A LANDMARK-BASED APPROACH TO AUTOMATIC VOICE ONSET TIME ESTIMATION IN STOP-VOWEL SEQUENCES

Stephan R. Kuberski, Stephen J. Tobin, Adamantios I. Gafos

University of Potsdam
Linguistics Department
Potsdam, Germany

ABSTRACT

In the field of phonetics, voice onset time (VOT) is a major parameter of human speech defining linguistic contrasts in voicing. In this article, a landmark-based method of automatic VOT estimation in acoustic signals is presented. The proposed technique is based on a combination of two landmark detection procedures for release burst onset and glottal activity detection. Robust release burst detection is achieved by the use of a plosion index measure. Voice onset and offset landmarks are determined using peak detection on power rate-of-rise. The proposed system for VOT estimation was tested on two voiceless-stop-vowel combinations /ka/, /ta/ spoken by 42 native German speakers.

***Index Terms*—** Acoustic phonetics, speech processing, landmark detection, voice onset time

1. INTRODUCTION

Voice onset time (VOT) is a major parameter defining linguistic contrasts in voicing across languages [1]–[3]. Often VOT measurement is carried out manually as part of laboratory work in experimental phonetics [4]–[6]. Following many decades of progress in digital computing, it has become increasingly easy to build and run experimental investigations of speech production. As a consequence, the amount and availability of digitally acquired speech data has reached a level at which manual measurement is no longer feasible or economical. Many hours of human transcription could be saved by using automatic measurement algorithms for this purpose. However, this requires both robust and accurate methods of machine-aided annotation.

By definition, VOT is the length of the interval between the release of an oral closure (e.g., in the production of a voiceless oral stop consonant) and the onset of vocal fold vibration associated with the following vowel. Acoustically, this is manifested as a burst or abrupt increase in energy and a subsequent initiation of periodicity during which formant structures emerge. On the basis of this definition, any automatic method of VOT estimation minimally needs to imply, explicitly or implicitly, a robust way of detecting the two landmarks of burst onset (+b) and voice onset (+g). Explicit methods generally make use of a set of rules which home in to the final set of landmarks after an initial phase of identification of candidate landmarks. In contrast, implicit methods commonly apply supervised statistical learning techniques to accomplish this task.

A first notable development among the explicit methods of robust, automatic landmark detection in the field of speech processing comes from the work of Liu [7] in the mid 1990s. Parts of her work are taken as a basis for the development of the current framework. More recently Stouten and van Hamme [8] used spectral reassignment methods with enhanced time-frequency resolutions to estimate VOTs of stops. Application of supervised machine learning techniques began with the work of Lin and Wang [9] and was further developed by Sonderegger and Keshet [10], and Ryant et al. [11]. These methods rely on the availability of manually measured data to gather systematicities between the acoustic signal and the measurements.

The present work returns the focus to explicit knowledge-based approaches to landmark detection and VOT estimation, and presents a framework that performs well on a dataset of monosyllabic stop-vowel sequences spoken by native speakers of German. The major advantage of using a landmark rule-based system for VOT estimation is that there is no manual labeling needed beforehand as is the case for implicit estimation methods using statistical learning.

2. PROPOSED ESTIMATION SYSTEM

The proposed VOT estimation system consists primarily of two activity detectors. Each of these activity detectors produce a set of candidate landmarks, which are finally validated by means of a series of rules. The algorithm is meant to work well on clean acoustic speech signals with high signal-to-noise ratio as recorded in laboratory environments. Input recordings furthermore need to be narrowly cut to the syllable of interest, either by experimental design or a preceding voice activity detection.

2.1. Release burst detection

Ananthapadmanabha et al. [12] recently presented a well-performing algorithm for stop and affricate release burst landmark detection by using a so-called plosion index measure. The results of their work indicate that this one-dimensional

temporal measure is highly correlated with the events of release bursts of acoustic energy accompanying the production of oral stops. Here, fundamentals of their method are taken up and modified.

Generally, the instant at which the oral closure of a stop consonant is released is accompanied by an abrupt increase of acoustic energy. This event could either be tracked directly in terms of the average power of the source acoustic signal or by means of a pre-processed, transformed version of that same signal. Ananthapadmanabha et al. argued wisely for the use of the Hilbert envelope of the signal due to its independence from a possible, initial phase shift occurring in the source. Using the transformed version of the signal together with an equal loudness pre-filtering [13], release burst detection comes down to detecting the instants at which the signal's amplitude exceeds some threshold in relation to the average of a preceding vicinity. This relation computed as the ratio between amplitude and vicinity average is named the plosion index. It is a dimensionless quantity and therefore independent of source recording level. The authors Ananthapadmanabha et al. [12] furthermore recommended computing the plosion index only for sequential subsets between consecutive zero crossings of the signal using the maximum amplitude therein instead of evaluating it for every sample value.

