THE RELATIONSHIP OF VOICE ONSET TIME AND VOICE OFFSET TIME TO PHYSICAL AGE

Rita Singh †, Joseph Keshet †, Deniz Gencaga ††, Bhiksha Raj †

† Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
†‡ Department of Computer Science, Bar Ilan University, Israel
†† Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

In a speech signal, Voice Onset Time (VOT) is the period between the release of a plosive and the onset of vocal cord vibrations in the production of the following sound. Voice Offset Time (VOFT), on the other hand, is the period between the end of a voiced sound and the release of the following plosive. Traditionally, VOT has been studied across multiple disciplines and has been related to many factors that influence human speech production, including physical, physiological and psychological characteristics of the speaker. The mechanism of extraction of VOT has however been largely manual, and studies have been carried out over small ensembles of individuals under very controlled conditions, usually in clinical settings. Studies of VOFT follow similar trends, but are more limited in scope due to the inherent difficulty in the extraction of VOFT from speech signals. In this paper we use a structured-prediction based mechanism for the automatic computation of VOT and VOFT. We show that for specific combinations of plosives and vowels, these are relatable to the physical age of the speaker. The paper also highlights the ambiguities in the prediction of age from VOT and VOFT, and consequently in the use of these measures in forensic analysis of voice.

Index Terms— Age, voice onset time, voice offset time, voice forensics, voice biometrics

1. INTRODUCTION

The human voice is increasingly being recognized to be a biomarker. Not only can it be used to match or verify speakers [1, 2], it has also been correlated with many characteristics of the speaker that could be descriptive of speaker’s self and surroundings. There is a burgeoning body of literature that addresses the problem of deriving such biodescriptive parameters from voice recordings, especially the speaker’s biophysical characteristics such as height, weight, age, race etc [3, 4, 5, 6].

The majority of current techniques that attempt to derive these biophysical parameters are based on macro characterizations of the speech signal, i.e. ensemble characterizations of spectral features derived from it. Typically, speech signals are parametrized into collections of Mel-frequency cepstral (or similar) vectors, which may also be used to obtain higher-level representations such as i-vectors which model their distributions [7, 8]. Yet other characterizations include ensembles of utterance- or segment-level measurements such as those obtained from the popular OpenSmile toolkit [9].

In contrast, many studies in the literature that have correlated biophysical parameters to voice are predicated on the fact that biophysical parameters most directly influence the speech-production mechanism. Age, height, weight, physical and psychological health status, etc., affect a variety of physical characteristics such as the size, tension and agility of the vocal cords, the length of the vocal tract, the power and resonance of the voice source, i.e. the lungs, the size and shape of the resonant cavities, muscle response in the vocal apparatus, and many other such factors. These influences manifest in the micro characteristics of the speech signal produced. By micro characteristics, we refer to localized, fine detail of the signal such as the nature of the individual pulses of excitation of the speech signal, the relative energy in periodic and aperiodic components of the excitation, the exact phoneme-specific positions of formants, their bandwidths, and their relation to one another, the width and energy in harmonic peaks, the degree of co-articulation in complex sound sequences, the degree of closure of the velum and cessation of voicing and a plethora of other features that are glossed over by the crude characterizations of macro representations.

Consequently, there is a prevalent belief amongst some researchers that computational approaches that directly key in on these micro features may be expected to be more effective at deciphering physical profiles. Some examples of such studies are [10] that employs estimates of sub-glottal resonances as an aid to estimating body size, [11] that utilizes formant positions, etc. For the most part, however, micro features have not featured prominently in the pantheon of features used for the deduction of biophysical parameters, often due to the difficulty in their accurate measurement at such small scales (typically 20-100ms).

