INTRODUCTION

As a by-product of networked digital computing, large and diverse digital samples of speech are becoming easier to collect and to manage and increasingly large and diverse samples of "found data" are available with almost no work at all. As a result speech data collection is no longer the limiting factor that it once was; rather, speech scientists are now often limited by the volume of human labor involved in making acoustic and articulatory measurements, even using modern interactive computer programs. It's easy to find or create speech datasets in which hundreds of thousands of measurements are in principle available for modeling, but even if each measurement takes only 20 seconds on average, 100,000 measurements will still take 185 grueling three-hour sessions.

Luckily, advances in speech technology and machine learning now generally make it possible to automate the measurement process. For example, a Bayesian approach can reliably automate the usual semi-automatic approach to formant measurement (Evanini *et al.*, 2009). Optional segment deletion or segment substitution can often be automatically detected using HMM matching techniques, as in the work of Fox (2006) on Spanish syllable-final /s/, or Yuan and Liberman (2011a) on English "g-dropping". And similar pattern-matching techniques can often provide non-traditional quantitative proxies for phonetic dimensions such as /l/-velarization (Yuan and Liberman, 2009) or vowel nasalization (Yuan and Liberman, 2011b).

One very common class of phonetic measurements involves time differences between nearby phonetic "events", these events being associated with things like vocal tract closures and releases, the start and end of voicing or frication, and so on. Such events generally correspond nicely to acoustic "edges": extrema or zero-crossings in (derivatives of) time-functions easily derived from speech signals. However, the correspondence depends crucially on choosing the right scale: if the signals are examined on too fine a scale, they will produce many false alarms, while looking on too coarse a scale will lead to events being missed.

An obvious solution is to do the analysis at many scales in parallel, with the idea that the truth will be found somewhere in the resulting "scale space". This leaves us with the problem of integrating information across scales, which remains an area of active research in image processing (e.g. Arbelaez *et al.* (2011). In speech processing, however, the use of scale-space techniques seems largely to have been abandoned.

We have found that in the case of speech-event detection, a standard max-margin classifier generally does a good job of integrating across scales. As a result, a simple scale-space expansion of a few relevant acoustic features, followed by a standard sort of classification step, often offers a simple and reliable way to detect events and to measure inter-event time differences in speech. In this paper, we present a case study: the detection of stop bursts and voice onsets, and the resulting measurement of "voice onset time" (VOT).

After showing that our technique works on three previously-collected datasets where previously-created human VOT measurements exist, we'll discuss the generalization of this approach to other inter-event time measurements, and suggest a set of relevant applications.

VOT MEASUREMENT

Architecture

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 between the burst onset and the voicing onset. Intuitively, it should be possible to measure VOT automatically using classifiers trained to discriminate frames immediately surrounding the

relevant acoustic events from more distant frames; indeed, both the stop burst and the point of voicing onset should be reflected as large positive peaks in the decision functions of these classifiers.

As is the case with edges in a gray-scale image, acoustic events such as a stop burst or voicing onset are highly variable in presentation and particularly in width. Stop bursts present as a brief instance of broad-band energy followed by two periods of frication noise – an initial period generated at the expanding constriction and a terminal period consisting of aspiration generated at the glottis – both of which may vary in duration as a function of stop, following segment, and speaker (Klatt, 1975). As such they are present at a range of intrinisic scales, suggesting no detector operating on a single scale representation can be optimal. Consequently, we adopt a multi-scale representation as the basis for our detection algorithm.

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 (energies in different bands, spectral entropy, spectral centroid, etc.) is extracted every millisecond, yielding a time-series of feature vectors, which, along with its first and second differences, is then projected into scale space via convolution with a series of gaussians. Following creation of this multiscale representation, at each frame we evaluate the decision function of a max-margin classifier 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 $t_v - t_b$.