The following algorithmic steps describe the proposed release burst detection method explicitly:

1) find the instants $n_1, n_2, \ldots$ of zero crossings in equal-loudness-filtered source signal $x[n]$, $n = 1, 2, \ldots$
2) compute the Hilbert envelope $H[n]$ of the signal using a time discrete Hilbert transform

$$H[n] = \left| x[n] + \frac{\mathrm{i}}{\pi} \sum_{\substack{k=-\infty \\ k \neq n}}^{\infty} \frac{x[k]}{n-k} \right| \tag{1}$$

3) in subsets between consecutive zero crossings, find the instants $m_{i,\max}$ at which the Hilbert envelope takes its maximum

$$m_{i,\max} = \underset{n_i \leq m \leq n_{i+1}}{\arg\max}\ H[m], \quad H_{i,\max} = H[m_{i,\max}] \tag{2}$$

4) consider the vicinity $[m_{i,1}, m_{i,2}]$ preceding that maximum $H_{i,\max}$ and its average value

$$H_{i,\mathrm{avg}} = \frac{1}{m_{i,2} - m_{i,1} + 1} \sum_{k=m_{i,1}}^{m_{i,2}} H[k] \tag{3}$$

5) set (non-zero) plosion indices $I[n]$ only at the beginning of that vicinity as the ratio between maximum and averaged Hilbert envelope

$$I[n = m_{i,1}] = \frac{H_{i,\max}}{H_{i,\mathrm{avg}}}, \quad I[n > m_{i,1}] = 0 \tag{4}$$

6) treat each non-zero plosion index as a candidate landmark ordered and prioritised by its specific value

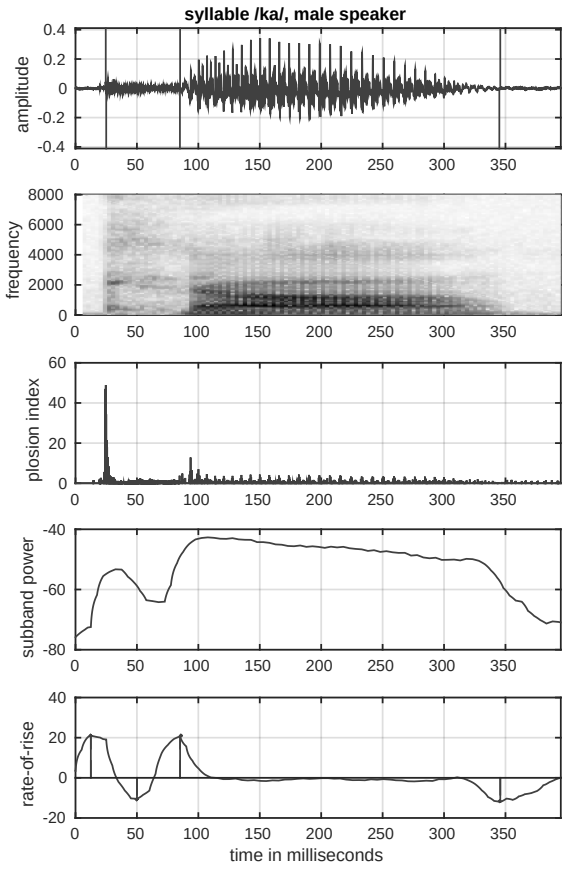

Figure 1: Waveform (top row) and spectrogram (second row) of an example syllable /ka/ spoken by a male subject. Third row shows the plosion index I given by equation (4). Fourth and bottom rows depict subband power P and power rate-of-rise R together with glottal candidate landmarks as computed by equations (7) and (8).

Given an example syllable /ka/ in Figure 1, with its waveform (first row) and spectrogram (second row), the so-computed plosion indices are shown in the third row. Clearly visible therein are two major series of peaks at around 25 ms and 90 ms, counted as the first two candidate landmarks for the occurrence of a release burst. The correspondence of the first candidate landmark with the actual release burst event is indicated by its higher value (resp. priority). However, possible appearances of additional, highly prioritised candidates, like the second one accompanying the beginning of glottal activity, need to be evaluated during a later stage of the estimation system as described in Section 2.3.

