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

Rita Singh[1], Joseph Keshet[2], Deniz Gencaga[3], Bhiksha Raj[1]

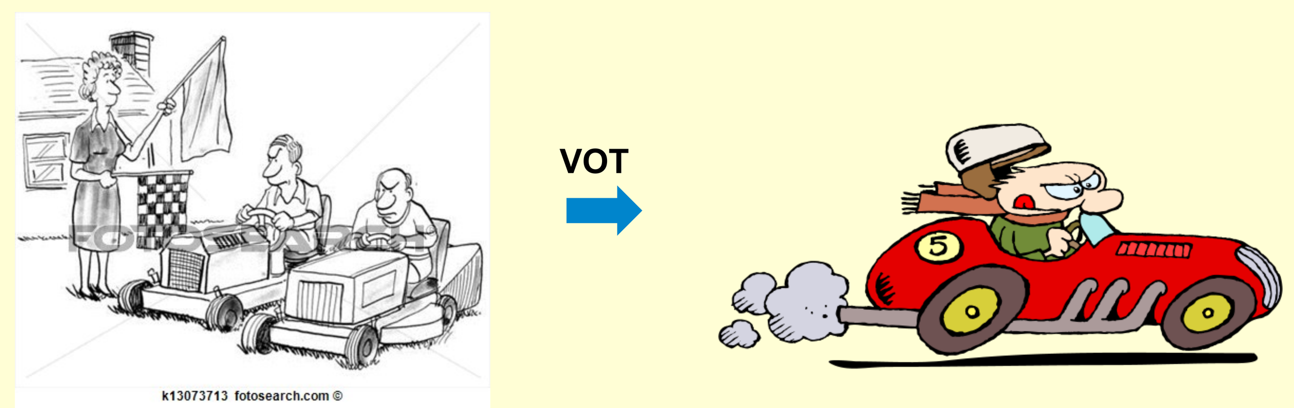
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
Department of computer Science, Bar Ilan University, Israel
Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
rsingh@cs.cmu.edu denizg@cs.cmu.edu joseph.keshet@biu.ac.il bhiksha@cs.cmu.edu

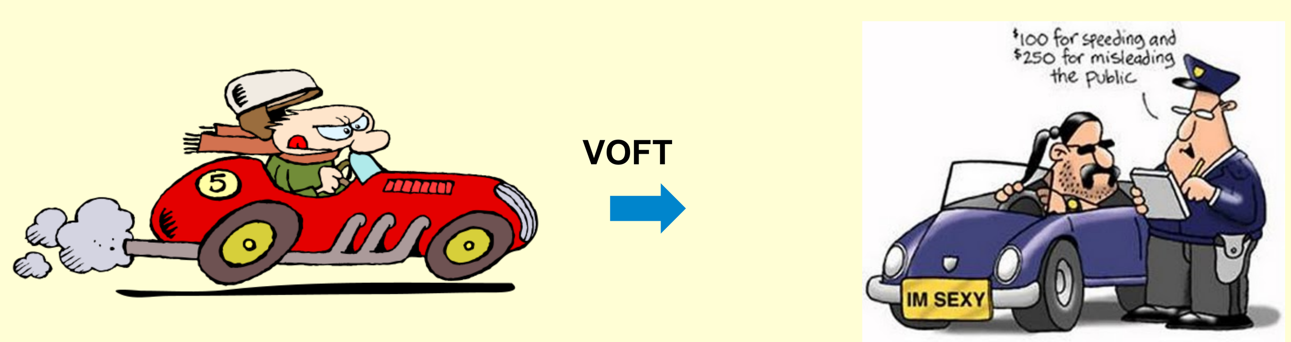
Abstract

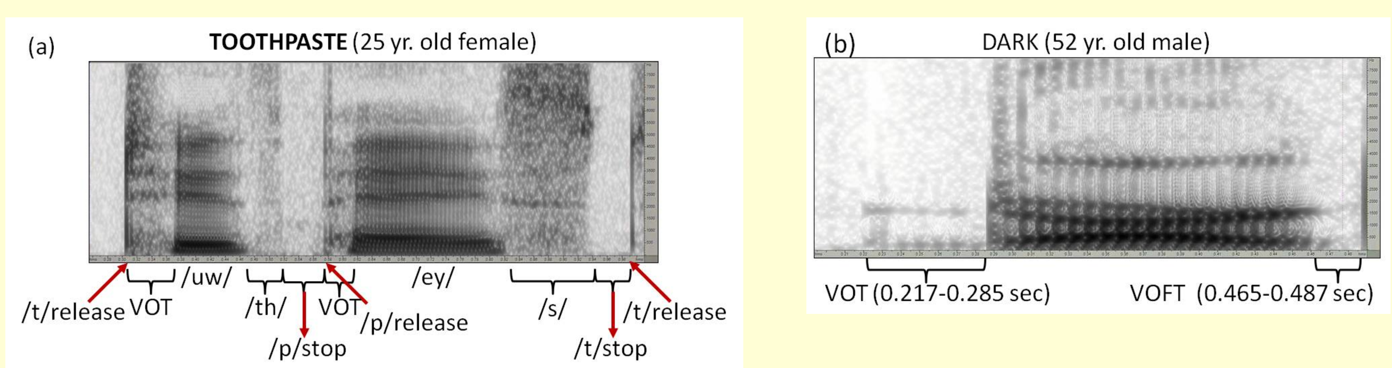
In a speech signal, Voice Onset Time (VOT) is the period between the release of a plosive and the onset of vocal cord vibrations in the production of the following sound. Voice Offset Time (VOFT), on the other hand, is the period between the end of a voiced sound and the release of the following plosive. 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.