Features and scale space representation

Five acoustic features (along with their first and second differences) were extracted every ms from the short-time power spectrum computed over a 5 ms gaussian window:

1. $\Delta\log\mathbf{E}(\mathbf{t}) = \log E(t) - \min_{t'} \log E(t')$
2. $\Delta\log\mathbf{E_l}(\mathbf{t}) = \log E_l(t) - \min_{t'} \log E_l(t')$
3. $\Delta\log\mathbf{E_h}(\mathbf{t}) = \log E_h(t) - \min_{t'} \log E_h(t')$
4. $\mathbf{H}(\mathbf{t}) = -\int p(f,t)\log_2 p(f,t)\mathrm{d}f$
5. $\mathbf{C}(\mathbf{t}) = \int f\,p(f,t)\mathrm{d}f$

where $p(f,t)$ is the short-time spectrum of the signal at frequency f and time t, normalized as a density.

The first three features – E, E_l, and E_h – correspond to energy below 8000 Hz, energy below 500 Hz, and energy above 3000 Hz, all normalized relative to the local floor. The fourth feature, $H(t)$ is the spectral entropy (computed as the Shannon entropy of the power spectrum normalized as a density) and measures flatness of the power spectrum. The fifth feature, $C(t)$ is just the spectral centroid, an indication of the center of mass of the power spectrum.

Let f be a feature and σ a scale parameter. Then, the value of f viewed at scale σ at time t is given by

$$L_f(t;\sigma^2) = \int f(t-t')g(t';\sigma^2)\mathrm{d}t' \tag{1}$$

where g is a one-dimensional gaussian with zero mean and standard deviation σ ms. For each of $f \in \{\Delta\log E, \Delta\log E_l, \Delta\log E_h, H, C\}$ and $\sigma \in \{0\text{ms}, 0.5\text{ms}, \ldots, 10\text{ms}\}$, we compute $L_f(t)$, $L_{f'}(t)$, and $L_{f''}(t)$ yielding a multiscale representation for input to the burst and voicing onset detectors.

Burst detector

Of the features described above, we retain the following for burst onset detection, yielding for each frame a 147-dimensional feature vector:

1. $L_{\Delta logE}(t;\cdot)$, $L_{\Delta logE'}(t;\cdot)$, and $L_{\Delta logE''}(t;\cdot)$
2. $L_{\Delta logE_h}(t;\cdot)$, $L_{\Delta logE'_h}(t;\cdot)$, and $L_{\Delta logE''_h}(t;\cdot)$
3. $L_H(t;\cdot)$

which, following (Rahimi and Recht, 2007), is then mapped to an 800-dimensional randomized feature space approximating an RBF kernel. This 800-dimensional representation forms the input to a max-margin classifier trained by Stochastic Gradient Descent (Bottou and Bousquet, 2008) on 1,774 voiceless stops randomly selected from the TIMIT training set (with γ set by grid-search using 5-fold cross validation). Labels for training were constructed by retaining the first two frames following the marked burst location as positive examples and all frames from 20 ms prior to the stop onset to 10 ms prior to the burst and from 10 ms post-burst to 20 ms post stop offset as negative examples.

Voicing onset detector

Training of the voicing onset detector proceeded similarly to that of the burst detector using the same 1,774 randomly selected voiceless stops. For each training instance the following features were retained, yielding a 189-dimensional vector

1. $L_{\Delta logE}(t;\cdot)$, $L_{\Delta logE'}(t;\cdot)$, and $L_{\Delta logE''}(t;\cdot)$
2. $L_{\Delta logE_l}(t;\cdot)$, $L_{\Delta logE'_l}(t;\cdot)$, and $L_{\Delta logE''_l}(t;\cdot)$
3. $L_C(t;\cdot)$, $L_{C'}(t;\cdot)$, and $L_{C''}(t;\cdot)$

which was then projected into an 800-dimensional randomized feature space approximating an RBF kernel (again, γ set by grid-search using 5-fold cross-validation). Labels for training were constructed by retaining the first two frames following the marked voicing onset as positive instances and all frames from 20 ms prior to the stop onset to 5 ms prior to voicing onset and 5 ms to 50 ms following the voicing onset as negative instances.

Performance

We evaluate the algorithm's performance with comparison to human measurements of VOT for three test sets:

TIMIT: We include all stops present in the standard 168 speaker TIMIT test set, whether word-initial or medial, resulting in 5,459 stops (3,158 voiceless and 2,301 voiced).