The control parameters of the proposed algorithm are the width $m_{i,2} - m_{i,1} + 1$ of the preceding vicinity and its temporal distance $m_{i,\max} - m_{i,2} + 1$ to envelope maximum $H_{i,\max}$. Ananthapadmanabha et al. suggested using values of 16 ms for vicinity width and 6 ms for its distance on the basis of detection performance (distance value) and statistics of burst transition length (width value). Throughout the present work, the fixed values of 10 ms for vicinity width and 1 ms for temporal distance were used. These different choices were made for reasons of detection performance with the current dataset.

2.2. Glottal activity detection

The basis of the proposed glottal activity detector is the estimation of the positions of two landmarks: one for voice onset (+g) and one for voice offset (−g), both flagging the region of vocal fold vibrations. Whereas only the former landmark is essential for further VOT estimation, the latter one comes as an algorithmic byproduct and can also be used to measure the duration of a vowel and to normalize VOTs by vowel length. Liu [7] presented a method of detecting these landmarks among some others. The fundamentals of her work are taken here as a basis and presented with slight modifications.

Vocal fold vibrations generally manifest themselves in the spectrogram of an acoustic signal as prominent bands of increased power (see Figure 1, second row). The existence of these characteristic bands, especially the lowest one referred to as fundamental frequency (F_0), can therefore be used as an appropriate indicator of glottal activity [14],[15]. By tracking the onset and offset of the fundamental frequency, candidates for the landmarks of voice onset and offset are obtained. To accomplish this, Liu suggested using the measure of spectral power rate-of-rise (ROR) of the most prominent frequency in a subband where F_0 is expected to be present (see Figure 1, last two rows). As a derivative-like measure, the ROR of power is associated with acoustic changes within this spectral subband. Hence, the peaks of the ROR that exceed an absolute threshold indicate the instants of most rapid change of spectral power and are treated as possible candidate landmarks where glottal activity turns on (+peaks) or off (−peaks). To ensure an expected natural sequence of alternating types of peaks (vocal fold vibrations must turn off before turning on again), peaks of reversed signs are inserted at the power ROR extrema between consecutive pairs of peaks having the same sign. Finally, leading −peaks and trailing +peaks are removed for the same reason of sequencing.

In the following, the explicit steps of the proposed algorithm of voice onset and voice offset landmark detection are listed:

1) compute the short time Fourier transforms of acoustic source signal $x[n]$, $n = 1, 2, \ldots$ at equally spaced instants m using window function w

$$X[m, \omega] = \sum_{k=-\infty}^{\infty} w[k-m]x[k]\,\mathrm{e}^{-\mathrm{i}\omega k} \tag{5}$$

2) follow the spectral power contour of the most prominent frequency in the subband $[\omega_{\min}, \omega_{\max}]$

$$P[m] = \max_{\omega_{\min} \le \omega \le \omega_{\max}} |X[m, \omega]|^2 \tag{6}$$

3) undo segmentation induced by short time Fourier transform by replicating power values of the same segments $P[m] \rightsquigarrow P[n]$
4) smooth the power contour by applying a box blur kernel $k[l]$, $l = 1, 2, \ldots, 2L$

$$P[n] = \sum_{l=1}^{2L} k[l]P[n+l-L] \tag{7}$$

5) approximate the derivative of the power contour by using the rate-of-rise (ROR) with a lookahead w_{a} and a lookbehind w_{b}

$$R[n] = P[n + w_{\mathrm{a}}] - P[n - w_{\mathrm{b}}] \tag{8}$$

6) find the peak positions in ROR exceeding the absolute threshold R_{thresh} using a Mermelstein-like peak detector [16]
7) pair consecutive peaks of the same sign by the inserting a peak with opposite sign between them at the extremum of ROR
8) remove any leading −peaks and trailing +peaks

The algorithm makes use of the following set of control parameters: the window width, overlap and function w of short time Fourier transforms, the spectral limits $\omega_{\min}$ and $\omega_{\max}$ of the subband under consideration, the values of lookahead w_{a} and lookbehind w_{b} for power ROR computation, and finally the threshold R_{thresh} of ROR peak detection. Liu [7] proposed a short time Fourier analysis using a 6 ms Hann window with an overlap of 5 ms. In the present work the different setting of a 15 ms Hann window with an overlap of 10 ms is used, resulting in a spectrogram with narrower bands and better detection performance. The spectral subband, originally set to a range of 0 . . . 400 Hz, was changed to the range of 150 . . . 500 Hz, permitting the removal of occasional mains hum and background noise from the source recordings while maintaining the inspection of the expected place of F_0. This also led to better detection rates. Both values of lookahead and lookbehind were set equally to 12.5 ms as recommended by Liu. The absolute threshold for power ROR peak detection was fixed to a value of 9 dB following the physiological arguments about sub-glottal and supra-glottal pressures by the same author.

