Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 777–789
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification

John H.L. Hansen *, Sharmistha S. Gray, Wooil Kim

Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080-1407, USA

Received 13 June 2009; received in revised form 13 April 2010; accepted 9 May 2010

Abstract

Articulation characteristics of particular phonemes can provide cues to distinguish accents in spoken English. For example, as shown in Arslan and Hansen (1996, 1997), Voice Onset Time (VOT) can be used to classify mandarin, Turkish, German and American accented English. Our goal in this study is to develop an automatic system that classifies accents using VOT in unvoiced stops[1]. VOT is an important temporal feature which is often overlooked in speech perception, speech recognition, as well as accent detection. Fixed length frame-based speech processing inherently ignores VOT. In this paper, 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/), is introduced. The proposed VOT detection algorithm also incorporates spectral differences in the Voice Onset Region (VOR) and the succeeding vowel of a given stop-vowel sequence to classify speakers having accents due to different ethnic origin. The spectral cues are enhanced using one of the four types of feature parameter extractions – Discrete Mellin Transform (DMT), Discrete Mellin Fourier Transform (DMFT) and Discrete Wavelet Transform using the lowest and the highest frequency resolutions (DWTlfr and DWThfr). A Hidden Markov Model (HMM) classifier is employed with these extracted parameters and applied to the problem of accent classification. Three different language groups (American English, Chinese, and Indian) are used from the CU-Accent database. The VOT is detected with less than 10% error when compared to the manual detected VOT with a success rate of 79.90%, 87.32% and 47.73% for English, Chinese and Indian speakers (includes atypical cases for Indian case), respectively. It is noted that the DMT and DWTlfr features are good for parameterizing speech samples which exhibit substitution of succeeding vowel after the stop in accented speech. The successful accent classification rates of DMT and DWTlfr features are 66.13% and 71.67%, for /p/ and /t/ respectively, for pairwise accent detection. Alternatively, the DMFT feature works on all accent sensitive words considered, with a success rate of 70.63%. This study shows that effective VOT detection can be achieved using an integrated TEO processing with spectral difference analysis in the VOR that can be employed for accent classification.

Keywords: Voice Onset Time (VOT); Voice Onset Region (VOR); Teager Energy Operator (TEO); Accent classification

1. Introduction: Importance of voice onset region (VOR)

The Voice Onset Region (VOR) of an unvoiced stop includes the unvoiced speech which starts after the pause of the stop (stop release area) and ends just before the onset of voicing of the following phoneme. A VOR for an egressive stop ideally includes two distinct regions:

- The initial burst of energy – this region signifies the initial opening of the articulatory organs to release the air

* Corresponding author. Address: Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA. Tel.: +972 883 2910; fax: +972 883 2710.
E-mail address: john.hansen@utdallas.edu (J.H.L. Hansen).
URL: http://crss.utdallas.edu (J.H.L. Hansen).

[1] A preliminary version of some of the work in this study was presented at the IEEE NORSIG-04 Symposium in Das (Gray) and Hansen (2004).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.05.004

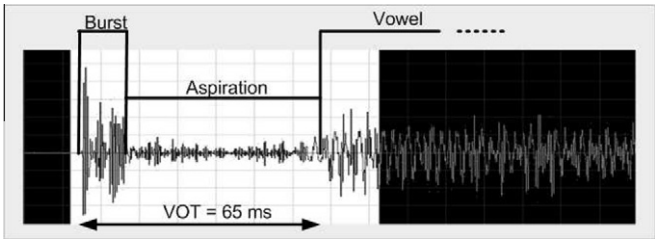

Fig. 1. The medial stop-vowel sequence /k/-/e/ part of the word 'communication'. The highlighted region is the VOR of /k/ together with the initial voicing of /e/.

pressure built up in the oral cavity (e.g., the opening of the lips while producing the /p/[2] to release the burst of air).
- The aspiration region – this region is marked by the drop of energy which succeeds the initial burst.

After the VOR, voicing of the following phoneme, typically a vowel (or a voiced approximant, /w/, /l/, /r/, and /y/), begins. In the initial part of the voicing of the vowel, there is formant movement prior to obtaining the steady state formant locations due to movement of vocal articulators from the stop to the vowel transition. Fig. 1 illustrates VOR for the stop /k/ and the initial voicing of the following vowel /e/. The length of the VOR is called the Voice Onset Time (VOT) and provides an acoustic cue to the listener as to which stop is produced. VOTs are generally ignored in fixed-length frame-based speech analysis because stop-consonant recognition is challenging, though VOT could offer a method to improve Automatic Speech Recognition (ASR).

Francis et al. (2003) have presented several methods for VOT detection with the most accurate method based on tracking the laryngographic signal. Rosner (1984) considered the perception of voice onset time from a signal detection strategy. A physiological study (Steinschneider et al., 1999) has shown that VOT can be directly recorded from the human auditory cortex (Heschl's gyrus, the planum temporale, and the superior temporal gyrus). These methods are only possible if a laryngograph, or other appropriate device to measure neural mechanisms, are employed simultaneously while recording speech from the speaker. Other methods are based on tracking formant frequencies (F1, F2 or F3), performing spectrographic analysis, or tracking the onset of speech (F0) periodicity in the acoustic waveform. These methods all require some manual interaction to obtain reliable VOTs. As recent studies for automatic VOT detection, Kazemzadeh et al. (2006) employed phone model based methods with forced alignment, and Stouten and Van hamme (2009) proposed a method using reassignment spectra. Mahadeva Prasanna et al. (2009) proposed a detection method for vowel onset point using source, spectral peaks and modulation spectrum energy. Finally, Lopez-Bascuas et al. (2004) considered VOT and "buzz-onset" time based on ROC curves.

In this study we employ the Teager Energy Operator (TEO) (Teager, 1980; Kaiser, 1990), which is a non-linear energy tracking signal operator that has been used in speech and signal processing. TEO based processing has been useful for detecting voiced/unvoiced speech (Sundaram et al., 2003), speech under stress (Zhou et al., 2001), and speech under vocal fold pathology (Hansen et al., 1998). Our focus in this study is to (i) formulate an automated sub-band frequency analysis method to detect VOT of unvoiced stops and to (ii) show that this method can be employed for accent classification. The Amplitude Modulation Component (AMC) obtained from the TEO (Maragos et al., 1993) is used to detect vowel plus VOR in different frequency bands, assuming the stop to vowel transition will have different amplitude envelopes for partitioned frequency ranges.

Based on the detected VOR, we will identify the VOT and the spectral differences in the VOR and the succeeding vowel of a given stop-vowel sequence. The spectral cues will be enhanced by one of four types of feature processing methods: Discrete Mellin Transform (DMT), Discrete Mellin Fourier Transform (DMFT) and Discrete Wavelet Transform (DWT), using the highest and the lowest frequency resolutions. Hidden Markov Models (HMM) are constructed with these extracted parameters. The VOT detection algorithm and HMMs are then applied to the problem of accent classification. While the diversity of the worlds' languages is significant (Comrie, 1990), there remains significant unanswered questions concerning accent traits for speakers of a primary language (L1) speaking a secondary/less familiar language (L2). It will be noted that a number of research studies have considered the problem of accent classification, and it should be emphasized that VOT estimation is would in general not be effective as the sole discriminating trait for general accent classification. However, this study suggests it can be employed to provide effective knowledge that can be leveraged in a multi-dimensional feature-set for future systems.

The paper is organized as follows: Section 2 describes the VOT detection algorithm with a brief background on the TEO, followed by assumptions made for developing the VOT detection algorithm. Section 3 presents a spectral analysis of the VOR of unvoiced stops and the succeeding vowels. We will assess whether the difference in frequency structure of the VOR and the possibility of the replacement of the vowel after the stop could be used for the problem of accent detection and classification, since VOT has been shown to be sensitive to accent structure in American English (Arslan and Hansen, 1996). Section 4 presents evaluation of the proposed VOT detection algorithm and accent classification with various spectral cues as discriminating factors including discussion on the results. Section 5 summarizes the findings of this paper and suggestions for future research.