Vot is VOT and VOFT?

When a plosive is followed by a voiced sound, the vocal cords go from a state of rest to state of motion (vibration) in a very short time. This is the voicing onset time (VOT). The time taken for vibrating vocal cords to stop is the Voicing Offset time (VOFT).







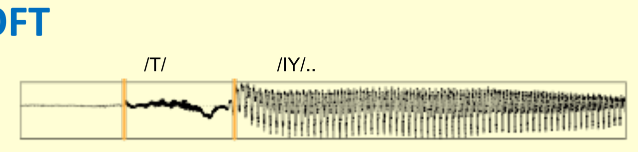
The hypothesis

It is generally accepted that VOT and VOFT are indicators of the ability of the vocal tract to move from one configuration to another. In other words, these entities measure the agility of the vocal tract, 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. We believe that with a better VOT/VOFT estimator, the correlations will be stronger than those reported in the literature.

The result

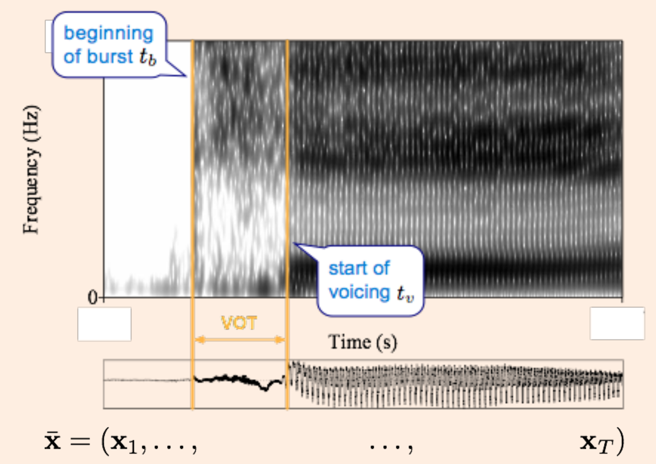
- In spite of multiple claims in the literature to the contrary, we did not see significant correlations between VOT and age. VOFT was better correlated

Estimation of VOT/VOFT



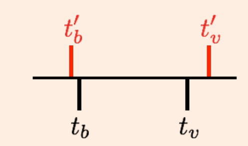
VOT/VOFT are difficult to estimate accurately : VOT and VOFT are of the order of milliseconds
We use a structure prediction algorithm that outperforms humans.

Estimation of VOT and VOFT



predicted VOT $= t'_v - t'_b$

manual VOT $= t_v - t_b$



$$L\left((t_b, t_v), (t'_b, t'_v)\right) = \max\{|(t'_v - t'_b) - (t_v - t_b)| - \epsilon, 0\}$$

Find VOT predictor such that

$$\mathbb{E}\left[L((t_b, t_v), (t'_b, t'_v))\right]$$

is minimized.

Define several "expert predictors"

- $\Phi_k(x, t_b, t_v)$
- Expected to (but may not actually) achieve a maximum when t_b and t_v are true start and end of VOT period

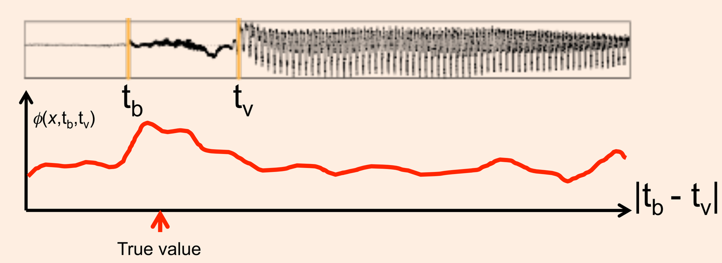
Define a weighted combination of experts

