Dynamic Mixtures of Splines: a Model for Saliency Grouping in the Time Frequency Plane

Stéphane H. Maes* Trevor Hastie †

March 11, 1997

Abstract

We describe a new approach for focusing images obtained by time-frequency analysis of speech signals. Our summary consists of a set of curves which include the formants. Due to the oscillatory nature of the two dimensional data, all the classical approaches to this problem have shortcomings. Our method is semi-automatic, requiring as input only the maximum number of curves. The summary curves are designed to correspond with human perception: they are not allowed to cross, they can die with time, new curves can be born, or the same curve can die and then be reborn.

Our approach is based on two modeling assumptions a) we view each normalized image as a probability density function, and b) we approximate these densities, conditional on time, as mixtures of Gaussians. The curves then represent the means of the Gaussians as a function of time. Within this framework, we fit the model by maximum cross-entropy, and enforce the constraints by adapting the standard fitting algorithm.

Keywords: curve fitting, EM algorithm, Kullback-Leibler distance, saliency extraction, mixtures of Gaussians.

* *IBM*, T. J. Watson Research Center, Human Language Technologies, Acoustic Processing Department, Room 23-148B, P.O. Box 218, Yorktown Heights, NY 10598, USA. Phone: +1-(914)-945-2908; Fax: +1-(914)-945-4490; e-mail: smaes@watson.ibm.com.

† Statistics Department, Sequoia Hall, Stanford University, CA94305. email: trevor@stat.stanford.edu

1 Introduction

For automatic speech and speaker recognition, a new representation was recently introduced, known as the *Synchrosqueezed Plane Representation* [11, 3]. This is a nonlinear transformation of the wavelet transform.

Wavelet transforms (WT) of speech provide a time-scale representations of the speech signal [16]. Typically low frequency events will appear at a high scale with poor temporal localization, while high frequency events will appear at a low scale but will be well localized in time. This representation is obtained by filtering the speech signal with a filter-bank of overlapping quasi-constant filters (i.e. filters obtained by successive dilations of a single filter) [10, 12]. This results in a quasi-continuous wavelet transform ($QCWT$), which is a intermediate between the classical discrete-time WT and the continuous WT [10, 12] Although this approach has strong physiological motivations, the primary components, such as formants, are not well separated in the time-frequency plane and very difficult to extract.

The synchrosqueezed representation is a nonlinear transformation of this plane that mimics some aspects of the human auditory system. The synchrosqueezed plane emphasizes coherent structures of the signal (i.e. primary components). The representation is robust to acoustic distortions [9, 3], a feature attributable to the average property of the auditory nerve representation and to its phase-derivation. Examples of synchrosqueezed images are given in Figs 1–3—the frequencies tend to band in ribbons that change smoothly as a function of time. Details of the synchrosqueezing process are beyond the scope of this article. The techniques reported here focus the representation even more, by extracting a curve corresponding to each primary component. In the speech literature this type of focusing is known as saliency grouping.

The output of the synchrosqueezing process is a grid of numbers, which can be viewed as an image as in Fig. 1. The modulation model of speech assumes that speech utterances can be written as a linear combination of primary components. Each of these primary component can be expressed as an amplitude and frequency modulated sine function [7, 13, 1]. In this framework, we must extract or track the center frequency and bandwidth of the primary components of the signal as a function of time. These would be curves in time passing through the ribbons in Fig. 1. Once extracted, this information is used to build cepstral features directly usable by classical speech or speaker recognizers.

The properties of the synchrosqueezed plane with its numerous oscillations make this tracking task difficult. Most of the the classical edge or saliency extraction methods explored by the first author were unsuccessful; we list them and their shortcomings in section 3. While it may have been possible to tune some of these methods to perform adequately on our problem, the numerous difficulties led to the more model-based and constrained alternative that we present here.

Here we model the primary components, or *ribbons*, as a mixture of Gaussian densities. This implicitly interprets the (suitably normalized) synchrosqueezed measure as a bivariate probability density. The center frequency curves for each ribbon correspond to the time-conditional means of the component densities. Our model has the following features:

- The conditional means change smoothly as a function of time; this smoothness is achieved by representing the means in a basis expansion of cubic-splines.
- Curves can die, be born, or die then be reborn in time.
- Curves are constrained not to cross.

As often is the case for mixture problems, maximum likelihood and the EM algorithm provides a natural framework for estimating our model. The procedure works well on the examples we have tried, some of which we report in this paper. The nature of the algorithm allows robust extraction of primary components, at least as long as they remain perceptually well defined.

2 Examples of Synchrosqueezed Representations

In this section we present some examples which characterize typical synchrosqueezed planes; see Figs. 1–5. They consist of pixels with rather strong variations in intensity in almost all regions and in all directions. This is so even along the course of a primary component. Because of the inherent noise present in physical signals, additional peaks appear in the representation. These peaks are sometimes superposed on a ribbon, located at the edge of ribbons, bridging two ribbons, or isolated.

