# Speech Recognition in Scale Space

**Richard F. Lyon**
Schlumberger Palo Alto Research
3340 Hillview Avenue
Palo Alto, CA 94304

**Abstract**

Scale-space filtering, proposed by Witkin (ICASSP 84) for describing natural structure in one-dimensional signals, has been extended for application to segmentation and description of vector-valued functions of time, such as speech spectrograms. By analyzing the rate of change of a vector trajectory at many different scales of time-smoothing, a tree of natural segments can be constructed. At various levels in the tree (i.e., at various scales), these segments are found to agree well with the kind of linguistically and perceptually important segments that spectrogram readers use to describe sound patterns of speech. Scale-space segmentations of cochleagrams (spectrograms based on a computational model of the peripheral auditory system) have been experimentally applied to word recognition. Recognition using fixed-scale segmentations with finite-state word models and a Viterbi search has led to speaker-independent digit recognition accuracies of greater than 97%, about the same as in tests with non-segmented cochleagrams. More complex recognition algorithms that use the segmentation tree are being developed, and scale-space experiments with connected digits and sentences are also underway.

## 1 Introduction

Before a speech signal can be recognized, it must be analyzed into a form suitable to the recognition algorithms. Most current systems use a simple sequence of frames (spectrum vectors, feature vectors, or vector quantization indices) as the representation to do pattern matching on, using a dynamic programming algorithm or similar matching approach. An approach that was popular ten years ago was to first do "segmentation and labelling", so that the recognition system could work on a sparser, more symbolic representation, applying more general kinds of knowledge than templates. This approach has been plagued by the difficulty of getting the right segments—any segmentation algorithm will occasionally either miss segments or split segments or both. The "correct" segmentation depends on the interpretation made by higher-level models, and is not always apparent from the input signal. The techniques presented in this paper are an attempt to find a lattice of alternative natural segments in a sequence of feature vectors, so that higher levels of knowledge can be successfully applied to the recognition problem.

Witkin's "Scale-Space Filtering" technique [1] provides a structured description of a signal at multiple scales. The structure chosen is a tree of segments, with segment boundaries representing supposed discontinuities in the signal. Since continuity of sampled signals is not defined, boundaries are placed at local maxima in the rate of change of the signal. But since rates of change are ill-defined, a space of different smoothing scales is explored, resulting in boundaries that exist over a limited range of scales, thereby inducing a tree of segments dividing into subsegments.

The segmentation tree appears to be an important innovation that will greatly help the acoustic-phonetic approach to speech recognition, and will also bring that approach into closer contact with statistical signal processing and pattern matching approaches. The segments provide natural units on which feature extractors can operate (i.e., they delimit reasonable intervals over which to measure features that characterize segments, rather than isolated spectral slices). The tree gets around the classic problem with segmentation by putting both too-coarse and too-fine segments together in a unified structure at different levels, letting the higher knowledge sources find the segments that they need to match a model.

We have implemented an interactive segmentation examiner in the ISP signal processing environment, so that the user can explore the segmentation tree and see when enough information is apparent to read words from the picture. For this purpose, each segment is described by the average vector within that segment of the original, so that a piecewise-constant approximation results; piecewise-linear and piecewise-parabolic approximations may better encode relevant transition structure, and are also being experimented with. Optionally, each segment's average spectrum vector can be vector quantized, resulting in a truly compact symbolic representation of the utterance that is reasonably adequate for recognition.

Finite-state word models similar to those used by Bush and Kopec [2] are used as the initial higher-level model of pronunciations. Recognition experiments with very simple matching algorithms give encouraging results of around 98% correct on speaker-independent isolated digit tests.