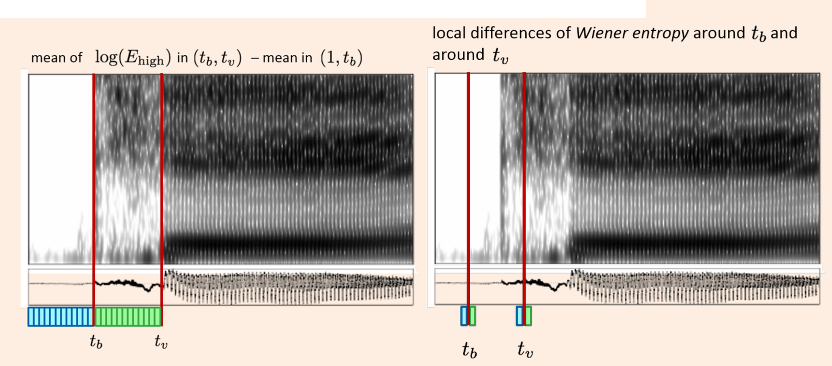
- $S(x, t_b, t_v) = \sum_k w_k \Phi_k(x, t_b, t_v)$

Overall predictor: find t_b and t_v where the expert's score is maximum

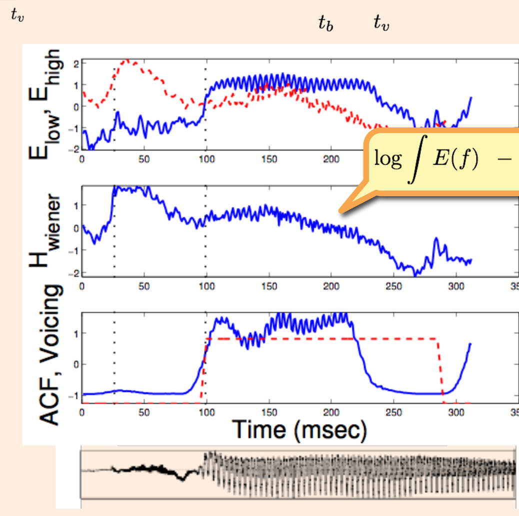
$$(t'_b, t'_v) = \arg\max_{(t'_b, t'_v)} S(x, t'_b, t'_v)$$

Challenge: What must w_k be for the mixture expert to really be a highly-accurate expert?





Define signal-characteristic motivated functions $\phi(x,t_b,t_v)$ which are expected to peak if we get t_b and t_v right, but may not necessarily do so. We define several such functions.



	Value at t_b	Value at t_v	$\Delta_a^{t_b}$	$\Delta_a^{t_v}$	Mean, max Δ_a^t in $[t_b, t_v]$	Mean, max Δ_a^t in $[t_b, t_v - 10]$	Mean in $[t_b, t_v]$ - mean in $[1, t_b]$	Max in $[1, t_b - 5]$	mean in $[1, t_b - 5]$
$\log(E_{low})$		x	x	x					
$\log(E_{high})$	x	x	x	x			x	x	x
$\log(E_{tot})$	x	x	x	x					
H_{wiener}	x	x	x	x			x	x	x
Autocor.		x	x	x	x	x			
Pitch		x							
Voicing		x							x

Local differences $\Delta_a^t(F)$ is the difference between mean of F In $(t-a, t)$ and $(t, t+a)$ where $a \in \{5, 10, 15\}$ ms

1. Define feature maps
2. Define mixture of expert predictor
3. Learn weights from training data
4. Employ on test data to identify and measure VOTs

How accurate is it?
More accurate than humans staring at spectrograms! Definitely more accurate than standard signal processing techniques (By up to 50ms)

Max-margin algorithm (large margin classifier)

Input: training set $S = \{(\bar{\mathbf{x}}_1, t_{b_1}, t_{v_1}), \ldots, (\bar{\mathbf{x}}_M, t_{b_M}, t_{v_M})\}$

Initialize: $\mathbf{w}_0 = 0$

For each example $(\bar{\mathbf{x}}_i, t_{b_i}, t_{v_i})$

Predict: $(t'_b, t'_v) = \arg\max_{(t_b, t_v)} \mathbf{w}_{i-1} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}_i, t_{b_i}, t_{v_i}) + L((t_b, t_v), (t_{b_i}, t_{v_i}))$

Update: $\mathbf{w}_i = \mathbf{w}_{i-1} + \tau_i[\boldsymbol{\phi}(\bar{\mathbf{x}}_i, t_{b_i}, t_{v_i}) - \boldsymbol{\phi}(\bar{\mathbf{x}}_i, t'_b, t'_v)]$

Output $\mathbf{w}^* = \sum_i \mathbf{w}_i$

Experiments

Data

TIMIT Database: 630 speakers. 10 utts/speaker
Training set: 462 speakers, 136 F and 326 M
Test set: 168 speakers, 56 F and 112 M