The synchrosqueezed plane appears as an image where the time axis is horizontal and the frequency axis is vertical. In accordance with the modulation model, the planes appear to contain groups of ribbons. Each ribbon has

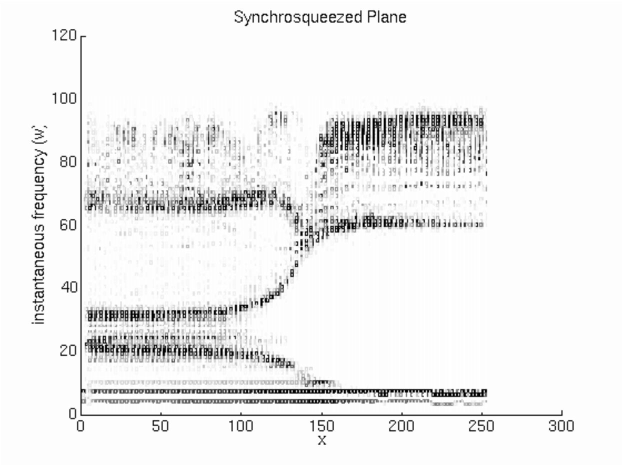

Figure 1: Synchrosqueezed plane representation for $/\underbrace{\cdots - a - a}-\underbrace{i - i - \cdots}/$. Colored noise is present with $SNR = 15\ dB$. Colored noise result from filtering white noise signals.

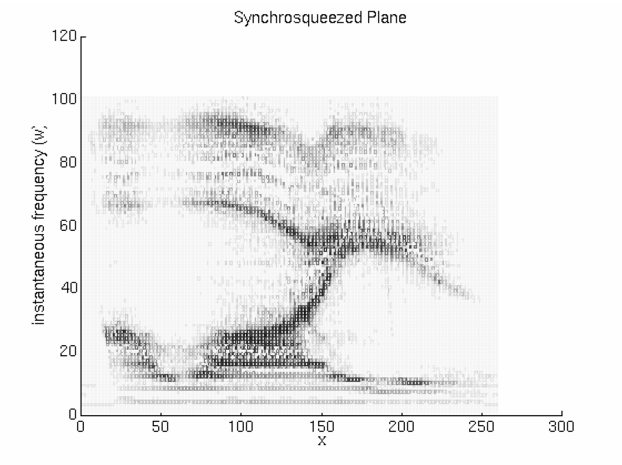

Figure 2: *Synchrosqueezed plane representation* for $/h - \delta - w - a - r - j - u\ ?/$. Colored noise is present with $SNR = 15\ dB$.

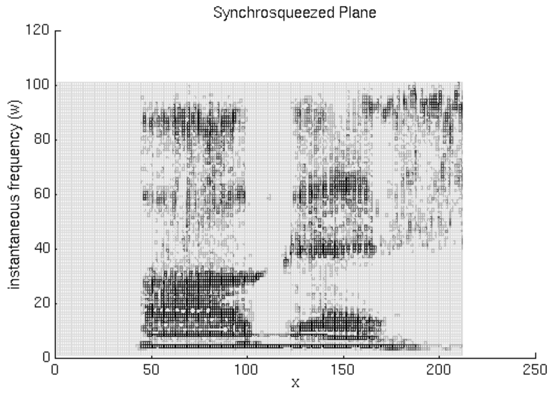

Figure 3: *Synchrosqueezed plane representation* for $/ow - g - \lambda - s - t/$. Colored noise is present with $SNR = 15\ dB$.

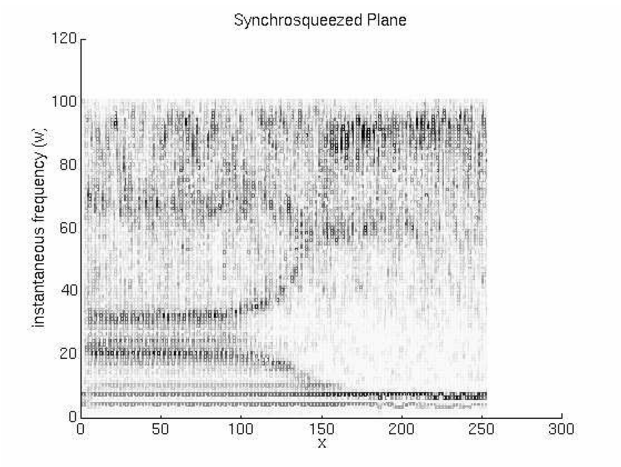

Figure 4: *Synchrosqueezed plane representation* for $/\underbrace{\cdots - a - a}-\underbrace{i - i - \cdots}/$. Colored noise is present with $SNR = 15\ dB$. Additional white noise is added with $SNR = 11dB$