One such micro feature that has repeatedly been reported to relate to many biophysical parameters is the Voice Onset Time (VOT). VOT measures the time between the burst in a plosive and the onset of voicing in the subsequent voiced phoneme. A number of studies have shown VOT to be relatable to the speaker’s age. Correlations between VOT and age have been closely studied in children, since the expectation is that because children’s vocal tracts change rapidly with age [12], VOT may show more changes across ages than seen in adults. In reality, however, this has not been the case [13]. Amongst adults, VOT has been reported to correlate with age but most studies have not evaluated its predictive potential for age. Most studies merely show a direct relation between VOT statistics and the speaker age, e.g. [14]. Some studies have found joint correlations of VOT with age and other parameters, e.g. gender [15], hearing loss [16], age of learning (a second language) [17], age of learning and speaking rate [18]. Along another dimension, joint correlations of VOT and other measures such as Formants and their bandwidths have been found with age [19].

The clear message from all of these studies is that under a variety of conditions VOT has statistical dependence on age, and consequently VOT measurements may be utilized to disambiguate the age
of the speaker, at least to some degree. In this paper, based on these studies, we investigate whether VOT estimates may be utilized to make predictions about age, and whether computational mechanisms may be useful for this purpose. We analyze the relation of VOT to the speaker’s age on a study of 630 speakers from the TIMIT corpus. We note that in contrast to other, previously reported studies on VOT, this analysis employs a much larger corpus of a much greater variety of speakers, while maintaining a phonetic balance and also a balance between genders. Note that the VOT is a fine detail of the speech signal and is hard to characterize accurately. In particular, for relatively large data such as TIMIT, hand-annotating VOTs is not feasible, and we require an automated algorithm that can do so. To this end, we also identify a high-accuracy algorithm that can be used to measure the VOTs in large corpora.

In addition to studying VOT, we also study the Voice Onset Time (VOFT). VOFT can be viewed as the complement of VOT, and measures the duration between the cessation of voicing in a voiced phoneme, and the onset of the burst of the subsequent plosive sound. Although VOFT has been studied in some medically relevant contexts e.g. [20], it is significantly less studied than VOT, VOFT, like VOT, also has a dependence on the physical parameters of the speaker, including, potentially, age. Like VOT, however, it is also hard to measure automatically, which is a requirement if we must analyze large quantities of speech. In this paper we also suggest an automated algorithm to obtain VOFT measurements from the speech signal.

Our experiments arrive at a surprising, if disappointing, outcome. Regardless of how we slice or dice it, VOT is unrelatable to age. Every result contradicts the large body of physiometric literature that claims the opposite. This is not a consequence of incorrect computation - in fact, in a separate exercise that compared manually marked VOTs to those derived by our automated algorithm, we have extensively verified that the automated VOT computation we use is better than human-judged VOT annotation.

The rest of this paper is organized as follows. In Section 2 we describe VOT, VOFT and their measurement in greater detail. In Section 3 we describe our experimental techniques and the results of our experiments, and in Section 4 we present our conclusions.

2. VOICE ONSET AND OFFSET TIMES

Speech sounds may be categorized along a variety of dimensions. One partition is based on voicing – whether the vocal cords vibrate or not during the production of the sound. Voiced sounds include vowel sounds such as /aa/, /uw/, /iy/ or not during the production of the sound. Voiced sounds include vowel sounds such as /aa/, /uw/, /iy/ or not during the production of the sound. Voice Onset Time measures a timing characteristic of sound pairs where a voiced sound is followed by a plosive. VOT is the duration between the cessation of voicing in the voiced phoneme and the onset of the following plosive. In this sense it is the complement of VOT. Fig. 1 shows an illustration of VOFT. Unlike the onset of voicing, the offset of voicing is often hard to discern, and consequently, VOFT is hard to measure.

Fig. 1. Illustration of VOT and VOFT (a) spectrogram of the word TOOTHPASTE showing micro-level variations between phonemes /b/ and /v/ on the spectrogram of an instance of the word DARK, as obtained by the structured-prediction algorithm (c) Example of negative VOT. VOT can have three variations: Zero VOT: the duration between the burst and the subsequent voicing pattern is zero; Positive VOT: there is a measurable duration between the two; Negative VOT: in rare cases, voicing begins before the onset of the stop. In this example the stop and release of /g/ are very faint.

