onsets for the home-based training of the soft voice onset. We focused on the initial oscillation characteristics of the open vocal folds at the soft voice onset, and adopted parameters that were previously correlated with breathy voice qualities. By using acoustic features that characterize the soft onset, a smaller error rate was obtained than previous reported. In addition, a "short" (rapidly rising) soft voice onset can be classified correctly by this method. This type of voice onsets is difficult to judge from the rise times, or even by the trained ears of speech specialists. If these parameters are suitably fed back, it is expected that speech patients would be able to avoid practicing unnaturally prolonged voice onsets.

In future, we will develop a home-based training system that visually feeds back the results of the automatic discrimination by displaying two proposed acoustic parameters in a two-dimensional map. The effectiveness and easiness of the feedback will be evaluated by human subjects.

8. ACKNOWLEDGMENT

This work was supported by a JSPS KAKENHI Grant-in-Aid for Young Scientists (B) (Grant Number: 20623713)

9. References

[1] B. Guiter, *Stuttering: An integrated approach to its nature and treatment,* Lippincott Williams & Wilkins, 2013.

[2] E. G. Conture, G. N. McCall and D, W. Brewer, "Laryngeal behavior during stuttering," *Journal of Speech, Language, and Hearing Research*, vol. 20, no, 4, pp. 661–668, 1977.

[3] E. K. Lee, and Y. I. Son, "Muscle tension dysphonia in children: voice characteristics and outcome of voice therapy," *International journal of pediatric otorhinolaryngology*, vol. 69 no. 7, pp. 911–917, 2005.

[4] L. E. Glaze, "Treatment of voice hyperfunction in the preadolescent," *Language, Speech, and Hearing services in schools*, vol. 27, no. 3, pp. 244–250.

[5] H. A. Euler, A. W. V. Gudenberg, K. Jung, and K. Neumann, "Computer-assisted therapy for speech disfluency: The long-term effectiveness of the Kassel Stuttering Therapy (KST)," *Sprache, Stimme, Gehör*, vol. 33 no. 4, pp. 193–202, 2009.

[6] H. F. Peters, L. Boves and I. C. van Dielen, "Perceptual judgment of abruptness of voice onset in vowels as a function of the amplitude envelope," *Journal of Speech and Hearing Disorders*, vol. 51, no. 4, pp. 299–308, 1986.

[7] A. R. Mallard, D. M. Hicks and D. E. Riggs, "A Comparison of Stutterers and Nonstutterers in a Task of Controlled Voice Onset," *Journal of Speech, Language, and Hearing Research*, vol. 25, no. 2, pp. 287–290, 1982.

[8] R. F. Orlikoff, D, D, Deliyski, R. J. Baken and B. C. Watson, "Validation of a glottographic measure of vocal attack," *Journal of Voice*, vol. 23, no. 2, pp. 164–168, 2009.

[9] J. Hillenbrand, R. Cleveland and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[10] Y. Koike, "Experimental studies on vocal attack," *Practica Oto-Rhino-Laryngologica*, vol. 6, no. 8, pp. 663–688, 1967.

[11] C. Bickley, “Acoustic analysis and perception of breathy vowels,” *MIT Speech Communication Group Working Papers*, pp. 71–81, 1982.

[12] M. Garellek, R. Samlan, B. R. Gerratt, and J. Kreiman, “Modeling the voice source in terms of spectral slopes,” *Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. 1404–1410, 2016.

[13] Praat: doing phonetics by computer (Version 5.3.53), http://www.fon.hum.uva.nl/praat/, Accessed: 2013-07-30

[14] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *Proceedings of the institute of phonetic sciences*, vol. 7 no. 1193, pp.97–110, 1993.

[15] A. Zeileis, A., K. Hornik, A. Smola, and A. Karatzoglou, “kernlab-an S4 package for kernel methods in R,” *Journal of statistical software,* vol. 11, no.9, pp. 1–20, 2004.

[16] H. M. Hanson, and E. S. Chuang, “Glottal characteristics of male speakers: Acoustic correlates and comparison with female data,” *Journal of the Acoustical Society of America,* vol. 106 no. 2, pp. 1064–1077, 1999.