2.3. Voice onset time estimation

The final estimation of VOT, based on the distance between previously detected candidate landmarks of release burst onset (+b) and voice onset (+g), is driven by the following ordered set of rules for candidate landmark validation:

1) any pair of consecutive candidate ±peaks lying completely in the first third of the utterance is rejected
2) all remaining, successive pairs of consecutive candidate ±peaks are merged into a single pair, having its +peak assigned to the landmark of voice onset (+g) and its −peak to the landmark of voice offset (−g)
3) any release burst candidate succeeding the validated voice onset landmark is rejected and the remaining candidate with highest priority is assigned to the final release burst landmark (+b)

The reason for rule 3) arises from the fact of processing voiceless-stop-vowel combinations in which voicing never precedes the release of the oral closure. The reasons for the first and second rule are derived from the assumption of processing appropriately cut recordings as stated in the beginning of Section 2. Occasionally the glottal activity

detector finds landmarks in the transition phase between the burst and voice onset when relatively large amounts of energy are present in the lower subband (see Figure 1, bottom row for an example). Application of the first rule compensates for this undesirable behavior. Furthermore, application of the second rule corrects for needless segmentation of glottal activity in case of emerging power fluctuations during the production of the vowel.

3. RESULTS

To evaluate the detection performance of the proposed VOT estimation system, its results are compared to manual measurements. Clean speech recordings (44100 Hz sampling rate, 16 bit depth, sound booth environment) of the stop-vowel sequences /ka/ and /ta/ were used as the test corpus. The total recordings consist of 40021 tokens (19881 /ka/, 20140 /ta/) spoken by 42 native German speakers (29 female, 13 male) with an average age of 23.7 years. In 3 tokens (2 /ka/, 1 /ta/) the release burst onset landmark detection method was not able to detect any burst. The glottal activity detection algorithm failed to detect any activity in 63 tokens (24 /ka/, 39 /ta/). Both kinds of detection misses yielded a total number of 63 tokens (24 /ka/, 39 /ta/) where no VOT estimation was possible. All other tokens were treated as properly detected landmarks or intervals.

To measure the accuracies of landmark detection and interval estimation the absolute deviations in millisecond from manual-labeled data were used. Figure 2 shows these accuracies graphically as the cumulative distributions of deviation between manual and automatic measurements. The graphs show the (cumulative) rate at which landmarks or intervals were correctly detected up to a specific level of tolerance expressed by the absolute deviation. Detection rates for landmarks at 10 ms tolerance are 96.1% (release burst onset), 97.3% (voice onset) and 73.3% (voice offset). At the same level of tolerance the interval estimation results are 94.1% (voice onset time) and 68.1% (vowel length).

The presented VOT estimation method was developed and tested on the basis of speech data from native German speakers. Although this dataset consists only of two stop-vowel combinations with the fixed vowel /a/, there appears to be no inherent reason for the proposed system not to perform well on other vowels too. Furthermore, VOTs do not differ substantially between American English, British English and German [1], [2], [4], [17]. In comparing the performance of the present system (94.1% overall VOT estimation accuracy at 10 ms tolerance) with different contemporary estimation techniques, it is worth mentioning that Stouten and van Hamme [8] achieved an accuracy of 76.1% based on the TIMIT database (cf. also Table 1), Lin and Wang [9] achieved 83.4% using the same database, the method of Sonderegger and Keshet [10] performed with an average accuracy of 87.6% on four different datasets including TIMIT, and Ryant et al. [11] achieved 91.7% averaged over three different datasets, also including TIMIT. However, it should also be noted that these approaches were developed on speech data from native English speakers and tested on larger subsets of consonant-vowel combinations (although in some cases with less tokens per combination than ours, e.g., the 168 speaker TIMIT set in Ryant et al. [11] had 5459 stops versus 40021 here). In future work, we aim to apply our approach to comparable dataset sizes (including word-medial stops which are not present in our dataset).