It is generally accepted that VOT and VOFT are indicators of the ability of the vocal tract to move from one configuration to another [21]. In other words, these entities measure the agility of the vocal tract [22, 23], which in turn is thought to be dependent on the age of the speaker, amongst other factors. It is therefore reasonable to expect VOT and VOFT to be statistically related to the speaker’s age, a hypothesis that seems to be borne out by the studies reported in Section 1.

2.2. Voice Offset Time

In contrast to the VOT, Voice Offset Time (VOFT) refers to a timing characteristic of sound pairs where a voiced sound is followed by a plosive. VOFT is the duration between the cessation of voicing in the voiced phoneme and the onset of the following plosive. In this sense it is the complement of VOT. Fig. 1 shows an illustration of VOFT. Unlike the onset of voicing, the offset of voicing is often hard to discern, and consequently, VOFT is hard to measure.

2.3. Estimation of VOT and VOFT

Automatic estimation of VOT is a challenging problem – the onset of voicing is typically a faint cue that is easily missed, as is the initial
burst that signals a plosive. A limited number of approaches have been proposed in the literature for the automatic estimation of VOT. Lin and Wang [24] employ random forest classifiers on cepstral features derived from the signal. Stouten and Hamme [25] propose the use of “reassignment spectra” to estimate the VOT. In all cases, the estimation errors can approach 20ms, which is sometimes as long as the VOT itself.

In this paper, we utilize a structured-prediction approach, originally proposed in [26] to estimate VOT, which has consistently been shown to result in VOT estimates with errors less than 5ms, provided finally proposed in [26] to estimate VOT, which has consistently been the VOT itself.

use of “reassignment spectra” to estimate the VOT. In all cases, the SP model computes a weighted combination \( \sum_{v \in V} \phi_i(X, T_p, T_v) \) of the evidence from all of the feature maps. The boundaries of the voice onset time are estimated as the instants \( t_{X,p} \) and \( t_{X,v} \) at which this score peaks.

\[
\hat{t}_{X,p}, \hat{t}_{X,v} = \arg \max_{t_{X,p}, t_{X,v}} S(X, T_p, T_v) \quad (1)
\]

A learning phase for the detector estimates the weights \( w = [w_1, w_2, \ldots, w_K]^T \) such that the expected error between the estimates given by the above predictor and the true boundaries of the VOT is minimized. To do so we define the following loss:

\[
L_X = \max\{|(t_{X,p} - t_{X,v}) - (\hat{t}_{X,p} - \hat{t}_{X,v})| - \epsilon, 0\} \quad (2)
\]

The weight \( w \) is estimated to minimize the empirical average of the above loss over a large number of training instances, leading to an iterative estimate with the following update rule.

\[
w \leftarrow w + \sum_X \Gamma(\Phi(X, t_{X,v}, t_{X,w}) - \Phi(X, \hat{t}_{X,v}, \hat{t}_{X,w}))
\]

where \( \Gamma \) is a diagonal matrix whose \( i \)th diagonal entry represents the learning rate corresponding to the \( i \)th feature map \( \phi_i(X, t_{X,v}, t_{X,w}) \). This update rule has been proven to converge to a local optimum in [27]. The learned weights may be used in conjunction with Equation 1 to estimate the VOT on test instances.

The same algorithm may be employed to estimate VOFT as well. The only distinction is in the feature maps used, which must now be characteristic for VOFT, rather than VOT. In practice we have found the same feature maps to be effective for both VOT and VOFT. A total of 59 feature maps are used. We refer the reader to [28] for a detailed description of the feature maps.