[2] In this study single symbol ARPAbet is used. See p. 117, Deller et al. (1999).

2. VOT detection algorithm

In this section, we formulate the TEO, which will be extended to VOT detection in Section 2.2. Limitations in automatic VOT detection will be discussed in Section 2.3.

2.1. TEO: Amplitude and frequency modulation

From a physics perspective, it is known that the energy, E, of a simple harmonic oscillation is proportional to the square of the product of amplitude and frequency:

$$E \propto A^2\omega^2 = \dot{x}^2 - x\ddot{x}, \tag{1}$$

where A is the amplitude, ω is the radian frequency and x is the displacement over time. Using this simple theory, Teager (1980), Teager and Teager (1983), and Kaiser (1990) formed a measure of the energy in any single frequency component signal as:

$$TEO[x(n)] = x(n)^2 - x(n-1)x(n+1), \tag{2}$$

where $TEO[\cdot]$ is the Teager Energy Operator applied to an input signal $x(n)$ and n is discrete time. Another way of measuring the energy is to consider the Differential *TEO* (*DTEO*) of the signal $x(n)$ which is the average of the Teager energy of two sequences: $y(n)$ and $y(n+1)$ (See Eq. (106) of Maragos et al., 1993), 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:

$$\begin{aligned} DTEO[x(n)] &= \frac{TEO[y(n)] + TEO[y(n+1)]}{2} \\ &= \frac{E_1 + E_2 + E_3 + E_4}{2}, \end{aligned} \tag{3}$$

where:

$$\begin{aligned} &E_1 = [x(n) - x(n-1)]^2, \quad E_2 = [x(n+1) - x(n)]^2, \\ &E_3 = -[x(n-1) - x(n-2)][x(n+1) - x(n)], \\ &E_4 = -[x(n) - x(n-1)][x(n+2) - x(n+1)]. \end{aligned}$$

In addition to *TEO* and *DTEO*, it is possible to extract the Frequency Modulation Component (FMC) and the Amplitude Modulation Component (AMC) of a band-limited amplitude/frequency modulated signal (for example, speech signal). The *FMC* and *AMC* of $x(n)$, represented by $x_f(n)$ and $x_a(n)$, can be separated using the following equations (DESA-1, see Eq. (107) and (108) of Maragos et al., 1993):

$$x_f(n) = \arccos\left(1 - \frac{DTEO[x(n)]}{2TEO[x(n)]}\right), \tag{4}$$

$$x_a(n) = \sqrt{\frac{|TEO[x(n)]|}{1 - \cos^2(x_f(n))}}. \tag{5}$$

It is assumed that $TEO[x(n)] > 0$ and $DTEO[x(n)] > 0$ for noiseless AM-FM and band-pass filtered speech signals for most of the cases. If $x_f(n)$ is calculated as zero (this can happen due to quantization or computational accuracy), then $x_a(n)$ is calculated from the previous sample.

The biggest advantage of DESA-1 is its very low-complexity of $O(N)$, making it an useful tool to detect VOR at close to real-time. Maragos et al. (1993) also showed that DESA-1 performs better than Hilbert Transform (HT), which is another possible method to decompose speech signals into AM-FM components, when HT is operated via shorter impulse response to match its complexity similar to DESA-1 to $O(N)$. HT implemented via an FIR filter or Fast Fourier Transform (FFT) can give smaller error than DESA-1, at the expense of the higher-complexity of $O(N^2)$ and $O(Nlog_2N)$, respectively. Another advantage of DESA-1 is that it uses a very short window (in order of 5 samples) which allows it to instantaneously adapt during speech transitions between phonemes. Therefore, DESA-1 is an ideal energy operator to note the abrupt amplitude/frequency changes at the phoneme junction; That is, transitions between an unvoiced stop and the following vowel. Readers are suggested to refer to the article by Maragos et al. (1993) for further details about AM-FM decomposition of speech signals.

2.2. Detection of the vowel and the preceding VOR

It is proposed to use the AMC of a given speech signal to detect stop-vowel clusters. The highest energy portion of the AMC is assumed to reflect the presence of the vowel in the low-frequency range. In our framework we also assume 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.

The assumptions made in our formulation are as follows:

1. Only unvoiced stops in the word-initial positions (with word-initial stressed syllables) are considered[3].
2. Prior knowledge about the structure of the word is assumed; i.e., in this study, a given word must have a word-initial unvoiced stop.[4]

[3] In English, all mono-syllabic words have stressed syllable and many multi-syllabic words have stressed first syllable. However, there are many exceptions among multi-syllabic words. Words with prefixes, such as, per- , pre- , and com-, col-, con- , verbs, such as, 'prefer', 'protect', 'collect', 'command', and 'connect', and many of the nouns and adjectives derived from them ('protective', 'collection'), tend to be stressed on the second syllable. Some other examples of unstressed word-initial syllables are: 'today', 'tomorrow', 'potato'.

[4] Due to co-articulation, the word-initial unvoiced stop, may not be fully articulated as an 'ideal' pronunciation should be; but, in this study we will assume that speaker has intended to pronounce it. For example, the vowel-initial formant structure of /A/ in the words 'cup' and 'up' are completely different. However, the vowel-initial formant structure of /A/ in 'cup' will be similar irrespective of the case whether the word-initial /k/ is aspirated or unaspirated. We will rely on the later statement to apply the spectral cues of VOR for the problem of accent classification.

3. The word-initial unvoiced stop will be followed by a vowel.[5]
4. Positive VOTs are only defined for unvoiced stops, /p/, /t/, or /k/.[6]
5. In order of low to high-frequency components of the voice onset region (VOR), stops are arranged as /k/, /t/, and /p/[7] (Ladefoged, 1993).
6. The VOR of a given stop for adult speakers will fall in a certain frequency band, as shown below:
   - /k/ 1500–2500 Hz
   - /t/ 2000–3000 Hz
   - /p/ 2500–3500 Hz
7. In addition, the vowel is assumed to have the most energy in the low-frequency band (first formant, i.e., 300 to 1200 Hz).

Fig. 2 illustrates the flow diagram of the proposed algorithm. The algorithm consists of primarily two processing phases: (i) Vowel emphasis processing, and (ii) VOT analysis. The subsequent steps are shown as follows:

Step 1: An input audio stream is submitted to a front-end ASR to detect words with unvoiced stops at the initial position followed by a vowel. The detected speech block is grouped into one of three categories based on whether the stop is *p*, *t* or *k*. This step is needed when operating on a spontaneous input audio stream but not needed if words with known stop-vowel sequence at the initial position are submitted in isolation.

Step 2: The entire detected word is band-pass filtered using a low-frequency band (300–1200 Hz) and then the AMC of the filtered signal is estimated.

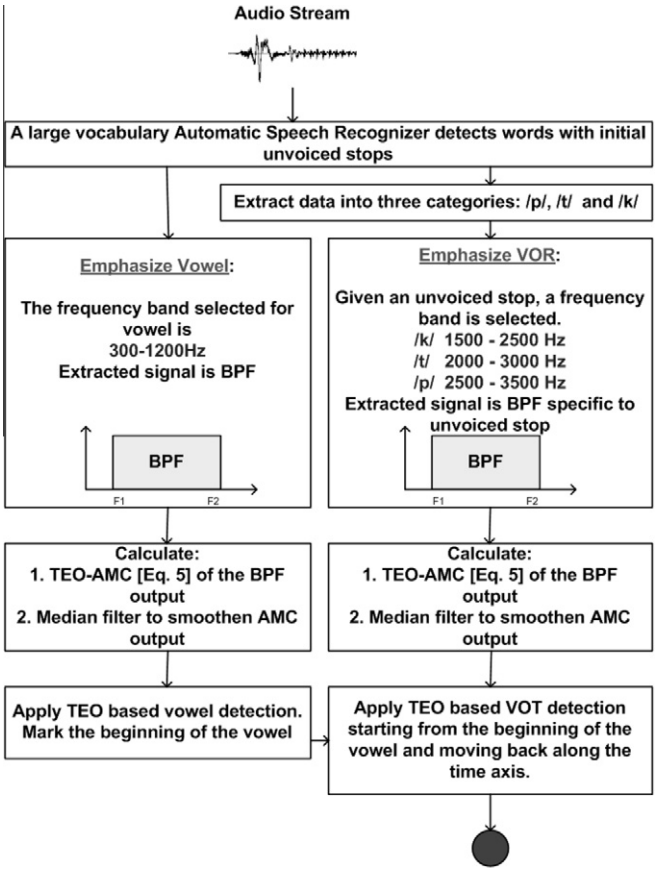

Fig. 2. Flow diagram of the VOT detection algorithm using TEO.

