Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

# Detection of Voice Onset Time (VOT) for Unvoiced Stops (/p/, /t/, /k/) Using the Teager Energy Operator (TEO) for Automatic Detection of Accented English

*Sharmistha Das*[1] *and John H.L. Hansen*[1,2]

[1]Department of Speech, Language and Hearing Sciences
[2]Robust Speech Processing Group, Center for Spoken Language Research;
University of Colorado, Boulder, Colorado 80309-0594, USA
sharmi@avaya.com, John.Hansen@colorado.edu

## Abstract

*Voice Onset Time (VOT) is an important temporal feature in speech perception and speech recognition. It also benefits for accent detection[1,2]. Fixed length frame based speech processing inherently ignores VOT. In this paper we propose a more effective VOT detection scheme using the non-linear energy tracking algorithm (Teager Energy Operator (TEO)) across a sub-frequency band partition for unvoiced stops (p, t and k). The VOT detection algorithm is applied to the problem of accent classification. Three different language groups (Indian, Chinese and American English) are used from CU-Accent-Corpus to compare VOT's of both accented and native American English. Some pathological cases are considered where speakers have breathy voices or other issues in recording procedure. The VOT is detected with less than 10% error when compared to the manual detected VOT. Also, pairwise English accent classification are 87% for Chinese accent, 80% for English accent, and 47% for Indian accent (includes atypical cases for Indian case).*

## 1. Introduction: Importance of VOT

Voice Onset Region (VOR) of a stop is the region of unvoiced speech, which starts after the pause of the stop (stop release area) and ends just before the onset of the voicing of the vowel. The length of the VOR is called Voice Onset Time (VOT) and represents a cue to the listener as to which stop is being produced. VOT's are generally ignored in fixed length frame based speech analysis, however it is known that VOT could help improve automatic speech recognition and understanding.

Among various applications of the use of VOT is the problem of accent detection. Foreign language can influence both length and quality of the VOR of English stops. In [1,2] it is shown that VOT's can be used to discriminate Mandarin, Turkish, German and English[1] accents. Let us take the example of Indian English. There are six stops in English: three unvoiced -- p, t, k and three voiced -- b, d, g. However, in Indio-Arian and Dravidian languages there are more than six stops. For example /k/ and /$k^h$/ are two different stops (with different alphabetic representations) used in Indian languages. Words with /k/ and /$k^h$/ will have significantly different meanings in Indian languages (In Bengali *kan* means 'ear' while *$k^h$an* means 'split'). In American English (AE), even though mild differences exist between k and $k^h$, they are not represented as different alphabet characters since either pronunciation does not change the meaning of the word. However, pronouncing words with /$k^h$/ in place of /k/, while it may not change the meaning in AE, does provide a cue for foreign accent. Indians, who learn AE by reading grammar books or from other Indian English speakers, have a tendency to pronounce words orthographically (lexically). For example, the word 'cake' can be transcribed as '$k^h$ E k' in AE. Unless an Indian person has learned the pronunciation of this word from an AE speaker, they will pronounce it as – 'k E k.' Only words with 'kh' as in 'khaki' are pronounced as '$k^h$ a k I'.

In [3], several methods of VOT detection are considered, with the most accurate method based on tracking the laryngographic signal. This method is not possible unless a laryngograph is used while recording speech from the speaker. The other methods are based on tracking formant frequencies (F1, F2 or F3), performing spectrographic analysis, or tracking the onset of speech ($f_0$) periodicity in the acoustic waveform. These methods all require some manual involvement to calculate the VOT's.

The Teager Energy Operator (TEO) [5] is a nonlinear energy tracking signal operator which has been used in speech and signal processing. It has been shown that TEO can be useful for detecting voiced/unvoiced speech[4], speech under stress[7], and speech under vocal fold pathology[8]. In this paper, an automated sub-band frequency analysis is performed to detect VOT of unvoiced stops. The Amplitude Modulation Component (AMC) [6] is used to detect vowel plus VOR's in different frequency bands assuming the stop to vowel transition have different amplitude envelopes for partitioned frequency ranges.