	Voiced			Unvoiced		
	B: /b/	LA: /d/	LV: /g/	B: /p/	LA: /t/	LV: /k/
VOT	0.19	0.16	0.16	*0.12*	0.18	0.20
VOFT	*0.46*	0.18	0.27	0.18	0.21	0.27

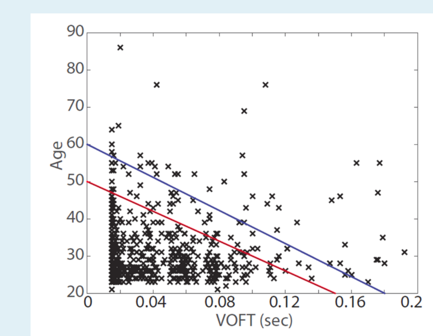
Table 1. Mutual Information in VOT and VOFT for different plosives. B: Bilabial; LV: Lingua-Velar; LA: Lingua-Alveolar. The italicized numbers were computed on fewer instances than others, using appropriately fewer histogram bins.

	Mutual Information					
	Voiced			Unvoiced		
Plosive	/b/	/d/	/g/	/p/	/t/	/k/
/b/	1.97	0.15	0.17	0.11	0.20	0.20
/d/		1.70	0.15	0.10	0.10	0.17
/g/			2.46	0.10	0.21	0.22
/p/				2.77	0.12	0.13
/t/					3.33	0.22
/k/						3.30

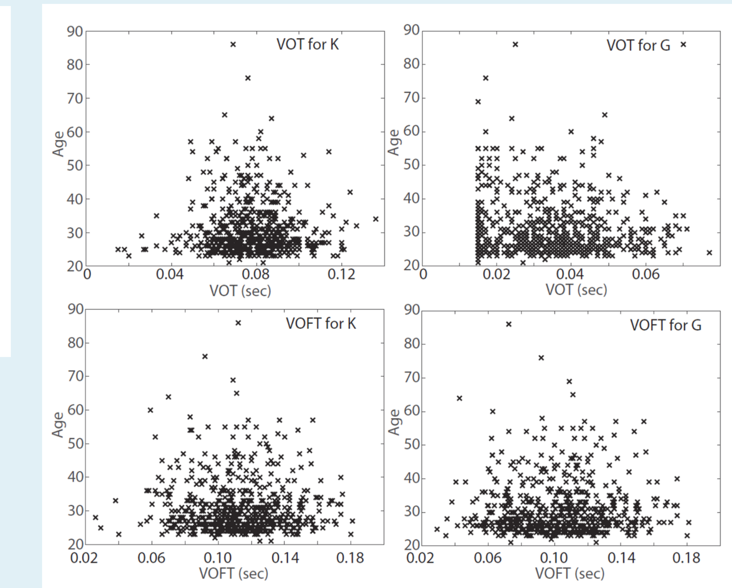
Table 2. Mutual Information in VOT measures across different plosives. The lower portion of the table is empty since MI is symmetric.

Measure	Mean	LR	RF	GPR	SLK	KNN
VOT: Ph	8.24	8.29	9.02	9.02	8.31	9.09
VOT: Wd	8.24	8.26	8.69	9.33	8.26	9.85
VOFT: Ph	8.24	8.21	8.78	8.89	8.40	10.96
VOFT: Wd	8.24	8.22	8.24	8.50	8.18	8.63

Table 3. RMS prediction errors on a 10-way jackknife test across phonemes (Ph) and words (Wd) using various regression models. Highlighted numbers are for the case where the predicted age is assumed to be the mean age of the training data partition.



Illusory age-limiting trend exhibited by VOFT for /d/ following the phoneme /ae/. For any given VOFT, it is possible to assign an upper limit to the age of the person with high accuracy. 86% of all instances lie below the lower line. 95% lie below the upper line.



Scatter plots for VOT and VOFT of plosives /k/ and /g/ against age.
Top: VOT. Bottom: VOFT.

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 /ae/, 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.

References

- Dennis H Klatt, "Voice onset time, frication, and aspiration in word-initial consonant clusters," Journal of Speech, Language, and Hearing Research, vol. 18, no. 4, pp. 686–706, 1975.
- Pascal Auzou, Canan Ozsancak, Richard J Morris, Mary Jan, Francis Eustache, and Didier Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," Clinical Linguistics & Phonetics, vol. 14, no. 2, pp. 131–150, 2000.
- H. Van Hamme and Stouten V, "Automatic voice onset time estimation from reassignment spectra," Speech Commun, vol. 51, no. 12, pp. 1194–1205, 2009.
- Morgan Sonderegger and Joseph Keshet, "Automatic measurement of voice onset time using discriminative structured prediction)," The Journal of the Acoustical Society of America, vol. 132, no. 6, pp. 3965–3979, 2012.
- David W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605-610, 1979.