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 most consonants (e.g., fricatives, stops, affricates). Some consonants, such as approximants (/w/, /l/, /r/, and /y/) have vowel-like characteristics with low-frequency content. Hence, unlike other consonants, approximants will not be completely de-emphasized; however, it is not necessary to find the boundaries between vowels and approximants in the current study.

Step 3: Similarly, the original signal is band-pass filtered with a high-frequency band. The appropriate stop frequency band is selected based on the identified stop from the recognizer. The AMC of the filtered signal is then calculated using Eq. (5). 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 length of the region that has the highest energy. This represents the final output of the algorithm.

[5] In English, the word-initial unvoiced stops (/p/, /t/, and /k/) can be followed by most vowels and the approximants, /w/, /l/, /r/, and /y/. Ladefoged (1993) has stated in his book that when approximants occur in consonant clusters with initial unvoiced stop-consonants (such as 'play', 'twice', 'clay', 'cue'), approximants are 'largely voiceless', due to co-articulation. The devoicing of the approximants is the manifestation of the aspiration that occurs after unvoiced stops. Ladefoged has shown the transcriptions of the words: 'play' as /p l_oe/ and 'pie' as /p^hY/. /p^h/ in 'pie' indicates that /p/ is aspirated (in word-initial stressed syllables) where as /p/ in 'play' indicates that /p/ is unaspirated at the expense of the following devoiced approximant /l_o/. Since approximants in clusters are usually shorter than about 60 ms, they can be entirely or almost entirely devoiced after syllable-initial voiceless stops because the approximant lies within the time-span before voicing starts. So technically the entire length of the approximant will be part of the VOR of the syllable-initial voiceless stops. Conversely, most vowels, except unstressed ones in rapid speech, last quite a bit longer than 60 ms, so they are only partly devoiced (the part before the onset of the voicing). Hence in this research to avoid confusion, we have assumed that VOT (length of the VOR) is only defined for the unvoiced stops which are followed by a vowel, and not a consonant.

[6] Negative VOTs are normal for voiced stops, /b/, /d/, and /g/, but are not considered in this paper.

[7] Although /p/ sound represents a wide-band characteristic in frequency, the frequency components with range of 2500–3500 Hz were most effective for its VOR detection in this study.

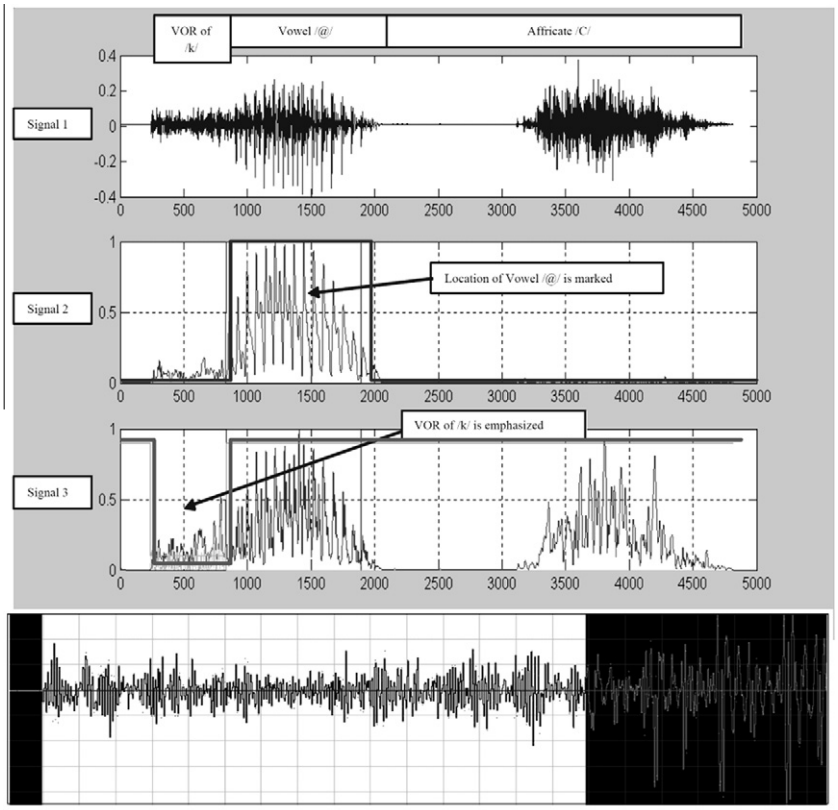

Fig. 3. Top: An example of using TEO to detect VOR of the initial /*k*/ of the word 'catch'. Bottom: The VOR of the /*k*/ (highlighted white region) is zoomed in to show the details of the temporal characteristics.

Step 4: The VOT and several spectral analysis components of the marked VOR and the following vowel are used for accent detection employing an HMM classifier structure.

To illustrate algorithm processing, consider the top portion of Fig. 3 which shows three signals. Signal 1 is the input speech signal of the word 'catch'. Signal 2 is the TEO based AMC of the input band-pass filtered speech over a 300–1200 Hz band (result from Step 2). Signal 3 is the AMC of the input band-pass filtered speech over a 1500–2500 Hz band (result from Step 3). In Signal 2, the high-energy portion is in the vowel /@/ portion only. The result from this figure shows that band-pass filtering with a low-frequency band has de-emphasized the VOR and affricate regions and made it easy to detect the vowel. In Signal 3, the VOR of /*k*/ is emphasized as well as the final affricate /*C*/. We see that in Signal 2 the beginning of the vowel is already marked, and from Signal 3 it is now possible to move back in time from the vowel to detect the VOR. At the bottom of Fig. 3 the final detected VOR of the /*k*/ in the word 'catch' is shown.

2.3. Atypical Cases for VOT detection algorithm

Automatic VOT detection schemes face challenges due to VOT's short duration and frequency structure, as well as the variability in production of unvoiced stops in isolated and continuous speech. Here we identify recognized limitations of the proposed algorithm:

1. If a stop is followed by a reduced or unstressed vowel, the energy and the duration of the vowel is often low, which reduces detection ability in Step 2. Similarly, 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.
2. If the VOR of the stop has low-frequency components, then it is not fully de-emphasized when band-pass filtered with the low-frequency band. In that case, the VOR is detected as a vowel in Step 2. Fig. 4 shows such an example, where Signal 1, 2 and 3 are obtained in a similar way to that in Fig. 3. In Signal 2 (as opposed to Signal 2 of Fig. 3), the VOR is not de-emphasized when band-passed filtered with the low-frequency band. Thus, the VOR is detected as part of the following vowel. As a result, the VOR is misdetected as shown in Signal 3. The bottom of Fig. 4 shows the actual VOR. If we compare the VOR from Figs. 4 and 3, it is clear that the VOR in Fig. 4 has a low-frequency area.
3. If the VOT is too short (on the order of 15 ms, which occurs for some Indian speakers in some words), detection performance with this algorithm is not as accurate as the typical cases.