CH2396-0/87/0000-1265 $1.00  

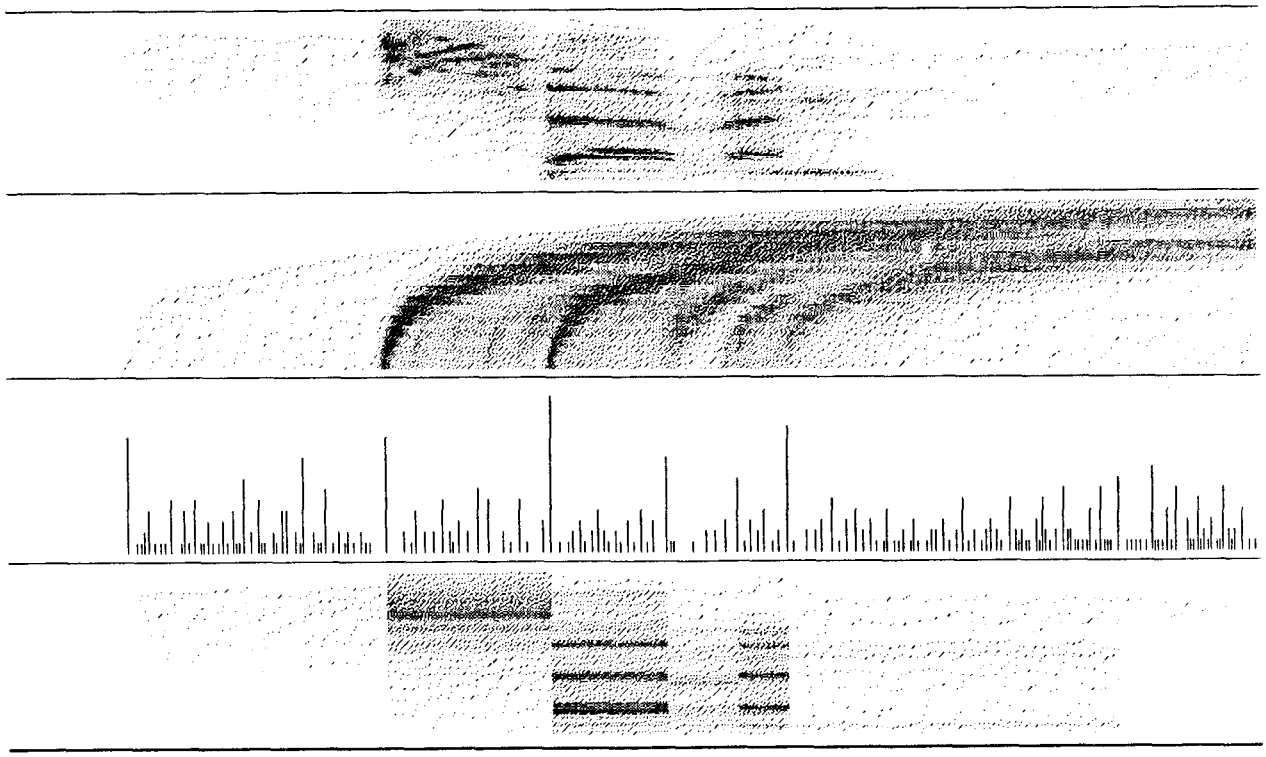

Figure 1: Scale-space segmentation of the isolated digit "seven". The original cochleagram (top) is smoothed and differenced to form its rates of change at various scales (second from top, plotting high rates of change as dark, from finest scale at bottom to coarsest scale at top); curve shapes reflect the causal smoothing filters. Local maxima in the rate of change at any scale are tracked back to the finest scale, yielding a set of boundaries that persist to various different scales (third from top). A reconstructed cochleagram (bottom) is made by averaging the spectra within segments (delineated by boundaries that persist to at least scale 6 in this example); notice that the final "N" sound is not resolved at this scale.

## 2 Scale-Space Overview

Witkin used zero-crossing of the second derivative (i.e., minima and maxima of the first derivative) of scalar-valued signals to segment them. The technique is generalized to vector-valued signals by using peaks of the magnitude of the vector derivative. A cascade of vector time-domain smoothers (first-order recursive filters) is used to get vector derivatives at a sequence of scales. A peak in a rate of change at any scale is tracked back through corresponding peaks at finer scales to find the underlying boundary. All computations have been implemented within the ISP/SRL signal processing environment [3] on the Symbolics lisp machine.

Figure 1 shows a typical isolated digit cochleagram along with a 2D plot of its rate of change vs. scale, a plot of the boundaries that correspond to the segmentation tree induced thereby, and a reconstructed segmented cochleagram at a particular scale (scale 6, with smoothing time constant of about 24 msec). At this scale, most segments correspond quite well to the usual phonemic interpretation.

## 3 Vector Quantized Cochleagrams for Non-Parametric Speech Recognition

For recognition of isolated digits from cochleagrams, we have used techniques based on vector quantization of cochlear spectrum slices; results are reported elsewhere in this proceedings [4]. To incorporate scale-space segmentations, we have so far experimented with only very simple techniques—in particular, we represent the spectral information in each segment only by the vector quantized average spectrum over the segment, using the same codebook that was used in the non-segmented experiments.