The paper is organized as follows: Sec. 2 develops the VOT detection algorithm. Sec. 3 presents evaluation results with application to accent classification. Sec. 5 summarizes the work and suggestions for future work.

## 2. VOT Detection Algorithm

In this section, we formulate the TEO and then extend this to VOR detection. Limitations for automatic VOR detection are then discussed.

### 2.1. TEO: Amplitude & Frequency Modulation

From a physics perspective, we know that the energy (E) of a simple harmonic oscillation is proportional to the square of the product of amplitude and frequency:

$$E \alpha A^2 \omega^2 = \dot{x} - x\ddot{x},$$

[1] Throughout the paper 'English accent' and 'American English' (AE) are used for 'standard native American English accent'.

where A is the amplitude and $\omega$ is the radian frequency and $x$ is the displacement over time. Using this simple theory, Teager formed a measure of the energy in any single component signal as:

$$TEO[x(n)] = x(n)^2 - x(n-1)x(n+1),$$

where *TEO* is the Teager Energy Operator applied on input signal *x(n)*, and *n* is the discrete time. Another way of measuring energy (*DTEO*) of the signal *x(n)*, is to calculate average of *Teager energy* of two sequences: *y(n) and y(n+1)* where,

$$y(n) = [x(n) - x(n-1)]$$
$$y(n+1) = [x(n+1) - x(n)],$$

Therefore, the *DTEO* of *x(n)* is given by:

$$DTEO[x(n)] = \frac{E1 + E2 - E3 - E4}{2}$$
$$E1 = [x(n) - x(n-1)]^2$$
$$E2 = [x(n+1) - x(n)]^2$$
$$E3 = [x(n-1) - x(n-2)][x(n+1) - x(n)]$$
$$E4 = [x(n) - x(n-1)][x(n+2) - x(n+1)]$$

The Frequency Modulation Component (FMC) and the Amplitude Modulation Component (AMC) of $x$ ($x_f$ and $x_a$), can be separated using the following equations:

$$x_f(n) = \arccos ine\left(1 - 0.5\frac{|TEO[x(n)]|}{|DTEO[x(n)]|}\right)$$
$$x_a = \sqrt{\frac{|TEO[x(n)]|}{1 - x_f^2}}$$

## 2.2. Detection of Vowel and VOR

It is proposed to use the AMC of a given speech signal to detect stop-vowel clusters. The highest energy portion of AMC is assumed to be the vowel in the low frequency range. It is also assumed that a stop will be followed by a vowel. Therefore, once a vowel is detected, the beginning is marked, and we move back in time to detect the VOT of the preceding unvoiced stop.

1. Only initial stops are considered.
2. A stop will be followed by a vowel (as in most English words with initial stops).
3. VOT are only defined for unvoiced stops -- /p/, /t/, /k/.
4. In order of low to high frequency components of the voice onset region (VOR), stops are arranged as /k/, /t/, /p/.
5. The VOR of a given stop will always fall in a certain frequency band.
   - /k/ → 1500 – 2500 Hz
   - /t/ → 2000 – 3000 Hz
   - /p/ → 2500 – 3500 Hz
6. Likewise, a vowel will have most energy in the low frequency band, i.e. 300 to 1200 Hz.

Table 1: List of assumptions for VOT detection.

Table 1 lists assumptions made for the following the proposed processing steps to detect VOT.

**Step 1**: A recognizer is used to detect words with unvoiced stops at the initial position followed by a vowel. The detected speech block is grouped into one of the three categories based on whether the stop is p, t or k.