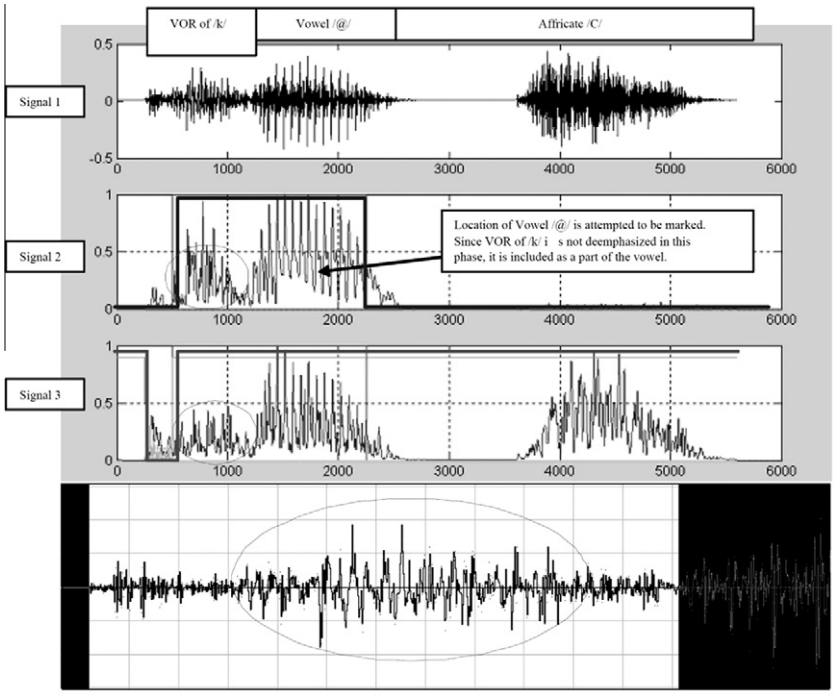

Fig. 4. Top: An example showing TEO has misdetected the VOR of the initial /*k*/ of the word 'catch'. Bottom: The VOR (highlighted white region) of the /*k*/ is zoomed in to show the details of the temporal characteristics. Note the VOR has vowel-like quality-low-frequency area circled in the figure.

3. Spectral analysis of initial unvoiced VOR and vowel sequences: HMM-based accent classifier with various feature types

Through the analysis of spectral features, we have noted distinct spectral differences in the VOR regions of Indian accented speakers as opposed to Chinese and American English (AE) speakers. An ideal VOR, as shown in Fig. 1, has two distinct parts-the burst and the aspiration. It is observed that Indian speakers do not have these two separate regions. The VOR for an Indian speaker is just a short duration of low-amplitude region that appears noisy with no change of spectral or amplitude characteristics over the region. This is due to how stops are pronounced in Indian languages. If orthographically represented '*k*', '*p*', or '*t*', Indian English speakers generally consider it as an unaspirated stop regardless of its position in a syllable.

Chinese speakers, like AE speakers and unlike Indian speakers, when speaking English, do not make the word-initial stops unaspirated if the words begin with '*k*', '*p*' or '*t*' (it will be unaspirated only if the words are spelled with '*g*', '*b*' or '*d*'). This is due to the pronunciations of stops in Chinese language (Hoshino and Yasuda, 2003), which can be seen in the aspirated stops /k^h/, /p^h/ and /t^h/ stops for '*k*', '*p*' and '*t*', versus the unaspirated stops /*k*/, /*p*/ and /*t*/ for '*g*', '*b*', '*d*'. As a result, VOT, which is directly correlated with aspiration (Poser, 2004), will not be a discriminating factor for accent detection between AE and Chinese speakers. However, it is noted that the stressed vowel after the word-initial stops in English words are often replaced by similar sounding but different vowels by Chinese speakers. Contrasting differences in English versus Chinese (Cantonese) was previously considered by Chan and Li (2000) from a phonology perspective and Peng and Ann (2004) from a voicing versus devoicing issue. Consequently, the shift in formant location in the VOR to the following vowel region result in different tracks than that seen for AE speakers. Similar changes in formant structure have been seen in Flemish accent (Ghesquiere and Compernolle, 2002). It is also known that vowel height has an effect on consonant voicing structure as illustrated in a study on Italian (Esposito, 2002). Therefore, spectral analysis is necessary to differentiate between VORs of American English and Chinese speakers.

The distinct spectral characteristics of VORs among speakers of different origins are hard to capture with a closed-form mathematical formulation. In this study, as an analytical approach, Hidden Markov Model (HMM) is used to model the VOR and the subsequent vowel, aiming at capturing spectral and temporal characteristics among speakers of different foreign origins. We recognize that there is a spectral sensitive shift, primarily due to tongue articulator movement, for the second formant in the stop to vowel phoneme sequence production. Here, we investigate alternate methods of projecting this information into a more discriminating space for accent classification. Four types of features are considered: Discrete Mellin Transform (DMT), Discrete Mellin Fourier Transform (DMFT), and Discrete Wavelet Transform (DWT) (with

the highest and the lowest frequency resolution). The mathematical representation and motivation for employing these features to model the VOR is described in subsequent sections.

3.1. Discrete Mellin transform (DMT) analysis

It is expected that the VOR of different speakers will have different durations. To capture only the spectral changes in this region, we wish to eliminate the variability of duration in VORs (i.e., VOT is already determined, so processing here is focused only on the spectral content). This can be achieved by using the Mellin Transform, which is done by exponential time-warping of a given function followed by its Fourier Transform.

Hence, analysis using the Mellin Transform focuses on the overall shape structure of a signal independent of the duration (it has been used for image and pattern analysis (Zwicke and Kiss, 1983; Fang and Häusler, 1990; Levkovitz et al., 1997)), for classifying objects based on different depths of view and has also been used to assess duration independence of excitation structure (Cairns and Hansen, 1994). It might be argued that fixed length sampling would provide sufficient spectral characterization. However, it can be shown that the Mellin Transform creates a spectrum by normalizing time domain duration (using exponential time-warping) without losing any spectral information; alternatively, by sampling a time domain signal at a fixed number of equidistant points or the Fourier Transform spectrum at a fixed number of equidistant spectral points will lead to the loss of spectral information (in the case of down-sampling). We employ the Mellin Transform to parameterize speech before performing Hidden Markov Modeling for the following reasons. The Mellin transform followed by HMM modeling inherently incorporates duration variation normalization. The reason is that HMM modeling uses fixed-length, frame-based analysis, which overcomes the durational differences when the frame sequences are similar. However, for analyzing stops, in the VOR can have various abrupt and short duration temporal characteristic changes together with a change in the VOR duration (due to differences in accent), which an HMM will not be able to capture. In addition, previous studies have also shown differences between speakers in the production of voice-on-set times for stops (Allen et al., 2003). Therefore, the Mellin Transform offers an attractive method to represent spectral structure in a potentially more discriminating way which is less duration dependent.

If $\mathcal{F}[\cdot]$ represents the Fourier Transform operator then the Mellin Transform of a function $f(t)$ is given by:

$$\Phi(z) = \mathcal{F}[f(e^t)] = \int_0^\infty t^{z-1} f(t)dt. \tag{6}$$

The Mellin Transform $\Phi(z)$ exists if the integral $\int |f(t)|t^{k-1}dt$ is bounded for some $k > 0$. The M-length Discrete Mellin Transform (DMT) for a N-length vector is then given by:

$$\Phi[m] = \sum_{n=1}^{N} f[n] e^{-j\frac{2\pi m \ln(n)}{M}}, \tag{7}$$

where $m = 0 \cdots (M-1), M \in \mathcal{R}$. Hence, DMT normalizes the duration of the VOR to a specified length by sampling the region at an exponentially spaced M number of intervals and then performing the Discrete Fourier Transform (DFT).

3.2. Discrete Mellin fourier transform (DMFT) analysis

In this feature extraction process, we use the DFT followed by the DMT, since the DFT captures the spectral information at fixed-length intervals. By performing the DMT afterwards, we effectively calculate the DFT again, except at an exponentially spaced fixed number of intervals. The first DFT captures the spectral information without distorting the duration, with the goal of retaining the temporal information. The second DFT (or the DMT) gathers further details of spectral information. Even though the second DMT normalizes the length, it retains more temporal information than by applying the DMT directly to the speech signal.

The Mellin Fourier transform has been applied for image recognition because its resulting spectrum is invariant in rotation, translation, and scale (Chen et al., 1994; Sheng and Arsenault, 1986; Sheng and Duvernoy, 1986; Sheng et al., 1988). In the case of stop-analysis, as mentioned earlier, VOR has abrupt and short duration temporal characteristic changes together with the variation of the entire length, and we aim to separate the impact on accent classification based on duration as represented by VOT versus discriminating changes in spectral structure.