For recognition with finite-state models, we use a Viterbi algorithm constrained to consider state transitions only at segment boundaries. The same algorithm works with either segmented or non-segmented data, by accepting run-length coded vector quantization index sequences as its input; with scale-space segmentation, each segment in a particular segmentation is simply converted to a run of a constant index.

The cost metric used by the Viterbi algorithm in finding a best model-based segmentation is a minus log likelihood

of occurrence of a codeword given the state. The state tables were initially trained using segmentations found by a previous-generation LPC-based recognizer; tables have been retrained and models modified to improve performance. In comparing the fits of the various word models, measures other than total cost (probability) were used to advantage. In particular, it was found that an average of the within-state time-average costs in each model gave much better results (by putting more relative weight on shorter segments). It was also found that adding in a minus log likelihood metric of a coarsely quantized state duration with a weight of about one-third gave further improvement; duration histograms were also bootstrap trained from the Viterbi fits. These measures are not exactly optimized by the Viterbi algorithm, and this is not a very good way to use durational information, but the results are still not bad.

In order to get good recognition results, we found it necessary to smooth the codeword distributions for each state, thereby avoiding pathological performance due to encountering codewords that were not seen in training. Very little smoothing was needed—we used a gaussian blur for each codeword with a standard deviation of only one-third the distance to its nearest neighbor codeword (using euclidean distances between codewords in the pattern space of cochlear spectra).

## 4 Experiments, Results, and Discussion

Initial experiments have used the weak technique of throwing out the segment tree and considering only the particular segmentations obtained by fixing the scale at different levels. The family of segmentations with scales from 6 to 24 msec (smoothing time constant before derivative) yield segment sequences that can be recognized about as well as the unsegmented originals (less than 1% error rate in multi-speaker tests, 2% to 3% in speaker-independent tests). Segment lengths average about triple the smoothing time constant, so are typically 18 to 72 msec at these scales; many digits are recognized correctly at higher scales, but error rates increase rapidly as important boundaries disappear.

Many experimental variants and retrainings on the basic weak technique have been tried, in an attempt to find conditions under which the coarser segmentations would yield accuracies significantly better than the finer segmentations. In the best results with testing on the same data as training (clearly not a fair test), scale 3 (8.5 msec) made fewer errors (a single error in 1232 tokens in the best case) than scales 2 (6 msec) and higher scales. In most other cases, the errors increased monotonically with scale.

The results of the weak technique can be combined by letting different scales vote. If the criterion is best two-out-of-three at scales 2 through 4, then no errors are made on the training set of 1232 tokens. With durational information omitted, our best results are 3 errors in 1232, either at scale 3 or by various multi-scale voting schemes. The system performance has been improved by bootstrap retraining the state statistics from the system's own best match segmentations, so it may be unfairly over-trained on the training set.

Fair experiments on additional data are in progress. Using half of the TI training talkers for training and half for testing, the best results so far are 12 errors in 616 (1.94% error) for training on the second half of the talkers (alphabetically) and testing on the first half, or 18 errors (2.92% error) for training on the first half and testing on the second half; on this latter condition but without scale-space segmentation, only 10 errors (1.62%) were made [4].

A problem with all of these tests is that the codebook was trained on all talkers, and without segmentation. In experiments trained with half the talkers, several of the codewords are never seen. Better results would be expected by using a codebook trained only on segment average spectra from the training talkers.

Better results are also expected from using both first and second repetitions of digits from all the talkers—the scale-space segmentations of the second repetitions are not done yet, but experiments with unsegmented cochleagrams gave much lower overall error rate when the second repetitions were included in the training and testing. The second repetitions apparently have more variability and perhaps informality than the first, since in test with both reps, the second reps cause twice as many errors as the first reps (totalling only 1.06% error in speaker-independent mode).

Expanding the tests to train on the entire TI training corpus and test on the testing corpus may actually further improve the accuracy, since twice as many training talkers will be represented.

## 5 Better Recognition Approaches

To fully take advantage of scale-space segmentation, better techniques are obviously required. The use of vector quantization, which is convenient in allowing simple non-parametric statistical approaches, may not be an appropriate way to take advantage of the extra structure provide by scale-space techniques. Explicit segment characterization by features may be a better alternative, and is certainly facilitated by the segment tree.

Constraint-based grammars to describe possible ways of putting segments together into words, based on compatible features [6], could be applied to parse the segment-tree. Experiments along these lines are just beginning, in collaboration with researchers at Xerox and MIT.