**Step 2**: The signal (for the entire word) is band-pass filtered using a low frequency band (300-1200 Hz), and then AMC of the filtered signal is estimated. Next the vowel detection algorithm is used to mark the boundaries by locating the highest energy regions of the entire signal. Filtering the signal with a low frequency band emphasizes the vowels and de-emphasizes consonants (e.g., fricatives, nasals, stops, etc). Since vowels are the only pronounced region in the energy sequence under test, it is easy to mark their boundaries.

**Step 3**: Similarly, the original signal is then band-pass filtered with a high frequency band (Note: the frequency band is selected based on the identified stop category selected by the recognizer). The AMC of the filtered signal is then calculated. Filtering the signal with a specified band will only accentuate the VOR of a given stop. Vowel boundary locations are already marked from Step 2. From the beginning of a vowel, the VOT detection algorithm moves back in time and detects the VOT of the stop – the region with highest energy. This represents the final output of the algorithm.

**Step 4:** The VOT and various spectral analysis components of the marked VOR and the following vowel (wavelet analysis, FFT, Mellin transform) are then used for accent detection employing an HMM classifier structure.

To illustrate algorithm processing consider Fig. 1, which shows three signals. Signal 1 is the original input speech signal of the word 'catch'. Signal 2 is the AMC of the original input band-pass filtered speech with a 300-1200 Hz band (result from Step 2). Signal 3 is the AMC of the original input band-pass filtered speech with a 1500-2500 Hz band (result from Step 3). Note in Signal 2, the only noticeable high energy is in the vowel /@/. Therefore, band-pass filtering with a low frequency band has de-emphasized the VOR and affricative regions and made it easy to detect the vowel. In Signal 3, the VOR is emphasized as is the final affricate /C/. In Signal 2 the beginning of the vowel is already marked. From Signal 3, it is then possible to move back in time from the vowel to detect the VOT.

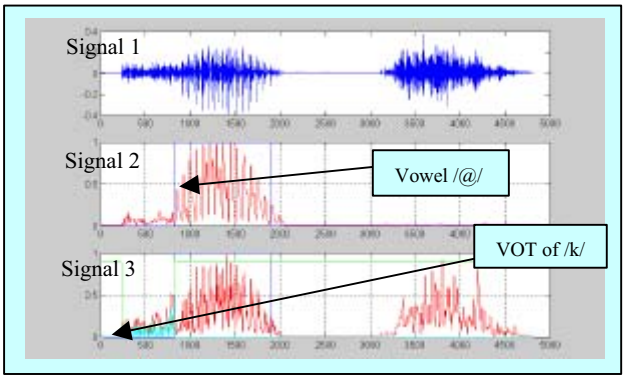


Fig. 1: VOT detection algorithm applied to the word "catch".

## 2.3. Atypical Cases for VOT Detection Algorithm

VOT detection in an automatic manner represents a challenge due to its short duration and frequency structure. Here we list several limitations of the proposed algorithm:

1. If a stop is followed by a reduced or unstressed vowel, then the energy and the length of the vowel is so low that it is often not detected properly in Step 2.
2. If the energy of the VOR of a stop is very high compared to the following vowel (often a reduced or

an unstressed vowel), then the VOR is incorrectly detected as part of the vowel in Step 2.

3. If the VOR of a stop has low frequency components like a vowel, then it is not fully de-emphasized when band-pass filtered with a low frequency band. In that case, the VOR is detected as a vowel in Step 2.
4. If VOT is too small (on order of 15msec) then it is very hard to reliably perform detection with this algorithm (e.g., occurs for some Indian speakers).