3.3. Discrete wavelet transform (DWT) analysis

As seen from Fig. 1, VOR has sharp changes in the temporal amplitude patterns in the time domain waveform representation and spectral changes (formant and frequency component changes) at the boundary of the stop-vowel transition. These changes can be best captured by a wavelet transform as oppose to the FFT which does not provide sufficient temporal information due to the infinite extent of the Fourier basis functions. The wavelet transform has been used for a wide range of speech processing tasks. Some of the applications of wavelet transform in speech are detection of speech under stress (Sarikaya and Gowdy, 1997), automatic recognition of speech (Erzin et al., 1996), speech enhancement (Bahoura and Rouat, 2001), to name only a few.

The continuous wavelet transform of a discrete sequence, $x[n]$, is defined as the convolution of $x[n]$ with a scaled version of a finite duration wavelet, $\Psi\left[\frac{n}{s}\right]$ (Torrence and Compo, 1998):

$$W_n(s) = \sum_{l=0}^{N-1} x[l] \Psi\left[\frac{(n-l)\Delta t}{s}\right], \tag{8}$$

Table 1
Speaker corpus information for vot based accent classification.

Chinese (CH)	CH-Spkr 1	CH-Spkr 2	CH-Spkr 3	CH-Spkr 4
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
Indian (IN)	IN-Spkr 1	IN-Spkr 2	IN-Spkr 3	IN-Spkr 4
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

where $x[n]$ is of length N, $\Psi\left[\frac{n}{s}\right]$ is the wavelet kernel function, scaled by the factor s, and $\Delta t = \frac{1}{F_s}$, and F_s is the sample frequency. The Morlet wavelet was employed for the kernel function in this study. To approximate the continuous wavelet transform, we first calculate the N-point FFT of $x[n]$ and $\Psi\left[\frac{n}{s}\right]$ separately and then employ the inverse N-point FFT of their product. $\widetilde{W}_n(s)$, the N-point wavelet transform of $x[n]$, which is the approximation of the continuous wavelet transform, $W_n(s)$, is given by:

$$\widetilde{W}_n(s) = \mathcal{F}^{-1}\left[\hat{x}(k)\,\widehat{\Psi}(s\omega_k)\right], \tag{9}$$

where the angular frequency is defined as:

$$\omega_k = \begin{cases} \frac{2\pi k}{N\Delta t} & : k \leqslant \frac{N}{2} \\ -\frac{2\pi k}{N\Delta t} & : k > \frac{N}{2} \end{cases},$$

and $\hat{x}(k)$ and $\widehat{\Psi}(s\omega_k)$ are the Fourier Transform of $x[n]$ and $\Psi\left(\frac{t}{s}\right)$ as follows:

$$\hat{x}(k) = \mathcal{F}[x[n]],\ \widehat{\Psi}(s\omega_k) = \mathcal{F}\left[\Psi\left(\frac{t}{s}\right)\right].$$

Eq. (9) represents the DWT transform of $x[n]$ for a given wavelet scaled with s. As s increases, the DWT results in higher-frequency resolution. Therefore, for a given range of s, smaller values result in higher-temporal resolution while larger values of s produce higher-frequency resolution. We kept the product $\hat{x}(k)\,\widehat{\Psi}(s\omega_k)$ intact and used it for our feature extraction process. Since we consider other features in this study that focus on the spectral domain, we did not perform the Fourier inverse in Eq. (9).

4. Experimental evaluation

Having established the proposed VOR detection algorithm and the parameterization of VOR with four spectral-analysis features, we now turn to an evaluation for VOT estimation and its application to accent classification using HMM models. A number of studies have considered automatic speech processing and/or acoustic/phonetic based analysis approaches to accent classification (Arslan and Hansen, 1996, 1997; Berkling, 2002; Das (Gray) and Hansen, 2004; Ghesquiere and Compernolle, 2002; Gray, 2005; Gray and Hansen, 2005; Kumpf and King, 1996, 1997). Work has also been focused on human perception of accent (Flege, 1984, 1988), as well as the perception of accent (McGory et al., 2001) and language structure based on intonation characteristics (Grover et al., 1987), and phonetic differences in early language differentiation (Johnson and Wilson, 2002). Major (2001) also provides a treatment of the ontogeny and phylogeny of second language phonology as it relates to non-native accent. Comrie (1990) provides an extensive treatment of the world's languages, which is estimated to exceed 6000. With the diversity of languages spoken worldwide, signal processing schemes for the effective characterization of accent is important. Manually marked VOT has previously been shown to be an accent sensitive feature by Arslan and Hansen (1996, 1997), however it has not seen use in automatic accent classification scenarios because of the absence of reliable VOT estimation strategies. It is suggested that the proposed automatic methods would offer a viable path to explore and incorporate this trait into accent classification systems.

Table 2
Results of VOT detection algorithm using TEO based processing (detection accuracy, %).

	Catch	Pump	Target	Total
American English (AE)	76.06	68.12	95.65	79.90
Chinese (CH)	83.33	86.76	91.55	87.32
Indian (IN)	54.72	36.36	45.61	47.73
Average	72.63	71.70	79.70	74.91

4.1. CU-Accent corpus and selected test-platform

CU-Accent corpus (CU-Accent, 2010; crss.utdallas.edu) was organized and collected to facilitate studies in accent modeling and automatic classification as well as speaker identification. The corpus consists of 181 speakers (72 males, 107 females, 2 unclassified speakers) from a large number of accent groups, such as native American English (AE), Chinese (Cantonese, Mandarin), Indian (Bengali, Hindi, Marathi, Tamil etc.), Romance languages (Italian, French, Spanish), Thai, Turkish and many others. The accented speech is collected using an online automated recording system via telephone connection. Subjects are asked to dial into the interactive system and produce speech at the audio prompts using a regular handheld telephone system (i.e., not a cordless or a cellular phone). The majority of the

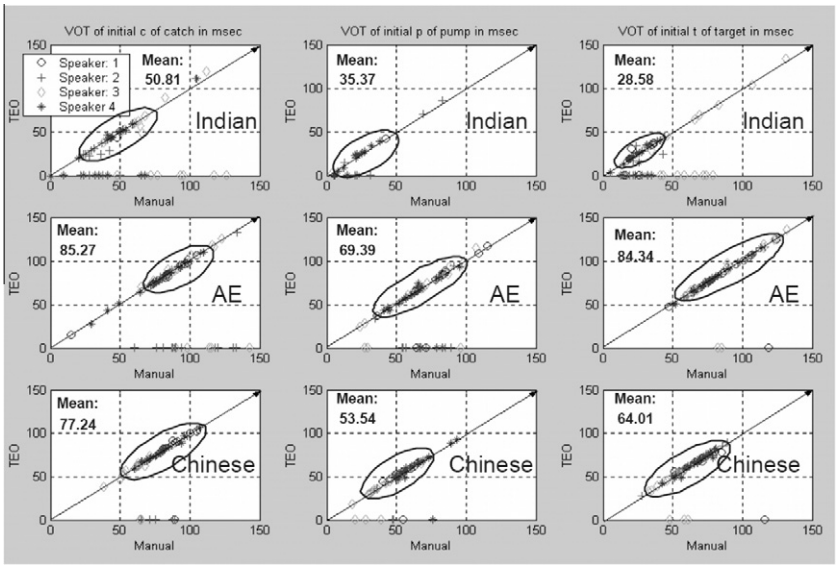

Fig. 5. VOT of /k/, /p/ and /t/ in 'catch', 'pump' and 'target' detected using TEO is plotted against VOT detected manually for all four speakers of three language groups, Indian, AE and Chinese.

speakers were from the Boulder, Colorado (USA) area. Speakers were asked to pronounce some predetermined isolated words (five times for each word), a few predetermined short sentences in English and in their native language, and one minute of spontaneous speech (on a topic of their choice). The subset of the words and sentences were previously employed for an earlier accent corpus collection (Arslan and Hansen, 1997). The speakers are motivated to make multiple calls with at least a day separation between each call, with up to four separate sessions. More than 50% of the entire database contains speakers with multiple sessions. The system consists of an ATT DSP32 based Gradient/DeskLab216 unit connected via an ISDN telephone line and SUN UNIX Operating System. The speech is digitally recorded and stored at 8 kHz sample rate and 16-bit linear PCM format as raw audio files.