3. EXPERIMENTS

Experiments were performed using the TIMIT acoustic-phonetic corpus [29]. The corpus consists of 630 speakers representing eight major dialects of American English. The recordings contain 16kHz sampled speech recordings of ten phonetically rich sentences that are read by each speaker. Nearly every stop consonant is represented at least once in both, the VOT and VOFT contexts for each speaker. In all cases, the speech was well-articulated, and the recordings are clean i.e., studio-quality with no noise present. The gender of the speaker and their age at the time of recording have been provided in this corpus.

Our first goal was to evaluate age-related trends in the VOT, such as those observed in other studies reported in the literature. In order to do this, we computed VOT for all words that began with a plosive leading into a vowel sound, distinguishing between voiced and unvoiced bilabial plosives (/b/ and /p/ respectively), voiced and voiceless lingua-alveolar plosives (/d/ and /t/ respectively), and voiced and voiceless lingua-velar plosives (/g/ and /k/ respectively). The top row of Fig. 2 shows the scatter of the actual VOT readings obtained for /k/ and /g/ against the age of the speaker. As we see, there are no noticeable trends with age. Similar lack of trend are observed for other plosives.

<table>
<thead>
<tr>
<th></th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B: /b/</td>
<td>LA: /d/</td>
</tr>
<tr>
<td></td>
<td>LV: /g/</td>
<td>B: /p/</td>
</tr>
<tr>
<td></td>
<td>LA: /t/</td>
<td>LV: /k/</td>
</tr>
<tr>
<td>VOT</td>
<td>0.19</td>
<td>0.16</td>
</tr>
<tr>
<td>VOFT</td>
<td>0.46</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 1. Mutual Information in VOT and VOFT for different plosives. B: Bilabial; LV: Lingua-Velar; LA: Lingua-Alveolar. The italicized numbers were computed on fewer instances than others, using appropriately fewer histogram bins, as suggested by [30].

On the other hand, we also find that the VOTs of the different plosives do not significantly predict one another. The mutual information between the various plosives is shown in Table 2. These two are of the same order as the MI between the plosives and age. Given this, we may speculate that although the VOTs for the individual plosives do not by themselves have significant MI with age, they might
do so jointly since, being effectively independent of one another, the information they individually provide may combine cumulatively.

### Mutual Information

<table>
<thead>
<tr>
<th>Plosive</th>
<th>/b/</th>
<th>/d/</th>
<th>/g/</th>
<th>/p/</th>
<th>/t/</th>
<th>/k/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>1.97</td>
<td>1.70</td>
<td>2.46</td>
<td>0.15</td>
<td>2.77</td>
<td>3.33</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>0.17</td>
<td>0.15</td>
<td>0.10</td>
<td>0.10</td>
<td>0.12</td>
<td>0.22</td>
</tr>
<tr>
<td>/b/</td>
<td>0.11</td>
<td>0.10</td>
<td>2.77</td>
<td>0.10</td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>/d/</td>
<td>0.20</td>
<td>0.21</td>
<td>0.12</td>
<td>0.22</td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>/g/</td>
<td>0.20</td>
<td>0.22</td>
<td>0.13</td>
<td>0.22</td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>/p/</td>
<td>0.20</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>/t/</td>
<td>0.20</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>/k/</td>
<td>0.20</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>3.33</td>
<td>3.33</td>
</tr>
</tbody>
</table>

Table 2. Mutual Information in VOT measures across different plosives. The lower portion of the table is empty since MI is symmetric.

### Table 3. RMS prediction errors on a 10-way jackknife test across phonemes (Ph) and words (Wd) using various regression models.