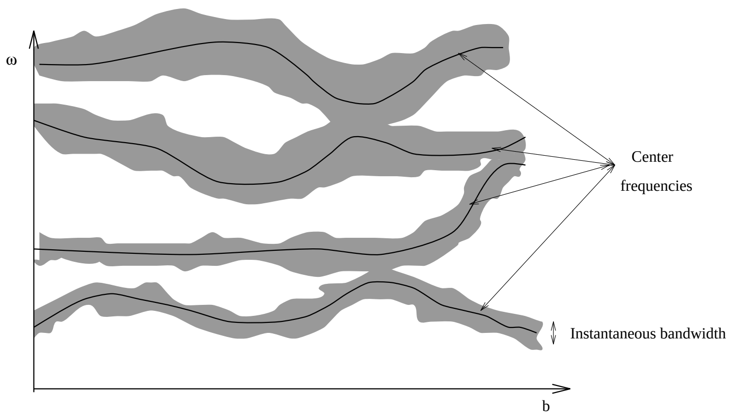

Figure 5: *Ribbons associated with speech signals in the* synchrosqueezed plane.

a center frequency and a bandwidth as illustrated in Fig. 5. These ribbons can easily be characterized by visual inspection of the images. Some of the ribbons have indeed the classical behavior of formants and the planes can be *read* by specialists just like spectrograms [5]. In accordance with the modulation model, and after studying many examples, the following constraints seem natural:

- Ribbons are not allowed to cross each other.
- Ribbons can die.
- Ribbons are allowed to merge, in which case one of them dies.
- New ribbons can be born.
- Alive ribbons have a smooth behavior and satisfy certain continuity conditions (to be defined later.)
- Dead ribbons can reappear later and behave as if they had smooth evolutions through their dead zone.
- The number of ribbons is limited but not constant.
- Ribbons propagate along the time axis: they are functions of time.
- Some ribbons are more dominant than others.

The model we propose in section 4 can be made to accommodate these constraints in a reasonably natural way.

3 Existing techniques

The numerous oscillations in the synchrosqueezed plane representation make it extremely difficult to correctly, efficiently and rapidly track the primary components. Details of the limitations of the classical image processing methods of extraction of saliencies can be found in [9], and are outlined here. The conclusion is that they all fail in many situations and therefore require supervision.

- The synchrosqueezed plane representation does not have the characteristics of a typical image. Oscillations and noise are encountered everywhere, often partially obscuring the relevant objects that constitute the ribbons. As a result, standard feature extraction and edge detection techniques for images are inefficient and inaccurate. The definition of features and edges as introduced by D. Marr [14] do not apply as almost every pixel is a singularity. Alternative definitions do not lead to any better results.

- Local non-linear variational methods for edge detection, and often closure, by minimization of pseudo-energy functionals [2] or by maximization of the energy content [17, 23], end up isolating the different pixels and therefore providing a useless granular image.

- Object oriented techniques exist for segmentation. These include morphological approaches as well as bottom-up and top-down approaches[22]. In our experiments these did not work on the synchrosqueezing measure. Texture-based method are not useful either, as there is no textural dissimilarities between the inside and outside of the ribbons that define each primary component.

- Simple low-pass filtering in the time direction also fails, with or without high-pass filtering in the frequency direction. It appears that, in order to be efficient, the low-pass filters must possess a very low cutoff frequency. This in turn blurs the behavior of the ribbons, loosing most of the gain obtained by the introduction of the synchrosqueezed measure.

- Restoration by noise cancellation techniques were applied to the synchrosqueezed plane, using the most efficient techniques available [18]. these also failed to produce any useful summaries.

- *Iterated pairing.* This method uses dynamic programming to implement saliency extraction by iterated pairing [20, 19, 21, 6]. Structures are built up locally by iteratively pairing together substructures. For the synchrosqueezed plane images, this method turned out to be very unreliable. Apart from being slow to implement, it does not have the temporal smoothness bias built in, and tends to find random structures pointing in any direction.

4 Dynamic Mixtures of Splines

None of the methods discussed capture the essence of what is visually evident in these images. Despite the noise, we see dark ribbons or bands evolving smoothly in time.

Here we model each ribbon, conditional in time, as a gaussian distribution. The conditional mean is represented by a cubic spline, and visually provides a curve running down the center of each ribbon (see Fig. 5.) The entire set of ribbons are represented as a mixture of gaussians, which naturally accommodates the regions of overlap.

While smoothness could be imposed in a number of ways, cubic splines are simple to implement, and have naturally-described smoothness properties in terms of integrated second-squared derivatives.

