SCALE-SPACE EXPANSION OF ACOUSTIC FEATURES IMPROVES SPEECH EVENT DETECTION

Neville Ryant, Jiahong Yuan, and Mark Liberman

Linguistic Data Consortium, University of Pennsylvania

ABSTRACT

In a system for detecting and measuring phonetic events (here bursts, voice onsets, and voice-onset times), we show that the addition of features smoothed at multiple scales can improve both recall (the proportion of events correctly identified) and measurement accuracy (the timing of events and the difference between event times, relative to expert human judgments). Multi-scale (or "scale space") features had an especially strong positive effect on robustness across datasets with different materials and recording conditions. Standard machine-learning classifiers were able to integrate information across scales, without any special treatment of the multi-scale features.

Index Terms— voice onset time, scale space, automated phonetic measurement

1. INTRODUCTION

The extrema in a signal and its derivatives often provide useful information about relevant regions and boundaries. Thus the maxima and minima in an appropriately-smoothed amplitude contour of a speech signal correspond approximately to vocalic and consonantal regions; maxima and minima in the derivative generally pick out points of transition from one phonetic segment to another.

This idea has been proposed for edge-detection in image processing for more than three decades [1, 2, 3], and for even longer in speech processing [4, 5]. However, finding just the right regions and edges requires smoothing to just the right scale: when the scale is too fine, there can be many false alarms, and when the scale is too coarse, relevant features will be missed.

One 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. The problem of integrating information across scales to derive segments and edges or contours remains an area of active research in image processing [6], but in speech processing, the use of scale-space techniques seems largely to have been abandoned. In this paper, we describe a case where deriving acoustic features at multiple scales improves performance substantially, and where the integration across scales is accomplished simply by a standard max-margin classifier, without any additional machinery.

We start with an algorithm for high-accuracy automated measurement of voice onset time (VOT) based on features such as energies in different frequency bands, spectral entropy, and spectral centroids, along with their first and second devatives, which serve as input to paired burst and voicing onset detectors. We find that scale-space expansion of these features yields a significant improvement in performance compared to the same input features spanning the same time regions but without smoothing at different scales.

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 Neal Fox and Sheila Blumstein for allowing us the use of their data.

2. VOT MEASUREMENT

2.1. 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 spanning burst onset and 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 induration as a function of stop, following segment, and speaker [7]. 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 be-

tween the recording and its transcript) a series of acoustic features (energies in different bands, spectral entropy, spectral centroid, etc.) is extracted every ms, yielding a timeseries 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 $t_v - t_b$.

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

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

$$L_\phi(t;\sigma^2) = \begin{cases} \int \phi(t-t') g(t';\sigma^2) \mathrm{d}t' & \text{if } \sigma > 0 \\ \phi(t) & \text{if } \sigma = 0 \end{cases} \quad (1)$$

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

2.3. Burst detector

Of the features described in Section 2.2 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 [8], is then mapped to an 800-dimensional randomized feature space approximating a radial basis function (RBF) kernel with $\gamma = 0.000152$[1]. This 800-dimensional representation forms the input to a max-margin classifier trained by Stochastic Gradient Descent [9] 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.

2.4. Voicing onset detector

Training of the voicing onset detector proceeded similarly to that of burst detection 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 with $\gamma = 0.0370$. 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.

3. EXPERIMENTS

We report results for two test sets:

TIMIT We consider all instances of the voiceless stops /p, t, k/ in the standard 168 speaker TIMIT test set (n=3,158).

[1] [8] propose approximating the implicit feature mapping of the RBF kernel using random Fourier features – cosines of random affine projections of the data. With sufficiently many such features it is possible to retain the ability of kernel machines to fit nonlinear decision surfaces while avoiding the high computational costs incurred in calculation of the kernel matrix.

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.

3.1. Single vs multi-scale

Figure 1 depicts the cumulative distribution of differences between automatic and human VOT measurements for our multi-scale system and for a series of single-scale systems (trained as described in Section 2.1, but with smoothing restricted to a single scale σ) on both test sets. As is the case in the edge-detection literature, smoothing of any kind is beneficial with $\sigma = 0$ by far the worst performer for both test sets. For TIMIT, accuracy of the VOT measurements increases with increasing σ up to a point then plateaus, while for LAB the situation is somewhat more complicated; though the overall percentage of errors $<$10 ms decreases with increasing σ, this pattern does not hold for errors $<$5 ms. Regardless, no single-scale system ever outperforms the multi-scale system. Moreover, while for large σ the error-distributions of the single-scale systems closely approximate that of the multi-scale system on TIMIT, this performance does not generalize to LAB. This suggests another advantage of using multi-scale features: increased generalizability to novel domains.