Differences are for the case where the predicted age is assumed to be the mean age of the training data partition.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>LR</th>
<th>RF</th>
<th>GPR</th>
<th>SLK</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT: Ph</td>
<td>8.24</td>
<td>8.29</td>
<td>9.02</td>
<td>9.02</td>
<td>8.31</td>
<td>9.09</td>
</tr>
<tr>
<td>VOFT: Ph</td>
<td>8.24</td>
<td>8.21</td>
<td>8.78</td>
<td>8.89</td>
<td>8.40</td>
<td>10.96</td>
</tr>
<tr>
<td>VOFT: Wd</td>
<td>8.24</td>
<td>8.22</td>
<td>8.24</td>
<td>8.50</td>
<td>8.18</td>
<td>8.63</td>
</tr>
</tbody>
</table>

Table 3. RMS prediction errors on a 10-way jackknife test across phonemes (Ph) and words (Wd) using various regression models.

To test this hypothesis, we attempted to develop several regression models to predict age from VOT. In each case, the input to the regression was a set of six values, consisting of the mean VOT times for each of the six plosives for the speaker. The predicted variable was the age of the speaker. In each experiment we ran a 10-way jackknife test – we partitioned the 630 speakers into ten sets of 63. For each set, we trained the regression from the remaining 9 sets and used the trained regression to predict the age of that set. Table 3 shows the mean-squared error of prediction obtained with six different age predictors – linear regression (LR), random forest regression (RF), Gaussian process regression (GPR), support-vector regression with a linear kernel (SLK), and a KNN regression (KNN). Support vector regression with other kernels was generally worse than with a linear kernel. For reference, we also show the MSE when we simply predict the age as the global mean (MEAN). For reference, the standard deviation of age in the TIMIT data is 8.23 years. In each case, the mean squared error of prediction is comparable to the standard deviation of the age variable itself, and is, in some cases, actually larger. None of the predictors are able to make any reliable estimates of age.

We considered that we may be losing information by aggregating the VOTs for all instances of a plosive without regard to the following vowel, and that the dependence of the VOT on the following vowel may be significant. So we focused on the four following VOTs: /d/-/aa/ (from “DARK”), /d/-/ow/ (from “DON’T”), /t/-/ax/ (from “TO”) and /k/-/ae/ (from “CARRY”), each of which was uttered by every subject. Table 3 also shows the mean-squared prediction error obtained with each of the regression models, when age was predicted on the basis of these four VOTs. Once again, the prediction error is comparable to the standard deviation of age itself.

This could well be from competing effects introduced by other factors such as height, weight, gender, dialect etc. To gauge the extent of these effects, in other experiments, we partitioned subjects by gender, dialect and height, and attempted to perform predictions.

While we do not report detailed results here for lack of space, segregation of the data by any of these factors did not result in improvement of predictions – the RMS prediction error remained greater than the innate standard deviation of age in all cases.

---

### 4. CONCLUSIONS

From our experiments we conclude that contrary to popular belief, VOT is not predictive of the age of the speaker across a large ensemble of speakers. Note that this observation does not preclude the presence of predictive VOT-age trends for much more carefully selected groups of speakers, as have been chosen in most earlier studies. In addition, our results indicate that VOFT may also be worth exploring in more detail as an age-profiling tool.

The fact that the results in this paper largely do not support those in most reported literature may be due to two factors. The first is that most earlier results were obtained on smaller amounts of data from subjects who were carefully selected to eliminate secondary factors. Some trends may be purely illusory. Fig. 3 shows one such example. For the voiced lingua-alveolar plosive /d/ in the context of /æ/, we appear to observe a trend that allows us to use the VOFT value to establish an upper limit on the age of the speaker. Closer inspection shows the VOFT to segregate into two groups, a high-occurrence cluster between 15-18ms, and a second more spread out one. Once separated, the trend disappears. A likely second factor is the aggregate error made in the estimation of VOT (and VOFT). Although our VOT predictor is highly accurate, with a mean error of less than 5ms, for micro-features small errors may eliminate patterns. Unfortunately both of these factors are likely to affect characterizations based on any micro-factor. This does not imply that micro features in general may not be useful for profiling. Rather, this work may be viewed as a caution that patterns observed in small-scale human studies may not appear in larger-scale automated analyses.
5. REFERENCES