## 3. Evaluation

Having established the proposed VOT detection algorithm we now turn to an evaluation of the algorithm applying it for accent classification. Four male speakers from each of the three language groups – Indian, Chinese and American English, were selected. Table 2 describes each speaker (from Indian and Chinese accent group). Note, American English (AE) is considered as native English (without accent). Three words – 'catch', 'pump' and 'target' are chosen to detect VOT of initial /k/, /p/ and /t/ respectively. Each speaker was asked to speak each word 5 times in 4 separate recording sessions (for a maximum of 20 tokens per speaker)[2] using an online automated recording system.

| | Spkr #1 | Spkr #2 | Spkr #3 | Spkr #4 |
|---|---|---|---|---|
| ***Indian*** | | | | |
| **First Language** | Marathi | Marathi, Hindi | Tamil, Hindi | Marathi |
| **Lived in USA** | 5 months | 6 months | 3 months | 3 months |
| **Exposure to English** | 25 years | 20 years | 18 years | 15 years |
| ***Mandarin*** | | | | |
| **First Language** | Mandarin | Mandarin | Mandarin | Cantonese |
| **Lived in USA** | 4 months | 1.5 years | 5 years | 10 years |
| **Exposure to English** | 20 years | 5 years | 17 years | 30 years |

Table 2: Speaker Corpus information for VOT based Accent Classification

Table 3 shows VOT detection algorithm results for each speaker in each accent group. Each entry in this table is formatted as – (total number of tokens which are detected with less than 10% error using the automated VOT detection algorithm) \ (total number of tokens produced by the $n^{th}$ speaker). Error is calculated as:

$$error = \frac{|x - \hat{x}|}{x + .001} * 100$$

where $x$ is the manually (human) calculated VOT and $\hat{x}$ is the VOT detected using the proposed algorithm. Across the three accents, the VOT detection rate is 72.6% for /k/, 71.7% for /p/, and 79.9% for /t/ (rates were better for AE than non-native speakers for /k/ and /t/). The VOT detection rate using TEO is fairly low for Indian English speakers (47%), but is much better for Chinese and English speakers with rates of 87% and 80% respectively. English Speaker 2 has VOR for /k/ and /p/ with low frequency components (we suspect his mouth was too close to the microphone), which explains why VOT detection algorithm could not detect it properly (Sec. 2.3, case 4). If /k/ and /p/ for English Speaker 2 are excluded, then the detection rate for English speakers VOT is 88% (consistent with Chinese speakers VOT detection).

| | Spkr 1 | Spkr 2 | Spkr 3 | Spkr 4 | Total |
|---|---|---|---|---|---|
| **Indian** | | | | | |
| **c**atch | 2\2 | 7\14 | 9\18 | 11\19 | 29\53 |
| **p**ump | 1\1 | 2\5 | 0\1 | 5\15 | 8\22 |
| **t**arget | 3\7 | 6\12 | 7\18 | 10\20 | 26\57 |
| **English** | | | | | |
| **c**atch | 12\12 | 8\20 | 15\20 | 19\19 | 54\71 |
| **p**ump | 11\13 | 8\20 | 12\16 | 16\20 | 47\69 |
| **t**arget | 12\13 | 19\19 | 17\19 | 18\18 | 66\69 |
| **Chinese** | | | | | |
| **c**atch | 9\10 | 11\19 | 18\20 | 17\17 | 55\66 |
| **p**ump | 8\10 | 18\20 | 17\20 | 16\18 | 59\68 |
| **t**arget | 10\12 | 20\20 | 17\20 | 18\19 | 65\71 |

Table 3: Individual Stop Consonant and Speaker Results of VOT Detection Algorithm using TEO based processing.

Fig 2 shows scatter plots of human vs. automatic VOTs for Indian, American English (AE), and Chinese speakers. The three columns represent the VOT of /k/, /p/ and /t/ in 'catch', 'pump' and 'target' respectively. The three rows represent three language groups -- Indian, American English and Chinese respectively. VOT detected manually is plotted against the VOT detected using TEO. The tokens for each of the four speakers are plotted with different symbols. If the VOT detection algorithm fails to detect the VOT within a 10% error level, then that VOT is plotted on the floor (x-axis). Otherwise, if the VOT is properly detected then it approximately lies on the diagonal line. This implies that except for some exceptional cases where TEO has failed significantly, the TEO based algorithm has detected VOT with very high accuracy. We suggest that the 10% is reasonable level to employ, since in cases where VOT error is greater than 10%, it is significantly larger or smaller, so it would fairly easy to establish confidence boundaries in the VOT estimation procedure to ensure that the estimated VOT would be reliable for accent classification; if not, the VOT time could be set aside and an alternative characteristic employed for accent classification. The result from Fig. 2 scatter plots confirms that when we estimate the VOT, it conforms to human manual measurements.