Lab Speech (LAB) This is a corpus of speakers reading sentence lists under controlled lab conditions (originally collected by Neal Fox and Sheila Blumstein for another study). Each sentence ends in a word containing word-initial /p/ or /b/, which served as the targets of VOT measurement. Data comes from 6 speakers whose VOTs were manually measured by the first author of the present paper, coming to 2,264 stops.

BU Radio Speech (BU) The third dataset comes from a study by Cole *et al.* (2007), who examined word-initial VOT in the BU Radio Speech corpus (Ostendorf *et al.*, 1996). Data comes from 4 speakers in the "lab news" portion analyzed, where professional radio news announcers read news stories in a laboratory setting. The ground-truth VOT measurements are those provided by Cole *et al.* (2007) and comprise 931 stops.

TIMIT

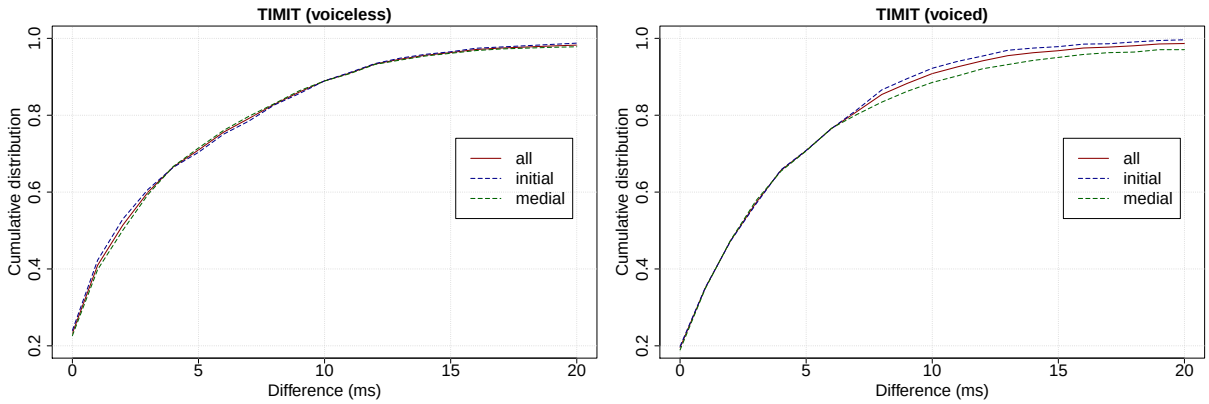

FIGURE 1: Cumulative distributions of absolute differences between human and system VOT measurements on voiceless (left) and voiced (right) stops in the full TIMIT test set. Mean absolute differences: all stops, 4.99 ms; voiceless, 4.76 ms; voiced, 5.3 ms

Overall, performance on TIMIT is generally good with ≈ 65% of measurements within 5 ms and > 85% within 10 ms of human. Figure 1 depicts the cumulative distribution function of the absolute differences between system and human VOT measurements for voiceless (left panel) and voiced (right panel) stops. Overall performance for voiceless and voiced stops is nearly identical. Similarly, there is no drastic difference between performance on initial and medial stops, though in the voiced case initial stops appear to be slightly easier than medial ones.

LAB

Despite being trained on TIMIT, the algorithm performs even better on the stops in LAB. As can be seen in the left panel of Figure 2, > 80% of the system measurements are within 5 ms and nearly 100% within 10 ms of the human-marked values. Performance is rather better for /b/ than for /b/, with nearly 90% of the /b/ measurements withing 5 ms of human as compared to 80% for /p/.

