Automating Phonetic Measurement: The Case of VOT

Neville Ryant, Jiahong Yuan, and Mark Liberman
Linguistic Data Consortium, University of Pennsylvania

Of the 58 papers published in volume 40 of Journal of Phonetics, 16 (28%) feature Voice Onset Time (VOT) or related measurements, confirming that VOT remains a central concern in the field. However, phoneticians' VOT measurements generally continue to rely on human judgment, which requires significant labor, makes even large laboratory experiments onerous, and prevents the field from taking full advantage of the millions of hours of digital speech now becoming available. Consequently, we present an algorithm for accurate automatic measurement of VOT and compare its measurements with those of humans for a series of corpora.

At its core the VOT measurement process reduces to accurately locating two acoustic events in the stop region: the initial burst of energy accompanying the stop release and the point at which voicing begins for the following vowel. VOT, then, is just the duration of the interval spanning burst onset and voicing onset. Underlying our algorithm is the intuition that both the stop burst and point of voicing onset should be reflected as large positive peaks in the decision functions of classifiers trained to discriminate frames immediately surrounding the relevant acoustic events from more distant frames. The measurement process, then, becomes a simple exercise in signal processing and machine learning.

Specifically, the algorithm proceeds as follows. First, within the stop region (identified via forced-alignment between the recording and its transcript) a series of acoustic features (log energy in the 0-8 kHz, 0-500 Hz, and 0-3 kHz bands, spectral entropy, and spectral centroid) is extracted every ms, yielding a timeseries of feature vectors, which, along with its first and second differences, is then projected into scale space (Witkin, 1984) via convolution with a series of gaussians. Following creation of this multiscale representation, at each frame we evaluate the decision function of a support vector machine burst onset classifier with radial basis function kernel[1] trained on 1,774 stop bursts from the TIMIT training set and the time t_b of the largest positive peak in this decision function is recorded. We then evaluate the decision function of a similarly trained voicing-onset classifier at each frame following the burst, recording the time of its highest positive peak as t_v. If either burst onset or voice onset detection fails, VOT measurement fails; otherwise, the VOT is recorded as $t_v - t_b$.

In the left panel of Figure 1 we plot the cumulative distribution of the system errors for stops in the full TIMIT test set, a corpus consisting of clean lab speech collected for another experiment (hereafter, LAB), and the BU Radio Speech Corpus (hereafter, BU). System performance is excellent, with >85% of errors under 10 ms for the TIMIT test set (mean: 4.67 ms), a number comparable to previous attempts on this data by Stouten and Van hamme (2009) and Sonderegger and Keshet (2010). Promisingly, performance does not degrade when evaluated on a novel domain; indeed, for BU 85% of errors fall within 10 ms (mean: 5.88 ms) and for LAB nearly 100% of errors are within 10 ms (mean: 2.8 ms). Moreover, on a subpart of LAB (n=229 stops) independently annotated by two of the authors, the system agreed with the two human annotators as well as they agreed with each other (see right panel of Figure 1). Given that the algorithm is designed to not attempt VOT measurement in cases where the bottom up information is ambiguous, the question naturally arised as to whether this introduces any bias. However, this does not appear to be the case, as seen in Figure 2, where the the 5th, 25th, 50th, and 95th percentiles for human and system measurements of the voiceless stops in TIMIT are essentially identical.

Our initial forays into automating VOT measurement have been promising, with a system trained on TIMIT achieving human-like performance for that corpus and generalizing quite well to a previously unseen selection of lab speech. This strongly suggests that in the near future human measurements could be replaced by automated measurements, opening the door to extremely large scale work that was previously impossible.

[1] In point of fact following Rahimi and Recht (2007) we approximate the implicit feature mapping of the RBF kernel using random Fourier features – cosines of random affine projections of the data. With sufficiently many such futures it is possible to retain the ability of kernel machiens to fit nonlinear decision surfaces while avoiding the high computational costs incurred in calculation of the kernel matrix.