Finally, Table 4 shows the mean of VOTs calculated manually and using the TEO algorithm. Each entry is formatted as: (mean of VOT detected manually) | (mean of VOT detected using TEO) all in msec.

## 4. Analysis of VOTs for Accent Classification

As we can see from Table 4 and Fig. 2, VOT is accent sensitive. VOT's for Indian speakers are low compared to English and Chinese, with the exception for Speaker 3.

[2] Multiple recording sessions ensured session-to-session variability. In some cases speakers did not say each word five times, or did not complete four separate sessions, or the words were not recorded properly. Therefore, some speakers have less than 20 tokens.

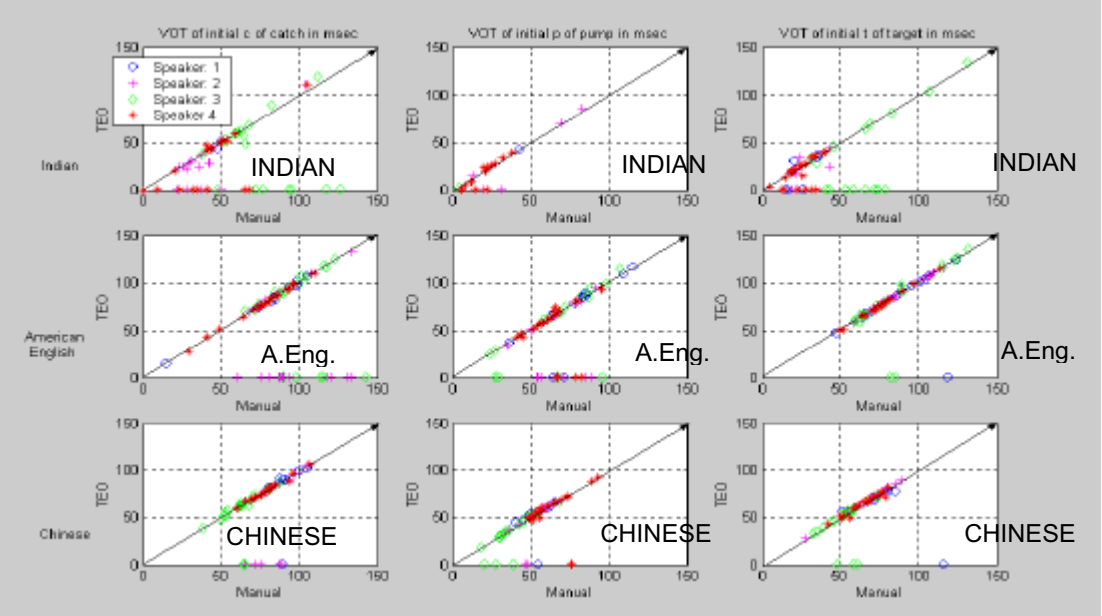


Fig 2. VOT detected using TEO is plotted against VOT detected manually for all four speakers of each language group.