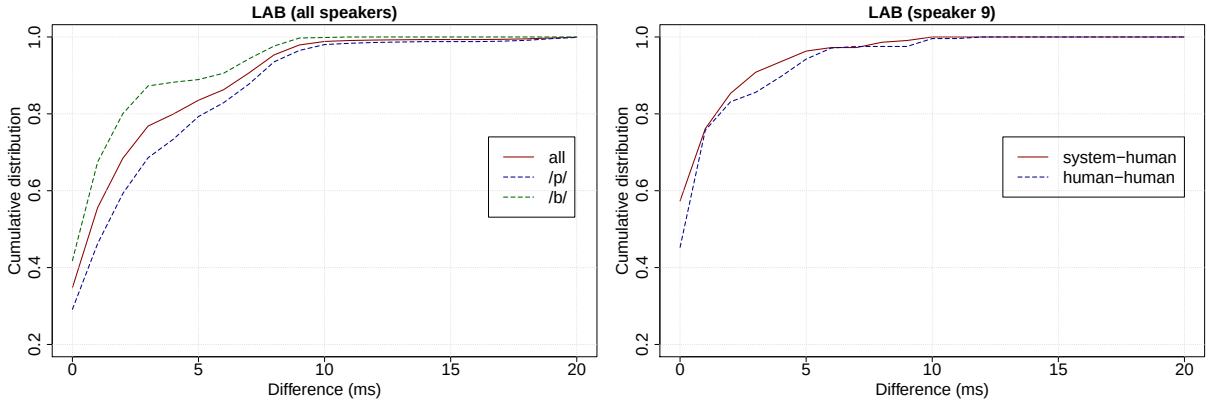

FIGURE 2: *Left*: Cumulative distribution of absolute differences between human and system VOT measurements for all instances of /p/ and /b/ in LAB. Mean absolute differences: all stops, 2.84 ms; voiceless, 3.44 ms; voiced, 2.08 ms. *Right*: Cumulative distributions of human/system and human/human differences for speaker 9. Mean absolute differences: system-human, 1.49 ms; human-human, 1.50 ms.

The error distribution in Figure 2 is certainly promising and suggests that the system's VOT measurements could replace those of human annotators. As a test this idea two of the authors independently annotated a subpart of LAB (corresponding to all 229 stops of speaker 9) and

their measurements were compared to that of system. The distributions of these system-human and human-human differences for this speaker are depicted in the right-panel of Figure 2. Strikingly, the system-human differences actually tend to be lower than the human-human differences.

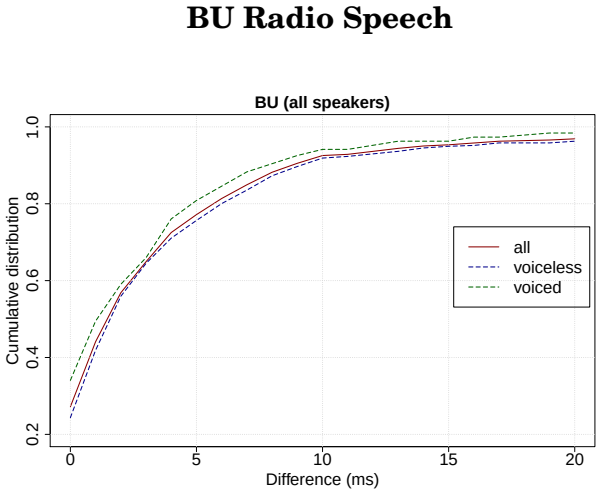

FIGURE 3: Cumulative distribution of absolute differences between human and system VOT measurements on voiceless and voiced stops in BU Radio Speech. Mean absolute differences: all stops, 4.57 ms; voiceless, 4.94 ms; voiced, 3.68 ms

We see similarly good generalization to a novel domain when we apply the system to the BU data, as seen in Figure 3. For BU > 70% of the differences are within 5 ms and > 90% within 10 ms compared to the measurements in Cole *et al.* (2007). As was the case in LAB, performance is somewhat better for voiced than for voiceless stops, particularly below 10 ms.