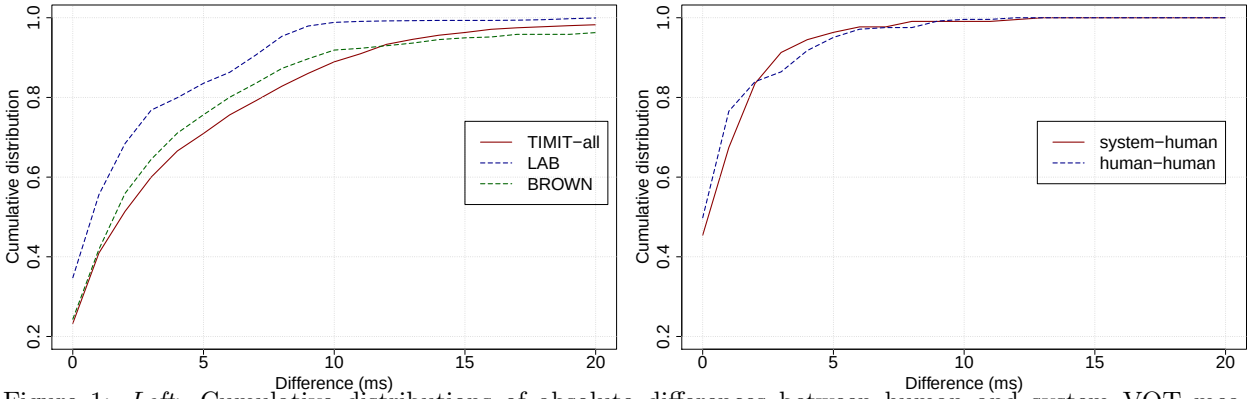

Figure 1: *Left*: Cumulative distributions of absolute differences between human and system VOT measurements for all occurrences of /p, t, k/ in the full TIMIT test set (n=3158), word-initial /p, b/ in LAB (n=2263), and /p, t, k/ in BU Radio Speech (n=931). *Right*: Cumulative distributions of human/system and human/human differences for speaker 9 in LAB (mean differences: system-human, 1.77 ms; human-human, 1.79 ms).

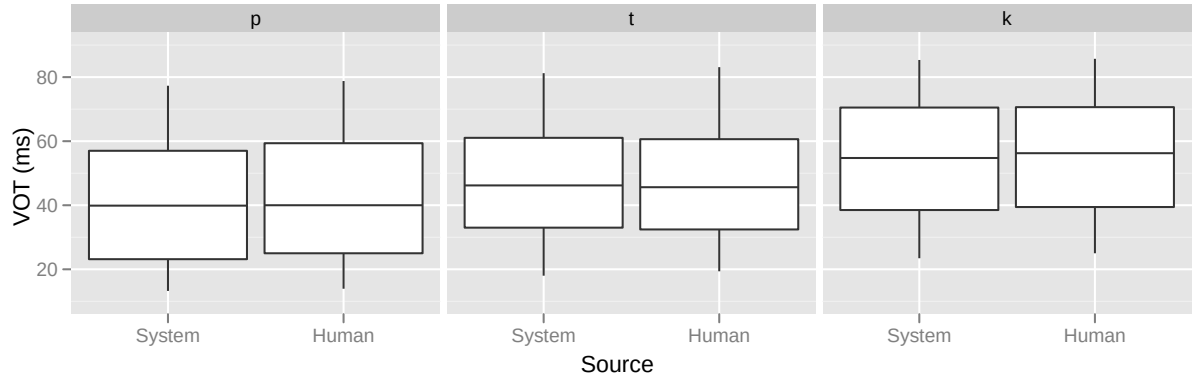

Figure 2: Boxplots displaying 5th, 25th, 50th, 75th, and 95th percentiles for human and system VOT measurements of the stops /p, t, k/ in the full TIMIT test set. Mean (s.d.): human /p/ 43.5 ms (21.7 ms), /t/ 50.8 ms (22.8 ms), /k/ 58.9 ms (22.2 ms); system /p/ 41.5 ms (22.0 ms), /t/ 49.9 ms (23.1 ms), /k/ 57.3 ms (22.8 ms).

References

Bottou, L., Bousquet, O., 2008. The tradeoffs of large scale learning. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems. Vol. 20. pp. 161–168.

Rahimi, A., Recht, B., 2007. Random features for large-scale kernel machines. Proceedings of NIPS 2007.

Sonderegger, M., Keshet, J., 2010. Automatic discriminative measurement of voice onset time. In: Proceedings of Interspeech 2010. pp. 2242–2245.

Stouten, V., Van hamme, H., 2009. Automatic voice onset time estimation from reassignment spectra. Speech Communication, 1194–1205.

Witkin, A., 1984. Scale-space filtering: A new approach to multi-scale description. In: Proceedings of ICASSP 1984. pp. 150–153.