As is often the case with mixture models, we rely on a version of the iterative EM-algorithm [4] to fit our mixture model by maximum likelihood estimation. While EM algorithms are notoriously slow, a combination of good starting values and the spatial smoothness ensure rapid convergence. In addition, our algorithm evolves in time, and allows us to enforce the additional constraints on our model in a natural, albeit *ad-hoc* way.

Our statistical model seems to be more robust to noise in the images than the other approaches; indeed, noise is explicitly present in the statistical model, and is averaged out in the estimation process. Although tailored to suit our application here, the method has potential applications for edge and feature extraction, and contour closing in classical image processing.

4.1 Description of the mixture model

We interpret the synchrosqueezed measure $\mu(t, \omega)$ as a bivariate probability density function in the synchrosqueezed plane $\mathcal{V}$. Our actual observations are discrete, and treated as frequency counts at each of the lattice points $\mathcal{V}_{(t,\omega)}$. This measure is approximated by a statistical model

$$P(\omega, t) = P(\omega|t, \nu)P(t), \tag{1}$$

where the conditional distribution $P(\omega|t)$ is modeled as a mixture of Gaussians:

$$P(\omega|t, \nu) = \sum_{k=1}^{N(t)} \pi_k(t)P(\omega|t, k, \nu) \tag{2}$$

The notation is as follows:

- $$P(\omega|t, k, \nu) = \frac{1}{\sigma_k(t)\sqrt{2\pi}} e^{\frac{-(\omega - c_k(t))^2}{2\sigma_k^2(t)}} \tag{3}$$

  is a gaussian density for ω, given t and that it belongs to the kth ribbon.
- $\nu = \{N(t), \pi_k(t), c_k(t), \sigma_k(t)\}$ is a placeholder for all the parameters in the model.
- Initially, the model assumes that $N(0) \leq N_{pc}$ primary components are present in the synchrosqueezed plane. The number $N(t)$ can change with time.
- $c_k(t)$ is a function that describes the center frequency of the kth ribbon at time t.
- $\sigma_k^2(t)$ is the variance associated with the kth ribbon at time t. It can be interpreted as the bandwidth of the associated primary component.
- $\pi_k(t)$ is the mixture proportion associated with the kth ribbon, which can also change with time.

The marginal distribution of time t is left intact:

$$P(t) = \int_{\omega \in \mathcal{V}} \mu(t, d\omega) \tag{4}$$

We represent the center frequencies $c_k(t)$ by an expansion in a basis of cubic splines with pre-chosen knots:

$$c_k(t) = c_k(t, \theta_k) \tag{5}$$

where θ_k stands for all the parameters that define the kth cubic spline. In compliance with the definition of a formant, these c_k will be constrained not to cross each other.

The parameters $\pi_k(t)$, θ_k and $\sigma_k^2(t)$ are fit by maximizing the cross entropy

$$\int_{(\omega,t)\ \in\ \mathcal{V}} \log\left[P(\omega|t,\nu)\right] P(t)\mu(d\omega, dt) \tag{6}$$

We can view the (normalized) synchrosqueezed measure $\mu(\omega, t)$ as a set of empirical frequency weights defined over the discretized domain $\mathcal{V}_{(t,\omega)}$. Then the cross-entropy integral (6) becomes a sum

$$\sum_{(\omega,t)\ \in\ \mathcal{V}_{(t,\omega)}} \mu(\omega,t)\log\left[P(\omega|t,\nu)\right] \tag{7}$$

and the cross entropy looks more like a traditional (weighted) log-likelihood.

4.2 Maximizing the likelihood

Since the log-likelihood factors into two pieces, we see that the maximizing value for $P(t)$ is simply the marginal observed measure as given in (4).

This leaves the conditional part, which we maximize over the parameters ν. We use the *Estimation-Maximization* algorithm[4], which simplifies the optimization process, and allows us to easily incorporate the extra bells and whistles required to enforce the constraints we have outlined. The generic EM algorithm consists of iterating the two steps:

E-step: Compute the *responsibility* of the kth ribbon for the pixel at (t, ω):

$$P(C_k|\omega,t) = \frac{\pi_k(t)P(\omega|t,k,\nu)}{\sum_{\ell=1}^{N(t)} \pi_\ell(t)P(\omega|t,\ell,\nu)} \tag{8}$$

These apportion the contribution $\mu(t,\omega)$ to each of the ribbons, using the current estimates of the components of the mixture densities. The essential information is how close ω is to each of the $c_k(t)$, measured in the appropriate units.