In Figure 4, we see the plots of mean VOT by voicing and accent and by place of articulation (POA) and accent, comparing the system's results to the original (human) measurements from Cole et al. 2007. The patterns are qualitatively identical, and quantitatively quite similar – the

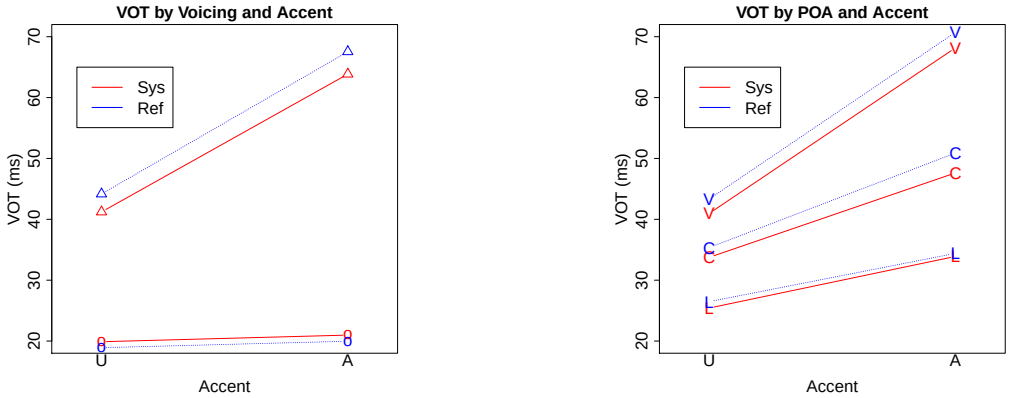

FIGURE 4: Interaction graphs illustrating the effect of Accent ((A)ccented/(U)naccented) on VOT. *Left*: ○, Voiced; Δ, Voiceless. *Right*: V, velar; C, coronal; L, labial.

human VOT measurements are systematically a few ms longer than the system's measurements, apparently due to the fact that Cole et al.'s definition of "voice onset" ("the zero crossing nearest to the onset of the second formant of the following vowel") tended to place this point slightly later than the automatic system did.

Recall

From the edge-detection literature, we also know that while boundary placement accuracy tends to increase with increasing scale, recall also decreases. Consequently, we also consider the recall of the system, defined as the percentage of stops with a human marked VOT where the system attempts a measurement. Table 1 gives these rates for all three test sets. For all test sets overall recall exceeds 80% with 90% exceeded in both TIMIT and LAB.

	voiceless	voiced	all
TIMIT	94.4	86.0	90.9
LAB	93.7	96.1	94.7
BU	84.4	80.3	83.2

TABLE 1: Recall (% stops detected) across test sets.

Comparison to previous systems

Automatic VOT measurement has been treated previously (Niyogi and Sondhi, 2002; Das and Hansen, 2004; Stouten and Van hamme, 2009; Sonderegger and Keshet, 2010), with the work closest to the current approach being that of Sonderegger and Keshet (2010) (hereafter, S&K). S&K report performance for word-initial voiceless stops in the core and full TIMIT test sets (excluding calibration sentences) using two metrics: root mean square error (RMS) of the burst placement and percentage of cases where the manual and automatic measurements differ by at least 10% (≥10%). Table 2 gives these metrics for our multi-scale algorithm on the same sets and compares them to those achieved by S&K. Our algorithm does better than S&K for proportion of errors ≥ 10% on both test sets and for RMS error on the full TIMIT test set, with S&K achieving somewhat better RMS error for the core test set. Both the features and the machine-learning algorithms were somewhat different for the two approaches. Our point here is just that our choices are competitive.

	system	mean (ms)	RMS (ms)	≥ 10%
TIMIT	multi-scale	4.67	6.14	31
(all)	S&K	–	8.66	35
TIMIT	multi-scale	4.11	5.87	28
(core)	S&K	–	5.28	34

TABLE 2: Comparison of system performance on core and full TIMIT test using metrics from (Sonderegger and Keshet, 2010). Additionally, we depict mean error.

CONCLUSION

We have argued that a standard max-margin classifier, operating on a scale-space expansion of a set of sensible input features, provides a good general architecture for detecting phonetic events and measuring time intervals between them. We used the example of Voice Onset Time as a case study, and showed that a simple implementation of our architecture works well on three existing datasets for which human VOT measurements are available.