Our evaluations here consist of three language groups-AE, Chinese and Indian. Four male speakers from each language group were selected from the CU-Accent Corpus. Table 1 describes each speaker (from Chinese and Indian accent groups). Note, AE is considered standard English (without non-native accent) and all AE speakers in the CU-Accent Corpus were born in the United States. Three words-'catch', 'pump' and 'target' are chosen to detect VOT of the word-initial /k/, /p/ and /t/ respectively. Each speaker spoke each word five times in four separate recording sessions (for a maximum of 20 tokens per speaker). Multiple recording sessions ensured session-to-session variability. The total data size employed consists of 546 tokens.

4.2. Evaluation of VOT detection algorithm using TEO

First, we consider the accuracy of VOT detection for all speakers. Table 2 summarizes VOT detection algorithm results for each accent group. The accuracy was calculated as (total number of tokens which are detected with less than 10% error using the automated VOT detection algorithm) over (total number of tokens). The error in VOT estimate is calculated as:

$$\text{Error} = \frac{|x - \hat{x}|}{x + 0.001} \times 100\%, \tag{10}$$

where x is the manually determined VOT and $\hat{x}$ is the VOT detected using the proposed algorithm. The value 0.001 ms is added to the denominator to avoid the cases where x is 0.

The VOT detection rates (for the average of the three accents shown in Table 2) are 72.63% for /k/, 71.70% for /p/, and 79.70% for /t/ (with detection rates better for the native AE versus than non-native speakers for /k/ and /t/). The VOT detection rate (for the average of all the stops /p/, /t/ and /k/) using TEO is low for Indian English

Table 3
Mean VOTs: (Human calculated – TEO estimation) in ms.

American English (AE)	AE-Spkr 1	AE-Spkr 2	AE-Spkr 3	AE-Spkr 4
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 (CH)	CH-Spkr 1	CH-Spkr 2	CH-Spkr 3	CH-Spkr 4
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
Indian (IN)	IN-Spkr 1	IN-Spkr 2	IN-Spkr 3	IN-Spkr 4
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

Table 4
Successful accent detection/classification rates for VOT-accent model. AE: American English, CH: Chinese and IN: Indian accented English.

Stops	Pairwise detection				Classification
	AE vs. CH	AE vs. IN	CH vs. IN	Pairwise Avg	AE-CH-IN
/k/	64.23	80.80	78.99	73.71	61.38
/p/	70.80	81.32	67.07	72.24	58.49
/t/	68.57	84.13	83.59	77.48	67.01
Average[a]	67.89	82.26	77.42	74.75	62.76

[a] All average performance in this study was calculated accounting for the different number of tokens for each stop and language.

speakers (47.73%), but improves for Chinese and native AE speakers (i.e., 87.32% and 79.90% respectively).

As noted earlier, there are a total of 546 tokens, consisting of 3 words from 12 speakers. Among the 546 tokens, 409 tokens (74.91%) were detected with less than 10% error, with an average ms mismatch between automatic and hand labeled VOT of 0.735 ms (1.15% mismatch).[8] For the atypical cases (the remaining 137 tokens), the mean mismatch is 20 ms, with one such example presented in Fig. 4. Fig. 5 shows scatter plots of human labeled versus automatic detected VOTs. The nine scatter plots represent the VOTs of /*k*/, /*p*/ and /*t*/ in 'catch', 'pump' and 'target' for three accent groups Indian, AE and Chinese respectively. The manually detected VOT is plotted against the VOT detected using the proposed TEO based algorithm. 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 (i.e. for 137 of 546 tokens), then that VOT is plotted on the floor (x-axis) in the plots. Otherwise, if the VOT is detected within a 10% error level (i.e. for 409 of 546 tokens) then it approximately lies along the diagonal line. This implies that aside from some exceptional cases where the TEO based method has failed significantly, the TEO based algorithm has detected VOT with high-accuracy. We suggest that the 10% is a reasonable threshold to employ, since in cases where the VOT error is greater than 10%, it is significantly larger or smaller.[9] Hence, it would be fairly easy to establish confidence boundaries in the VOT estimation procedure to ensure that the estimated VOT would be reliable for accent classification. The result from scatter plots confirms that when we are able to estimate the VOT, it conforms to human manual measurements.

Finally, Table 3 shows the mean VOTs calculated manually and using the TEO algorithm for the 409 tokens (out of 546 token), which were the detected within a 10% error level. Each entry is formatted as: (mean of VOT detected manually) vs. (mean of VOT detected using TEO), all in ms. The results are very encouraging, given a 1.15% mismatch in ms for the automatic scheme, and considering time savings via machines vs. human labeling.

4.3. Evaluation of VOT Model-based accent classifier

In this section, the accent detection/classification based on the VOT obtained by our VOT detection algorithm is evaluated. The VOT detected using TEO for a given accent group is modeled by the best fit probability distribution function, chosen from the Gamma, Weibull and Gaussian distributions. It is noted that the Gaussian distribution, which is symmetric, does not always accurately represent the set of VOTs from a given accent group. Hence, Gamma and Weibull distributions are also employed together with Gaussian distributions. The Kolmogorov–Smirnov test (KS-test) is used to find the appropriate distribution. The KS-test is applicable to unbinned distributions that are functions of a single independent variable (see NIST/SEMATECH (2005) *e*-Handbook website for details). Thus, for comparing $S_N(x)$, the cumulative probability distribution of the given data set under test, $\{x\}$, to the i^{th} known theoretical distribution (Gamma, Weibull etc.), whose cumulative probability distribution is $P_i(x)$, the KS statistic is used as,

$$D_i = \max_{-\infty<x<\infty} |S_N(x) - P_i(x)|.$$

The 'significance level' of an observed value of D_i (as a 'disproof of the null hypothesis' that the distributions are the same) is calculated using a monotonically decreasing function, which is inversely proportional to D_i. The distribution with the highest 'significance level' (implying that $D_i \to 0$) is selected as the best fit.

Accent detection/classification is performed for each possible pair of accents and also among the three given accents. Open speaker accent classification is employed for every test by training the VOT distribution model with tokens from two speakers and tested on the two outside speakers in a round-robin fashion.

Fig. 6 shows the distribution of manually extracted VOTs for AE, Chinese and Indian, when all speakers of a given language group are pooled together. It shows that Indian VOT distributions are well separated from Chinese and AE distributions. VOTs for Indian speakers are shorter compared to American and Chinese speakers. On the other hand, AE and Chinese VOT distributions have measurable overlap.

[8] Recent studies on VOT detection have shown 20–30% error rates (Kazemzadeh et al., 2006), and about 80% detection rate (Stouten and Van hamme, 2009) with 10 ms precision.

[9] Using a prior probability density distribution of VOTs, it is possible in theory to automatically determine these outliers without true VOT knowledge. However, in this study we did not employ this part in order to assess accent dependencies.

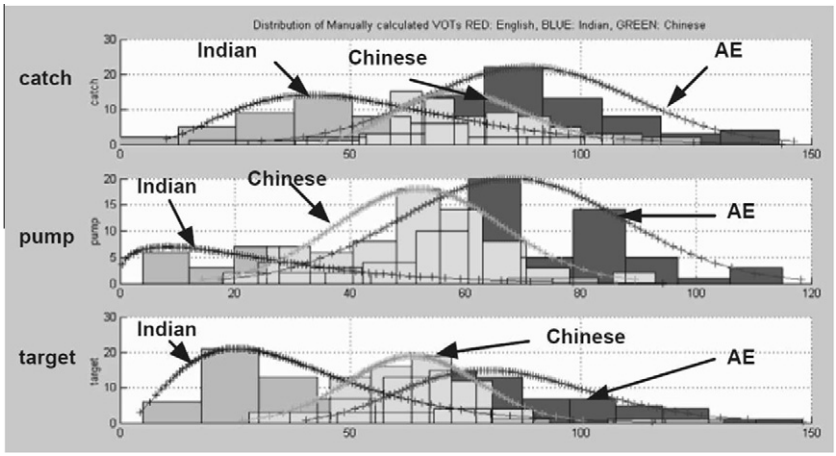

Fig. 6. Distributions of manually extracted VOTs for three accent groups-AE, Indian and Chinese when all speakers in a language (American English, Chinese, and Indian accent) are pooled together.