From the edge-detection literature, we also know that while accuracy tends to increase with increasing scale, recall decreases. Consequently, we also report recall – the percentage of stops with a human marked VOT where the system attempts a measurement. From Table 1 it is evident that for both TIMIT and LAB, recall for single-scale systems is highest for lower σ and quite poor for $\sigma = 8$ and $\sigma = 9$. Moreover, no single-scale system comes close to the recall of the multi-scale system on either test set.

		σ (ms)					
	multi	0	1	2	5	8	9
TIMIT	91.5	85.1	80.2	86.4	86.2	64.4	51.8
LAB	93.1	22.0	72.1	88.2	74.8	10.6	2.5

Table 1. % recall compared for multi-scale and single scale systems on TIMIT and LAB.

3.2. Sensitivity to number of kernels

Having established the general utility of multi-scale representation for VOT measurement, the question naturally arises as to how sensitive this approach is to the exact number of gaussian kernels used in creating the multi-scale representation. Figure 2 depicts cumulative error distributions for a series of multi-scale systems whose maximum scale σ varies from 1 to 20. While the effect in going from $\sigma = 1$ to $\sigma = 5$ is marked, adding additional scales above this has very little effect on the error distribution. Nor does it markedly impact recall, which, as is seen in Table 2, quickly climbs to 90%, then levels off.

	$\sigma_{\max}$ (ms)				
	1	5	10	15	20
TIMIT	67.7	88.3	91.5	91.2	92.5
LAB	49.8	92.4	90.0	88.6	88.9

Table 2. % recall compared for different maximum scale sizes on TIMIT and LAB.

3.3. Comparison to previous systems

Automatic VOT measurement has been treated previously [10, 11, 12, 13], but the work closest to the current approach is that of Sonderegger & Keshet (S&K; [13]). 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% ($\geq$10%). Table 3 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 a proportion of errors $\geq 10\%$ 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, and that versions of our algorithm with multi-scale features outperform versions with single-scale features[2].

	system	mean (ms)	RMS (ms)	$\geq 10\%$
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
LAB	multi-scale	2.80	1.82	29

Table 3. Comparison of system performance on core and full TIMIT test sets and LAB using metrics from [13]. Additionally, we depict mean error.

3.4. Comparison to human

The error distribution in Figure 1 is certainly promising and suggests that the system's VOT measurements could replace

[2]Indeed S&K do include some scale-space like features (consisting of differences in means of sequences of frames). We suggest that performance would be worse without these derived features and that performance would improve were scale-space features more consistently and systematically used.

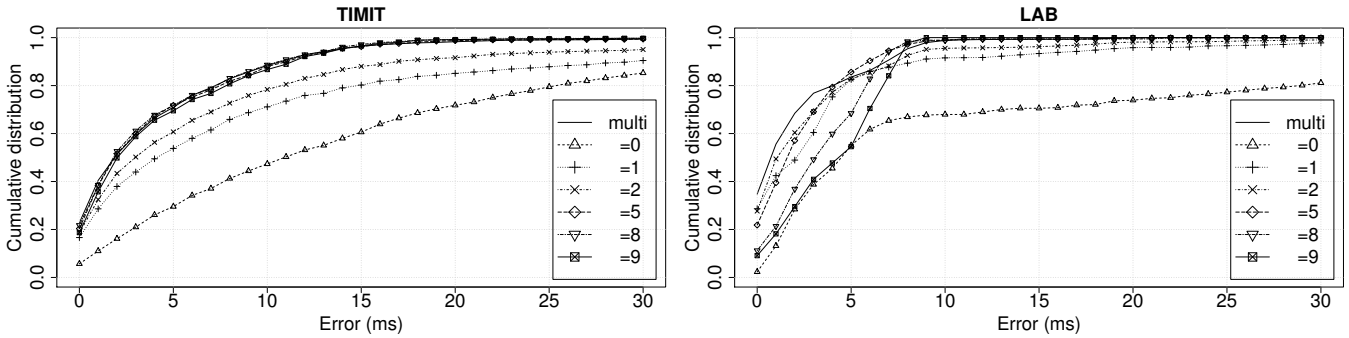

Fig. 1. Cumulative distributions of absolute differences between human and system VOT measurements on TIMIT (left) and LAB (right) test sets for multi-scale and single-scale systems.