Speaker 3 has a breathy voice and displays a lisp, which may explain why his VOTs are larger than other Indian speakers. Speaker 2 produced the word 'pump' only 5 times for his 4 four sessions making overall VOT detection of /p/ a coarse estimate (i.e. other speakers have nearly 20 samples). The mean VOT's for Indian speakers (excluding Speaker 3, /p/ for Speaker 2) are [50.81 35.37 28.58] for /k/, /p/ and /t/ respectively. VOT's for Chinese speakers are comparable but lower than English speakers. The mean VOT's for Chinese speakers are [77.24 53.54 64.01] for /k/, /p/ and /t/ respectively. VOT's for English speakers are highest. The mean VOT's for English speakers are [85.27 69.39 84.34] for /k/, /p/ and /t/ respectively.

| | Spkr 1 | Spkr 2 | Spkr 3 | Spkr 4 |
|---|---|---|---|---|
| **Indian** | | | | |
| catch | 49.0\|47.1 | 42.0\|43.1 | 65.1\|66.6 | 47.1\|47.4 |
| pump | 42.1\|42.3 | 76.0\|78.1 | NA | 28.6\|29.1 |
| target | 30.0\|30.5 | 28.7\|29.1 | 76.1\|75.7 | 27.0\|27.6 |
| **English** | | | | |
| catch | 82.3\|83.1 | 88.6\|89.4 | 93.3\|95.3 | 77.0\|77.5 |
| pump | 80.5\|82.3 | 66.3\|66.4 | 70.2\|72.2 | 60.7\|61.6 |
| target | 85.7\|86.3 | 84.6\|85.2 | 91.4\|93.2 | 75.7\|76.3 |
| **Chinese** | | | | |
| catch | 87.1\|87.2 | 78.6\|78.3 | 64.0\|63.9 | 79.2\|78.8 |
| pump | 57.0\|57.9 | 53.3\|53.4 | 40.5\|40.7 | 63.4\|63.5 |
| target | 71.4\|70.3 | 67.2\|67.4 | 53.1\|53.1 | 64.4\|64.5 |

Table 4: Mean VOTs from (Manual | TEO estimation).

## 5. Summary & Conclusions

The ability to detect VOT in speech is a challenging problem because it combines temporal and frequency structure over a very short duration. In this study the AMC of TEO, a sub-frequency band based non-linear energy tracking algorithm operator, is developed to track VOR and vowel energies. This algorithm is applied to accent classification using English, Chinese, and Indian accented speakers. Using 546 tokens, consisting of 3 words from 12 speakers, the average msec mismatch between automatic and hand labeled VOT is 0.735 msec (excluding the atypical cases). This represents a 1.15% mismatch. It is also shown that the average VOT's are different among three different language groups, hence making VOT a good feature for accent classification.

## 6. References


[1] L.M. Arslan, J.H.L. Hansen, "Language Accent Classification in American English," *Speech Communication,* 18:353-367, 1996.

[2] L.M. Arslan, J.H.L. Hansen, "A Study of Temporal and Frequency Characteristics in American English Foreign Accent," *J. Acoustical Society of America*, 102(1):28-40, July, 1997.

[3] A.L. Francis, V. Ciocca, J.M.C. Yu, "Accuracy and Variability of Acoustic Measures of Voicing Onset," *J. Acoustical Society of America* pg. 1025, 2003.

[4] N. Sundaram, B.Y. Smolenski, R. Yantorno, "Instantaneous Nonlinear Teager Energy Operator for Robust Voiced – Unvoiced Speech Classification;" http://www.temple.edu/speech_lab/sundaram.PDF

[5] J.F. Kaiser, "On a Simple Algorithm to Calculate the'Energy' of a Signal," IEEE ICASSP-90 – Inter. Conf. Acoustics, Speech, and Signal Processing, Albuquerque, NM, pp. 381-384, Apr. 1990.

[6] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Energy Separation in Signal Modulations with Applications to Speech Analysis," IEEE Trans. on Signal Processing, 41(10):3024-3051, Oct. 1993.

[7] G. Zhou, J.H.L. Hansen, J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress," IEEE Trans. Speech & Audio Processing, 9(2):201-216, March 2001.

[8] J.H.L. Hansen, L. Gavidia-Ceballos, J.F. Kaiser, "A nonlinear based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. Biomedical Engineering*, 45(3):300-313, March 1998.

[9] CU-Accent: http://cslr.colorado.edu/accent/