M-step: The parameters for each ribbon are computed by maximizing the appropriate weighted likelihood. The spline parameters for the kth ribbon are estimated by weighted least squares:

$$\tilde{\theta}_k = \mathrm{argmin}_{\theta_k} \int_{(t,\omega)\in\mathcal{V}} \frac{(\omega - c_k(t,\theta_k))^2}{\sigma_k^2(t)} P(C_k|\omega,t)\mu(dt,d\omega) \tag{9}$$

The variance function is computed by weighted least squares at each point in time:

$$\tilde{\sigma}_k^2(t) = \frac{1}{\int_{\omega \in \mathcal{V}|t} \mu(t, d\omega)} \int_{\omega \in \mathcal{V}|t} (\omega - c_k(t, \tilde{\theta}_k))^2 P(C_k|\omega, t)\mu(t, d\omega) \quad (10)$$

The mixture proportions are estimated from the total weight allocated to each ribbon at each point in time:

$$\tilde{\pi}_k(t) \sim \int_{\omega \in \mathcal{V}|t} P(C_k|\omega, t)\mu(t, d\omega), \quad (11)$$

and normalized to sum to one.

While this sets the general framework, there are many additional details required to make the procedure work effectively. Even standard mixture models pose challenges, such as sensitivity to starting values. We now present the special implementation details that were required to fit these models in the context of our speech data.

4.3 Implementation Details

Our approach is to fit the model sequentially in time. The advantage of this is that we can build in the various restrictions particular to this problem in a more natural fashion. These include the birth and death of ribbons, as well as the crossing restrictions.

We segment the time axis into $N + 1$ overlapping bands $\mathcal{V}_{n(t,\omega)}$ with $\cup_{n=0}^N \mathcal{V}_{n(t,\omega)} = \mathcal{V}_{(t,\omega)}$, illustrated in Fig. 6. We grow the ribbons from left to right, advancing one band at a time. In the first round the fitted model in the previous band provides initial conditions for the estimates in the current band. The left to right pass is repeated several times, until the curves stabilize.

4.3.1 Initialization in the left-most band

The method requires initial values for the $c_k(t)$ and $\sigma_k^2(t)$ and the number of components $N(0)$. We start of in the left-most band $\mathcal{V}_{0(b,\omega)}$, and initialize and run a K-means clustering algorithm on the marginal distribution of ω, confined to that band.

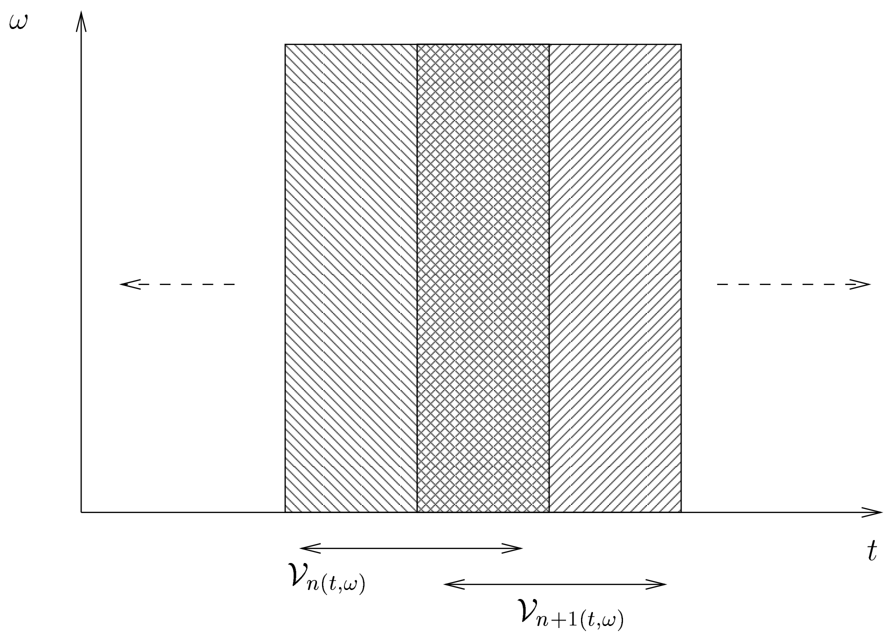

Figure 6: *Segmentation of the* synchrosqueezed plane*. The plane is segmented along the t axis, by creating N overlapping bands as shown.*

The K-means algorithm itself requires initialization, which in our case is based on the marginal histogram $\mu(t, \omega)$ in this band. The number $N(0) \leq N_{pc}$ and location of the dominant peaks of this histogram determine the initial guesses for $c_k(t) = c_k$ in this first band.

When modeling speech, in order to guarantee the detection of the pitch, its first harmonics and the first formant, one usually selects the first maxima of the histogram, even if they are less dominant, before sorting the peaks and selecting the remaining dominant peaks.

These are then used to seed the Lloyd's *K-mean* algorithm [8, 15], which is similar in spirit to the EM algorithm, but much simpler. K-means alternates between the two simple steps:

Partition: Assign each pixel to the closest (in ω) center c_k.