Author (and technique)	Accuracy
Stouten and van Hamme (reassignment spectra)	76.1%
Lin and Wang (random forests)	83.4%
Sonderegger and Keshet (structured prediction)	87.6%
Ryant et al. (support vector machines)	91.7%

Table 1: Comparison of different contemporary methods of automatic VOT estimation along with their detection performances. Detection accuracies are specified at a 10 ms level of tolerance. The proposed detection system achieved an accuracy of 94.1% on a different dataset.

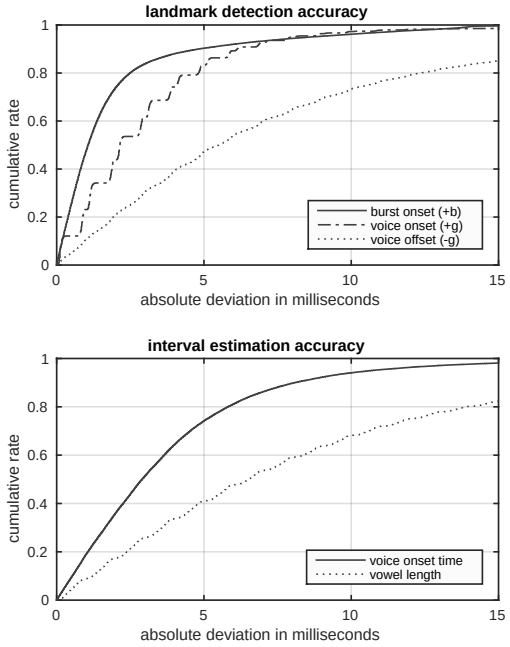

Figure 2: Cumulative distributions of absolute deviations between manual measurement and automatic detection of landmarks (upper graph) and automatic estimation of intervals (lower graph) resp. Periodic variations of rates for voice onset and voice onset time are mainly caused by the short time Fourier segmentation in step 1) of glottal activity detection.

4. CONCLUSION

The present work provides a robust method of automatic VOT estimation based on two well-performing landmark detection procedures. Whereas implicit techniques use methods of statistical learning, the above proposed explicit method does not depend on any manual measurements. Even without training on an already labeled data set, the present framework performs in the range of the above cited methods.

5. REFERENCES

[1] L. Lisker and A. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *WORD*, vol. 20, no. 3, pp. 384–422, 1964.

[2] A. Abramson and L. Lisker, "Discriminability along the voicing continuum: Cross language tests," in *Proc. 6th Int. Congr. Phon. Sci.*, 1967, pp. 569–573, Prague.

[3] A. Abramson, "Laryngeal timing in consonant distinctions," *Phonetica*, vol. 34, no. 4, pp. 295–303, 1977.

[4] C. A. Fowler, V. Sramko, D. J. Ostry, S. A. Rowland, and P. Hallé, "Cross language phonetic influences on the speech of French–English bilinguals," *J. of Phonetics*, vol. 36, no. 4, pp. 649–663, 2008.

[5] S. J. Tobin, *Phonetic accommodation in Korean-English and Spanish-English bilinguals: a dynamical approach*, Ph.D. thesis, Univ. Connecticut, 2015.

[6] E. Klein, K. D. Roon, and A. I. Gafos, "Perceptuo-motor interactions across and within phonemic categories," in *Proc. 18th Int. Congr. Phon. Sci.*, 2015, Glasgow.

[7] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3417–3430, 1996.

[8] V. Stouten and H. van Hamme, "Automatic voice onset time estimation from reassignment spectra," *Speech Comm.*, vol. 51, no. 12, pp. 1194–1205, 2009.

[9] C. Y. Lin and H. C. Wang, "Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection," *J. Acoust. Soc. Am.*, vol. 130, no. 1, pp. 514–525, 2011.

[10] M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured prediction," *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3965–3979, 2012.

[11] N. Ryant, J. Yuan, and M. Liberman, "Automating phonetic measurement: The case of voice onset time," in *Proc. Mtgs. Acoust.*, 2013, vol. 19, Montreal.

[12] T. V. Ananthapadmanabha, A. P. Pratosh, and A. G. Krishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 460–471, 2014.

[13] R. Robinson, "Replay gain—a proposed standard," `http://replaygain.hydrogenaud.io/proposal/equal_loudness.html`, 2001.

[14] K. Stevens, *Acoustic phonetics*, MIT Press, 2000.

[15] G. Fant, *Speech Acoustics and Phonetics: Selected Writings*, Kluwer Academic, 2004.

[16] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am.*, vol. 58, no. 4, pp. 880–883, 1975.

[17] M. Jessen, *Phonetics and Phonology of Tense and Lax Obstruents in German*, J. Benjamins Publ. Co., 1998.