## 6 Training Issues

All of the recognition experiments so far have relied on the best model fits, determined by a Viterbi search, of a previously trained recognizer to provide labelled training data. Retraining a recognizer based on its results has provided a performance improvement in most, but not all,

cases. Many errors occur with utterances that do not provide very good matches to any of the models, and retraining on these simply reinforces the problem in some cases. In a few cases we have hand-modified data structures to provide better fits within the training data, with limited success. A good corpus of hand-marked training data could provide further improvement, or perhaps a Baum-Welch algorithm or similar technique could be used to optimize the models.

## 7 Scale-Space Filtering Details

The scale-space filtering implementation used in this work includes a novel filter structure for approximating a sequence of derivatives at increasing scales. Each stage of smoothing is a single-pole lowpass filter with unity gain at DC, yielding a delay equal to the smoothing time constant. Each vector lowpass filter simply filters the components of the vector independently (the 92 frequency channels of the cochleagram). These stages, with time constants increasing by a factor of $\sqrt{2}$ per stage, are cascaded to yield a succession of increasingly smooth versions of the original signal. Each derivative is then taken as a simple difference across a stage, as shown is Figure 2; i.e., the filter stage delay is used as the time difference, yielding a transfer function with a zero at DC and a bunch of poles on the real axis. Derivatives at larger scales are effectively measured across larger time differences, so that the magnitudes of the derivative signals are not decreasing in inverse proportion to the scale, as they would be for a fixed time difference.

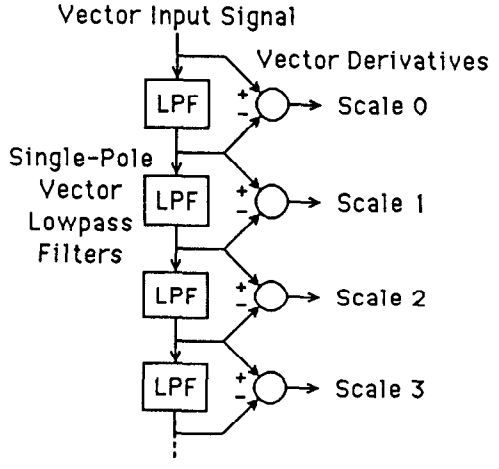


Figure 2: Multi-scale derivative filter structure.

This implementation has proved to be better in terms of roundoff noise amplification than other approaches that were tried, such that spurious local maxima to not occur. The cascade structure is also similar to models we have used for the cochlear wave propagation structure and for neural correlator delay lines. For any of these applications, the cascade structure provides the desirable feature that the characteristic time scales of the outputs are monotonically increasing, even if the inaccurate component values are not, as in analog implementations or neural wetware.

## 8 Conclusions

It had been hoped that the segmentations found by scale-space methods at fairly coarse scales would lead to improved recognition by eliminating the typical errors make by the Viterbi search—namely, allowing some model states to fit extremely short stretches of the unknown input. In fact, the coarse segments do help some in this direction, but the same problem was essentially solved without segmentation by weighting all within-state time-average costs equally, rather than in proportion to duration as is standard. On the other hand, the extremely data-compressed segmented representations (about 200 bits per second for scale 6 to 800 bits per second for scale 2) led to recognition accuracies almost as good as the unsegmented data, indicating that information is being fairly well preserved by these methods.

The power of the scale-space techniques should be further enhanced by describing trajectories within segments as low-order polynomials, rather than as constants (zero-order). Then features such as formant tracks and time patterns can be extracted from these smooth trajectories, probably much more easily than from the original data.

## 9 References

[1] Andrew P. Witkin, "Scale-Space Filtering: A New Approach to Multi-Scale Description," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, San Diego, March 1984.

[2] Marcia Bush and Gary Kopec, "Network-Based Connected Digit Recognition," to appear in *IEEE Trans. ASSP* **35**, 1987.

[3] Gary Kopec, "The Integrated Signal Processing System ISP," *IEEE Trans. ASSP* **32**:4, Aug. 1984.

[4] Eric Loeb and Richard Lyon, "Experiments in Isolated Digit Recognition with a Cochlear Model," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, April 1987 (this proceedings).

[5] Richard F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Paris, May 1982.

[6] Daniel Huttenlocher and Meg Withgott, "On Acoustic versus Abstract Units of Representation," *Proc. Montreal Symposium on Speech Recognition*, Montreal, July 1986.

**Research supported by DARPA contract #N00039-85-C-0585.**