Update: Replace the center c_k by the weighted mean of all the pixel locations assigned to it, using the measure $\mu(t, \omega)$ to assign the weights.

This is repeated until the change in the centers is negligible. At this point the fitted curves for $c_k(t)$ are taken to be constant in the first band. The next section shows how we convert these to piecewise linear curves, and finally piecewise cubic.

4.3.2 Progressing from band $n-1$ to band n

We now describe how the algorithm updates its estimates in the nth band, having computed the curve $c_k(t)$ through the rightmost edge of the $n-1$th band.

In the initial pass from left to right, we use piecewise-linear splines to represent the $c_k(t)$. The knots t_i are placed uniformly along the time axis; a typical number is $I = 8$ knots per band. The splines are linear between the knots and join continuously at the knots. Associated with these knots are piecewise-linear basis functions $\xi_j(t)$ that represent the family of linear splines. The center for the kth ribbon thus has a representation

$$c_k(t, \theta_k) = \sum_{j=1}^{J} \theta_{jk} \xi_j(t) \tag{12}$$

where J is the total number of such basis functions. If the boundary knots are included, this is equal to the number of knots for linear splines; in our case $J = I(N+1)/2$.

Among the many choices possible for the basis functions, we chose a particular simple form that suits our sequential scheme. We pick $\xi_i(t)$ such that

$$\begin{aligned} \xi_i(t_i) &= 1 \\ \xi_i(t_j) &= 0, \; j \neq i \end{aligned} \tag{13}$$

Figure 7 shows some examples. This means that the coefficients are the fitted

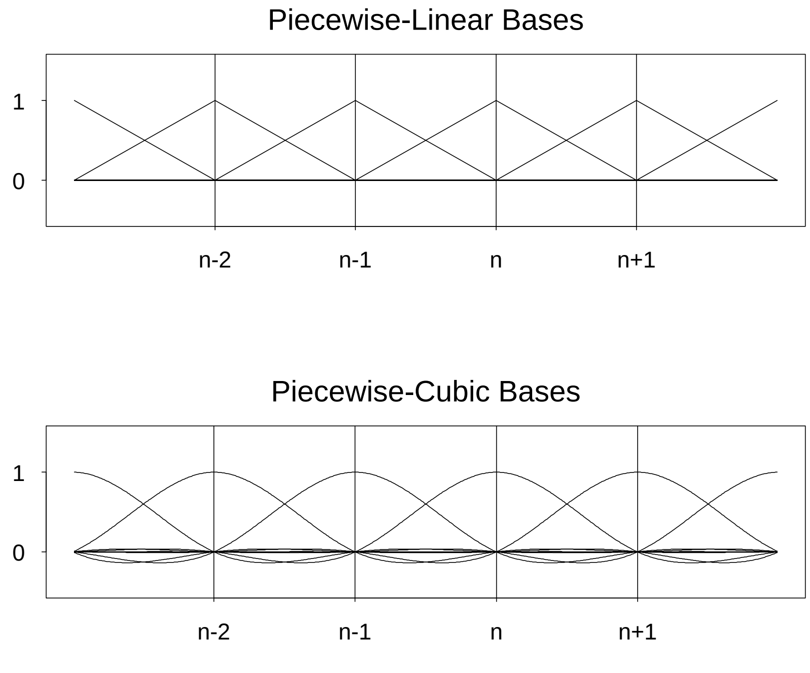

Figure 7: *Both panels shows $\xi_n(t)$ for a series of knots. The top are piecewise linear, the bottom are piecewise cubic. Notice how each function is exactly* 1 *at one knot, and zero at all other knots.*

values at the knots — $c_k(t_i) = \theta_{ik}$ — and hence

$$c_k(t, \theta_k) = \sum_{i=1}^{J} c_k(t_i)\xi_i(t) \tag{14}$$

This means that by changing the I fitted values corresponding to the knots in any given band, the continuity conditions with neighboring bands are automatically satisfied. This has even stronger consequences for cubic splines.

Upon entering the nth band, the $I/2$ leftmost values of $c_k(t_i)$ are set from band $n-1$. We initialize the rightmost $I/2$ values by extrapolating the linear spline in the left hand side of the band. This is the reason we initially use piecewise-linear rather than cubic splines, since they extrapolate in a more stable fashion.

In doing this, the following precautions are taken:

- if any two curves $c_k(t)$ cross during this extrapolation, we interchange their coefficients from the point of crossing onwards. Such crossings of primary components are not consistent with the speech model [1, 9].

- We repeat the histogram procedure in each band, to check for the birth of new ribbons not accounted for by the current model. If a new peak is located, a new ribbon is created and added to the model, and $N(t)$ is incremented

With these initial values the parameters are updated according to (9), which amounts to weighted least squares within the band.