Table 4 shows the results of accent detection/classification using VOT as the discriminating factor, with the highest pair-wise successful detection rate occurring for the detection between AE and Indian accents (82.26% on average). The successful pair-wise detection rate for AE vs. Chinese accents is the lowest (67.89% on average) due to the similarity in their VOTs. The average 3-way confusion rate combining errors between AE and Chinese is the highest (28.50%), followed by error between Indian and Chinese (17.93%), and AE and Indian is the lowest (7.89%). The average accent detection/classification rates are comparable among all unvoiced stops, /k/,/p/ and /t/, showing that VOT has equal importance irrespective of place of articulation (i.e., velar, bilabial, alveolar).

4.4. Evaluation of HMM-Based accent classifier with various feature types

This section presents the evaluation of spectral analysis using four feature processing based parameters – DMT, DMFT, DWTlfr (DWT with the lowest frequency resolution) and DWThfr (DWT with the highest frequency resolution) for the problem of accent classification. The estimated VOR using the TEO based algorithm and 128 ms of the following vowel region is used for spectral analysis to extract the desired feature parameters. A 16 ms analysis window with a 50% overlapping, consecutive rectangular window sequence is used for feature extraction. The spectrum of the 16 ms window, after the feature transform, is linearly partitioned into five sub-frequency bands and the energy of each sub-band is determined. Therefore, a 16 ms window is converted into a vector of length five. A five state HMM representing different parts of the VOR, the burst, the aspiration and the transition of the stop to the following vowel, along with a five mixture Gaussian output per state is used to model the VOR-vowel sequences for all speakers in each accent group.

As identical to the evaluation of VOT-accent model in Section 4.3, open speaker accent classification is employed for every test by training the HMM model with tokens

Table 5
Successful accent detection/classification rates for various features. AE: American English, CH: Chinese and IN: Indian accented English.

Stops	Features	Pairwise detection				Classification
		AE vs. CH	AE vs. IN	CH vs. IN	Pairwise Avg	AE-CH-IN
/k/	DMFT	80.42	67.50	71.41	73.79	50.94
	DMT	47.08	43.06	48.66	46.35	37.41
	DWTlfr	58.00	44.75	50.45	51.72	32.16
	DWThfr	50.02	54.46	57.99	53.78	38.94
/p/	DMFT	77.64	75.71	52.94	69.92	61.06
	DMT	60.76	78.89	73.87	68.99	53.50
	DWTlfr	71.91	72.22	53.92	66.70	56.55
	DWThfr	47.64	43.33	50.91	47.56	34.07
/t/	DMFT	78.83	64.57	57.14	68.37	53.31
	DMT	61.33	70.06	61.90	64.05	49.75
	DWTlfr	81.02	73.24	69.05	75.27	61.54
	DWThfr	48.92	53.58	49.21	50.37	33.35
Average	DMFT	78.97	68.29	60.93	70.63	54.67
	DMT	56.67	62.43	60.62	59.46	46.63
	DWTlfr	70.98	62.71	58.40	64.93	50.24
	DWThfr	48.87	51.45	52.73	50.74	35.44

from two speakers and tested on the two outside speakers in a round-robin fashion. Table 5 shows the average open test results for accent detection/classification. After collectively assessing the results from our evaluations using performance measures from Table 5, we can make the following observations:

1. DMFT feature shows promising results for every accent classification test. Since DMFT captures both duration and spectral changes of the VOR, it represents the VOR more accurately than any of the other features. The average success rates of this feature are 70.63% and 54.67% for pairwise and across all accents for detection/classification, respectively.
2. DMT feature, which captures only the spectral changes (by normalizing the duration of VOR), shows unsatisfactory results for the word 'catch' but produces comparable results to DMFT for words such as 'target' and 'pump'. This confirms that (as verified by an informal listening test) the word 'target' and 'pump' have substitution of the succeeding vowel after the stop (as oppose to the word 'catch') in accented speech which leads to spectral changes in the VOR. The average success rates of this feature are 59.46% and 46.63% for pairwise and across all accents for detection/classification, respectively. The rates are 66.13% and 51.33%, respectively if the word 'catch' is not considered.
3. DWTlrf, like DMT, also shows comparable performance for words such as 'target' and 'pump', but not for the word 'catch'. Since the words 'target' and 'pump' are associated with vowel substitutions, discriminative movement of articulators are more pronounced for these words than the word 'catch'. DWTlfr is a well established method, as described in several studies, to capture sudden temporal changes in speech versus other FFT based methods. DWTlfr has average success rates of 64.93% and 50.24% for pairwise and across all accents for detection/classification, respectively. The rates are 71.67% and 59.44%, respectively, if the word 'catch' is not considered.
4. DWThfr consistently shows poor performance for all tests, with an average success rates of 50.74% and 35.44% (barely over chance-rates, 50% and 33%) for pairwise and across all accents for detection/classification, respectively. Hence, this feature should not be considered in the future for accent classification.

5. Summary and conclusions

The ability to detect VOT in speech is a challenging problem because it combines temporal and frequency structure over a very short duration. To our knowledge, no successful automatic VOT detection scheme has yet been developed or published in the literature. In this study, the Amplitude Modulation Component (AMC) of the Teager Energy Operator (TEO), a sub-band-frequency based non-linear energy tracking operator, was employed to detect the VOR and estimate VOT. The proposed algorithm was applied to the problem of accent classification using American English, Chinese, and Indian accented speakers. Using 546 tokens, consisting of 3 words from 12 speakers, the average ms mismatch between automatic and hand labeled VOT was 0.735 ms[10] (among the 409 tokens (74.91%), which were detected with less than 10% error). This represents a 1.15% mismatch. It was also shown that average VOTs are different among three different language groups, hence making VOT a good feature for accent classification.

It was also shown that VOR carries spectral cues which can be used for accent detection. Indian speakers generally replace their aspirated stops with unaspirated stops, while Chinese speakers replace their entire stop-vowel sequence with similar sounding foreign stop-vowel phonemes. The ability to capture these spectral cues and changing articulatory movements were considered using one of four feature parameter transformation methods-Discrete Mellin Transform (DMT), Discrete Mellin Fourier Transform (DMFT), and Discrete Wavelet Transform (DWT) (with the highest and the lowest frequency resolution). It was noted that the DMT and DWTlfr transformed features were effective in parameterizing words which exhibit substitution of the succeeding vowel after the stop in accented speech. The average success rate of DMT and DWTlfr transformed features are 66.13% and 71.67%, respectively, when the words 'pump' and 'target' are used for pair-wise accent detection. On the other hand, the DMFT transformed feature worked on for all accent sensitive words. This feature has an average success rate of 70.63% for pair-wise accent detection.

This study has therefore shown that a TEO-spectral based signal processing strategy can effectively be used to estimate VOT, and that this feature shows promise for application in accent classification. Future research could consider automatic outlier rejection of the estimated VOTs based on confidence measures across unvoiced stop-vowel pairs. Furthermore, a wider range of accent types could be explored.

Acknowledgement

This work was supported by the US Air Force Research Laboratory, Rome NY, under contract number FA8750-04-1-0058.

References

Allen, J.S., Miller, J.L., DeSteno, D., 2003. Individual talker differences in voice-onset-time. J. Acoust. Soc. Amer. 113 (1), 544–552.

Arslan, L.M., Hansen, J.H.L., 1996. Language accent classification in American English. Speech Commun. 18, 353–367.

[10] For atypical cases the error was greater than 20 ms, as seen in Fig. 4.