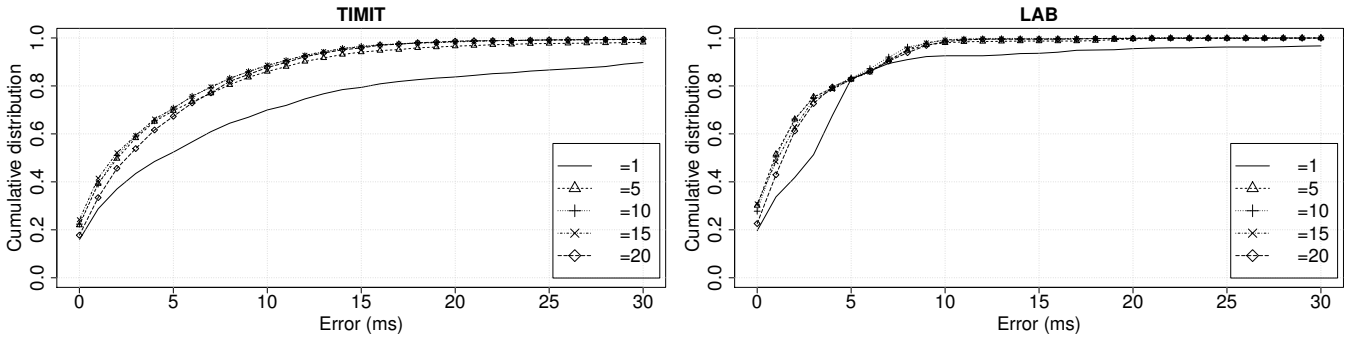

Fig. 2. Effect of maximum scale size on measurement accuracy for TIMIT and LAB.

those of human annotators. To test this posit two of the authors independently annotated a subpart of LAB (all 229 stops from speaker 9) and their measurements were compared to that of the multi-scale (Figure 3). Strikingly, the distributions of the human-human differences and the mean of the human-system differences are essentially identical.

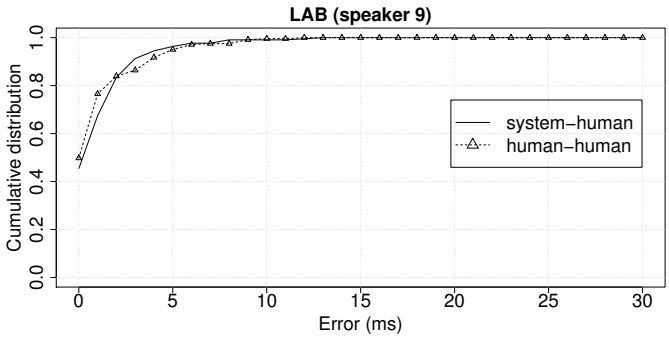

Fig. 3. Cumulative distributions of human/system and human/human differences for speaker 9 in LAB (mean differences: system-human, 1.77 ms; human-human, 1.79 ms)

4. CONCLUSION

Ever since the work of David Marr in the 1970s, researchers have been exploring the idea that animals and computer algorithms might locate edges in an image or in an acoustic signal by looking for zero-crossings in the second derivative of some kind of intensity signal, and that such methods could be made more robust by linear smoothing of the input signals at multiple scales. While such "scale space" approaches remain relevant in image processing and in studies of computer vision, they have largely dropped out of sight in acoustic analysis. Even research aimed at detecting phonetic "landmarks" generally does not use multi-scale inputs in any systematic way.

We believe that this omission is a mistake, and that detectors and classifiers for acoustic-phonetic events will generally benefit from "scale spacification" of their input features. We also believe that the problem of integrating information across scales can generally be handled by standard machine-learning techniques, without any special attention to the scale-space nature of some of the inputs.

We have shown that multi-scale features, fed into a standard machine-learning algorithm, provide significant benefits in the specific case of an algorithm for detecting bursts and voicing onsets, locating these phonetic events in time, and thereby measuring voice onset time.

5. REFERENCES

[1] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London. B.*, vol. 207, pp. 187–217, 1980.

[2] R. Haralick, "Digital step edges from zero crossing of second directional derivatives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 58–68, 1984.

[3] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679–698, 1986.

[4] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of the Acoustical Society of America*, vol. 58, pp. 880–883, 1975.

[5] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *Proceedings of ICASSP.*, 1984, pp. 150–153.

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

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

[8] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Proceedings of NIPS*, 2007.

[9] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20, pp. 161–168. 2008.

[10] P. Niyogi and M. Sondhi, "Detecting stop consonants in continuous speech," *Journal of the Acoustical Society of America*, vol. 111, pp. 1063–1076, 2002.

[11] S. Das and J. H. Hansen, "Detection of voice onset time for unvoiced stops using teager energy operator for automatic detection of accented english," in *Proceedings of NORSIG 2004*, 2004.

[12] V. Stouten and H. Van hamme, "Automatic voice onset time estimation from reassignment spectra," *Speech Communication*, pp. 1194–1205, 2009.

[13] M. Sonderegger and J. Keshet, "Automatic discriminative measurement of voice onset time," in *Proceedings of Interspeech 2010*, 2010, pp. 2242–2245.

[14] D. Marr, "Early processing of visual information," *Philosophical Transactions of the Royal Society of London. B.*, vol. 275, pp. 483–524, 1976.