In principal we could iterate the E- and M-step within each band. We have found one iteration of each per visit to be more efficient, given that we make several passes from left to right through the bands.

At the first pass, the probabilities $P(C_k|\omega,t)$ are computed using the initial values for $c_k(t)$ in the the nth band just described. After $c_k(t)$ is updated, so are the $P(C_k|\omega,t)$, and are available for the next pass through that band.

The procedure is slightly simpler for subsequent passes through the bands, although after the first we replace the linear splines with cubic splines. These are cubic polynomials between the knots, and join up with continuous derivatives of degree 0, 1 and 2. We use the same notation for the basis functions $\xi_{ik}(t)$, and they can be chosen to have the same properties (13) as the linear bases (see Fig. 7). Cubic splines require in addition some derivative boundary conditions in the outermost bands, but this is a detail and easily accommodated.

Again, the probabilities $P(C_k|\omega,t)$ used are the latest available: the left $I/2$ just computed for band $n-1$, and the right $I/2$ from the previous pass through the nth band.

4.3.3 Further details

After determination of the center frequencies, the variances $\sigma_k^2(t)$ and mixture proportions $\pi_k(t)$ are computed using equations (10) and (11). These are available, just like the probabilities, for the next pass through the present band, as well as all following bands. We do not explicitly build smoothness into the variance and mixture functions, and as stated they will be piecewise constant. In fact after a complete pass from left to right, the variances and mixture proportions are filtered in the time direction to create smooth functions of time as well.

Once the center frequencies $c_k(t)$ and variances $\sigma_k^2(t)$ are computed, the total synchrosqueezed measure of each extracted ribbon is computed in the current band:

$$\tilde{E}_k(n) \;=\; \frac{\int_{(t,\omega)\in\mathcal{V}_n} P(C_k|\omega,t)\mu(t,d\omega)}{\int_{(t,\omega)\in\mathcal{V}_n} dtd\omega}, \tag{15}$$

$\tilde{E}_k(n)$ is compared with the values of its closest neighbors $\tilde{E}_{k\,\pm\,1}(n)$, with the measure of the surrounding background region (not belonging to any extracted ribbon), and to the previous value along the ribbon $\tilde{E}_k(n-1)$. Based on these four values, a decision is taken (using a decision tree built by trial and error) to keep the ribbon or to kill it.

Special attention has to be paid not to inopportunely kill primary components. The behavior of the other ribbons is taken into account, and if it seems that the region is *unvoiced*, no component is killed, unless it does not rapidly reach another region where they are "well-defined". Actually, in unvoiced sounds, the resonances of the vocal tract still introduce peaks in the transfer function, which are still followed by the tracking algorithm, despite the fact they are smeared. Hence as a general strategy, unless the behavior becomes erratic, when the decision is taken to kill an extracted ribbon, it is rather put under watch and the decision is finally taken depending on its behavior in the next band. Of course, during the next pass through this band, the decision may be reversed based on the knowledge of the future of the ribbon (behavior in bands n+1, n+2 etc).

Figures 8 and 9 present examples of extracted ribbons. In Fig. 9, some non-formant components appear as explained in section 3. They can easily be eliminated with energy thresholds. By reducing these same thresholds, and by increasing the maximum number N_{pc} of primary components to be extracted, more and more negligible primary components are obtained. This

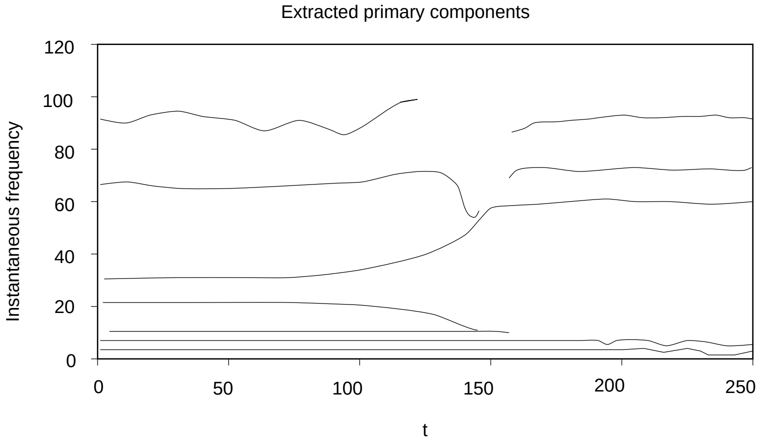

Figure 8: Extracted primary components for $/\underbrace{\cdots - a - a}-\underbrace{i - i - \cdots}/$. Colored noise is present with $SNR = 15\ dB$.

allows a continuous transition between a formant-based modulation model which represents speech only by its formants and a sine-wave representation which introduces a large amount of sinusoidals for the same purpose [9].