There are many other duration-measurement tasks for which this approach is suited: stop closure durations, the durations of fricatives or nasal murmurs, inter-obstruent vowel durations, and so on. And there are many reasons to want to measure such durations: linguistic questions like the distribution of raddoppiamento fonosintattico in Italian (Borrelli, 2000); psychological questions like the effects of phonological neighborhood density on speech production (Gahl *et al.*,

2012); clinical questions like the effects of parkinsonism on speech timing patterns (Rusz *et al.*, 2011).

Obviously there is a relationship between scale space features and more general trajectory modeling methods. One currently-popular way to incorporate more general trajectory information is to use so-called "deep belief nets" (Hinton *et al.*, 2006), operating on sequences of adjacent frames. We've used this approach for several speech-classification and event-detection problems, with generally positive results. However, the scale-space method described above often works just as well, if supplied with an appropriate set of inputs, and notably lacks the additional training requirements of the DBN features.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0964556. We would also like to express our gratitude to Jennifer Cole, Neal Fox, and Sheila Blumstein for allowing us the use of their data.

References

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. (**2011**). "Contour detection and hierarchical image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 898–916.

Borrelli, D. (**2000**). "Raddoppiamento Sintattico in Italian: a synchronic and diachronic cross-dialectal study", Ph.D. thesis, Cornell University.

Bottou, L. and Bousquet, O. (**2008**). "The tradeoffs of large scale learning", in *Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis, volume 20, 161–168.

Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. (**2007**). "Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from radio news speech", Journal of Phonetics **35**, 180–209.

Das, S. and Hansen, J. H. (**2004**). "Detection of voice onset time for unvoiced stops using Teager Energy Operator for automatic detection of accented English", in *Proceedings of NORSIG 2004*.

Evanini, K., Isard, S., and Liberman, M. (**2009**). "Automatic formant extraction for sociolinguistic analysis of large corpora", in *Proceedings of Tenth Annual Conference of the International Speech Communication Association*.

Fox, M. (**2006**). "Usage-based effects in Latin American Spanish syllable-final /s/ lenition", Ph.D. thesis, University of Pennsylvania.

Gahl, S., Yao, Y., and Johnson, K. (**2012**). "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech", Journal of Memory and Language **66**, 789 – 806.

Hinton, G., Osindero, S., and Teh, Y. (**2006**). "A fast learning algorithm for deep belief nets", Neural Computation **18**, 1527–1554.

Klatt, D. (**1975**). "Voice onset time, frication, and aspiration in word-initial consonant clusters", Journal of Speech Hearing Research **18**, 686–706.

Marr, D. (**1976**). “Early processing of visual information”, Philosophical Transactions of the Royal Society of London. B. **275**, 483–524.

Niyogi, P. and Sondhi, M. (**2002**). “Detecting stop consonants in continuous speech”, Journal of the Acoustical Society of America **111**, 1063–1076.

Ostendorf, M., Price, P. J., and Shattuck-Hufnagel, S. (**1996**). *The Boston University Radio Speech Corpus* (Linguistic Data Consortium, Philadelphia, PA).

Rahimi, A. and Recht, B. (**2007**). “Random features for large-scale kernel machines”, in *Proceedings of NIPS 2007*, 1177–1184.

Rusz, J., Cmejla, R., Ruzickova, H., and Ruzicka, E. (**2011**). “Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson’s Disease”, The Journal of the Acoustical Society of America **129**, 350–367.

Sonderegger, M. and Keshet, J. (**2010**). “Automatic discriminative measurement of voice onset time”, in *Proceedings of INTERSPEECH 2010*, 2242–2245.

Stouten, V. and Van hamme, H. (**2009**). “Automatic voice onset time estimation from reassignment spectra”, Speech Communication 1194–1205.

Yuan, J. and Liberman, M. (**2009**). “Investigating /l/ variation in English through forced alignment”, in *Proceedings of INTERSPEECH 2009*, 2215–2218.

Yuan, J. and Liberman, M. (**2011**a). “Automatic detection of “g-dropping” in American English using forced alignment”, in *Proceedings of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 490–493 (IEEE).

Yuan, J. and Liberman, M. (**2011**b). “Automatic measurement and comparison of vowel nasalization across languages”, in *Proceedings of ICPhS XVII*, 2244–2247.