Arslan, L.M., Hansen, J.H.L., 1997. A study of temporal features and frequency characteristics in American English foreign accent. J. Acoust. Soc. Amer. 102 (1), 28–40.

Bahoura, M., Rouat, J., 2001. Wavelet speech enhancement based on the Teager energy operator. IEEE Signal Process. Lett. 8 (1), 10–12.

Berkling, K., 2002. Scope, syllable core and periphery evaluation: Automatic syllabification and foreign accent identification. Speech Commun. 35, 125–138.

Chen, Q.S., Defrise, M., Deconinck, F., 1994. Symmetric phase-only matched filtering of Fourier-Mellin transforms for image registration and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 16 (12), 1156–1168.

Cairns, D., Hansen, J.H.L., 1994. Nonlinear analysis and detection of speech under stressed conditions. J. Acoust. Soc. Amer. 96 (6), 3392–3400.

Chan, A.Y.W., Li, D.C.S., 2000. English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' English pronunciation problems, language. Culture Curriculum 13 (1), 67–85.

Comrie, B., 1990. The World's Major Languages. Oxford University Press, New York.

CU-Accent, 2010. Formally at: <http://cslr.colorado.edu/accent>, now served at: <http://crss.utdallas.edu>.

Das (Gray), S.S, Hansen, J.H.L., 2004. Detection of Voice Onset Time (VOT) for Unvoiced Stops (/p/, /t/, /k/) Using Teager Energy Operator (TEO) for Automatic Detection of Accented English. In: IEEE NORSIG-04: Nordic Signal Processing Symposium, pp. 344–347.

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 1999. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, NY.

Esposito, A., 2002. On vowel height and consonantal voicing effects: Data from Italian, phonetica. Int. J. Phonetic Sci. 59 (4), 197–231.

Flege, J.E., 1984. The selection of French accent by American listeners. J. Acoust. Soc. Amer. 76, 692–707.

Flege, J.E., 1988. Factor affecting degree of perceived foreign accent in English sentences. J. Acoust. Soc. Amer. 84, 70–77.

Francis, A.L., Ciocca, V., Yu, J.M.C., 2003. Accuracy and variability of acoustic measures of voicing onset. J. Acoust. Soc. Amer. 113 (2), 1025–1032.

Fang, M., Häusler, G., 1990. Class of transforms invariant under shift, rotation, and scaling. Appl. Optics 29 (5), 704–708.

Ghesquiere, P.J., Compernolle, D.V., 2002. Flemish accent identification based on formant and duration features. In: ICASSP-02, pp. 749–752.

Gray, S.S., Hansen, J.H.L., 2005. An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system. In: IEEE Automatic Speech Recognition and Understanding Workshop.

Gray, S.S., 2005. Ph.D. Thesis: Speech Science Modeling for Automatic Accent and Dialect Classification. Department of Speech, Language and Hearing Sciences, University of Colorado, Boulder.

Grover, C., Jamieson, D.G., Dobrovolsky, M.B., 1987. Intonation in English, French, and German: Perception and production. Lang. Speech 30 (3), 277–295.

Hansen, J.H.L., Gavidia-Ceballos, L., Kaiser, J.F., 1998. A nonlinear based speech feature analysis method with application to vocal fold pathology assessment. IEEE Trans. Biomed. Eng. 45 (3), 300–313.

Hoshino, A. Yasuda, A., 2003. The evaluation of Chinese aspiration sounds uttered by Japanese students using VOT and power. In: IEEE ICASSP-03, pp. 472–475.

Johnson, C.E., Wilson, I.L., 2002. Phonetic evidence for early language differentiation: Research issues and some preliminary data. The Int. J. Bilingualism 6 (3), 271–289.

Kaiser, J.F., 1990. On a Simple Algorithm to Calculate the 'Energy' of a Signal. In: IEEE ICASSP-90, pp. 381–384.

Kazemzadeh, A., Tepperman, J., Silva, J., You, H., Lee, S., Alwan, A., Narayanan, S., 2006. Automatic detection of voice onset time contrasts for use in pronunciation assessment. In: Interspeech-2006.

Kumpf, K. King, R.W., 1996. Automatic accent classification of foreign accented Australian English speech. In: ICSLP-96, pp. 1740–1743.

Kumpf, K. King, R.W., 1997. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparison with human perception benchmarks. In: EUROSPEECH-97.

Ladefoged, P., 1993. A Course in Phonetics, Third ed. Harcourt Brace College Publishers, Fort Worth.

Levkovitz, J., Oron, E., Tur, M., 1997. Position-invariant, rotation-invariant, and scale-invariant process for binary image recognition. Appl. Optics 36 (14), 3035–3042.

Lopez-Bascuas, L.E., Rosner, B.S., Garcia-Albea, J.E., 2004. Voice-onset time and buzz-onset time identification: A ROC analysis. J. Acoust. Soc. Amer. 115 (5), 2465.

Mahadeva Prasanna, S.R., Sandeep Reddy, B.V., Krishnamoorthy, P., 2009. Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Trans. Audio, Signal, Lang. Process. 17 (4), 556–565.

Major, R.C., 2001. Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology. Lawrence Erlbaum Associates Publishers, New Jersey.

Maragos, P., Kaiser, J.F., Quatieri, T.F., 1993. Energy separation in signal modulations with applications to speech analysis. IEEE Trans. Signal Process. 41 (10), 3024–3051.

McGory, J., Frieda, E., Nissen, S., Fox, R.A., 2001. Acquisition of dialectal differences in English by native Japanese speakers. J. Acoust. Soc. Amer. 109 (5), 2474.

NIST/SEMATECH, 2005. e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/.

Peng, L., Ann, J., 2004. Obstruent voicing and devoicing in the English of Cantonese speakers from Hong Kong. World Englishes 23 (4), 535–564.

Poser, W.J., 2004. Language Log post of 2004-03-13T16:21. <http://itre.cis.upenn.edu/myl/languagelog/archives/000583.html, retrieved 2005-05-08>.

Rosner, B.S., 1984. Perception of voice-onset-time continua: A signal detection analysis. J. Acoust. Soc. Amer. 75 (4), 1231–1242.

Sheng, Y., Arsenault, H., 1986. Experiments on pattern recognition using invariant Fourier Mellin descriptors. J. Opt. Soc. Amer.

Sheng, Y., Duvernoy, J., 1986. Circular Fourier radial Mellin transform descriptors for pattern recognition. J. Opt. Soc. Amer.

Sheng, Y., Lejeune, C., Arsenault, H.H., 1988. Frequency-domain Fourier-Mellin descriptors for invariant pattern recognition. Opt. Eng.

Steinschneider, M., Volkov, I.O., Noh, M.D., Garell, P.C., Howard III., M.A., 1999. Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. The Amer. Physiol. Soc., 2346–2357.

Stouten, V., Van hamme, H., 2009. Automatic voice onset time estimation from reassignment spectra. Speech Commun. 51 (12), 1194–1205.

Sundaram, N. , Smolenski, B.Y. , Yantorno, R., 2003. Instantaneous Nonlinear Teager Energy Operator for Robust Voiced–Unvoiced Speech Classification, <http://www.temple.edu/speech_lab/sundaram.PDF>.

Teager, H., 1980. Some observations on oral air flow during phonation. IEEE Trans. Acoust. Speech, Signal Proc. 28 (5), 599–601.

Teager, H., Teager, S., 1983. A Phenomenological Model for Vowel Production in the Vocal Tract. In: Daniloff, R.G. (Ed.), Speech Science: Recent Advances. College-Hill, San Diego, pp. 73–109.

Torrence, C., Compo, G.P., 1998. A Practical Guide to Wavelet Analysis, Program in Atmospheric and Oceanic Sciences. University of Colorado, Boulder, Colorado. <http://paos.colorado.edu/research/wavelets/bams_79_01_0061.pdf>.

Zhou, G., Hansen, J.H.L., Kaiser, J.F., 2001. Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. 9 (2), 201–216.

Zwicke, P.E., Kiss Jr., I., 1983. A new implementation of the Mellin transform and its application to radar classification of ships. IEEE Trans. Pattern Anal. Mach. Intell. 5 (2), 191–199.