Figure 10 illustrates the successful performance of the algorithm when the synchrosqueezed plane is corrupted by additive noise, which as observed in Fig. 4 does not modify the behavior of the ribbons.

In practice, with the strategy used to determine the first guesses, only two passes through the whole $\mathcal{V}_{(t,\omega)}$ are usually required. This method can be used for any time-frequency or time-scale representations.

5 Conclusions

We have described a novel approach to curve fitting, which adapts the standard EM algorithm to a particularly tough speech modeling context. Although the background theory is relatively standard, we have shown how it can be adapted to solve the intricacies of the extraction of primary components. After struggling with many of the contemporary techniques with little success, this is the only method that gives satisfactory results. This is a case where standard models are unlikely to be successful in their raw form. The success of our approach was largely due to the *engineering* of the EM

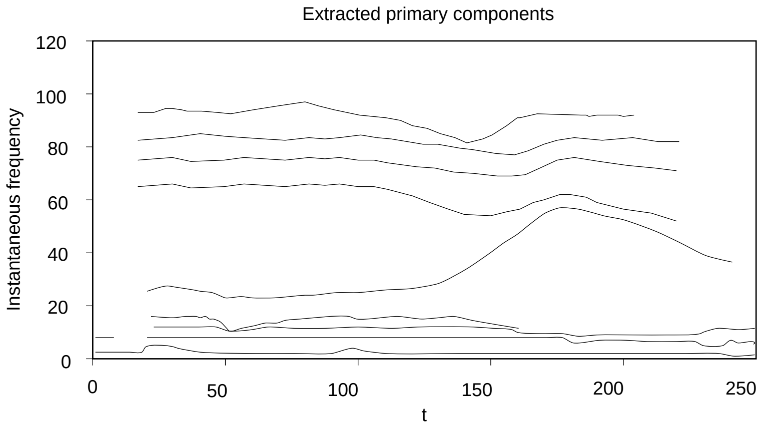

Figure 9: Extracted primary components for $/h - \delta - w - a - r - j - u\ ?/$. Colored noise is present with $SNR\ =\ 15\ dB$.

algorithm, to allow for the constraints and restrictions natural for the modulation model. It is unlikely that we will be able to prove any hard theory for this approach. In practice the algorithm will get refined and tailored as new and different idiosyncrasies occur.

We feel that the method may find some use (in some modified form) for other time-frequency or time-scale representations, especially in cases where classical methods fail.

The method may be worth pursuing in image processing contexts, for example for feature extraction, grouping and connection. We suggest the following generic algorithm:

- High-pass filter the signal with a radial filter.
- Apply the method in the x direction.
- Apply the method in the y direction.
- If after superposition of the two results bifurcations occur, decide to select the x or the y evolution.
- Iterate the algorithm.

The final output presents connected edges and closed contours wherever there are perceptually closed. To date we have only made a preliminary investigation of this method.

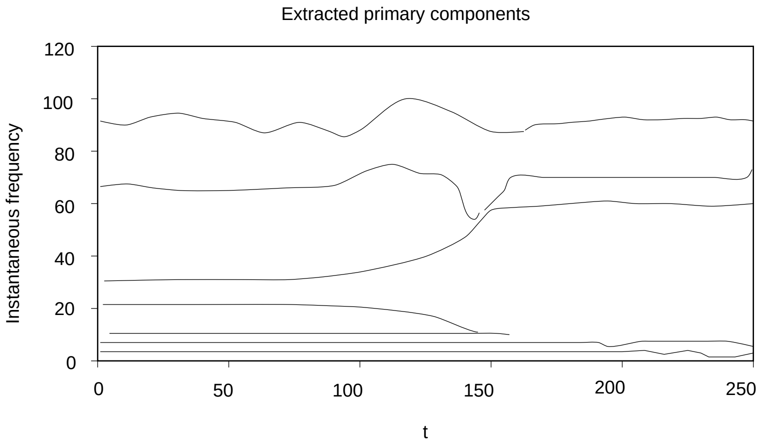

Figure 10: Extracted primary components for $/\underbrace{\cdots - a - a} - \underbrace{i - i - \cdots}/$. A pink noise is present with $SNR \ = \ 15 \ dB$. White noise is added with $SNR \ = \ 10 \ dB$.

Acknowledgments

Stéphane H. Maes thanks his research advisor, Professor I. Daubechies, for her help, encouragement and guidance during this research. Special thanks to Professor R. Mammone and Professor J. L. Flanagan for their comments, suggestions and hospitality in the speech recognition team at *CAIP*, Rutgers University. Finally many thanks to his thesis advisors, Professor J-P. Antoine and Professor P. Delogne. This work was performed when Dr. Maes was a Research Assistant of the Belgian National Fund for Scientific Research (*FNRS*). He was affiliated to *CAIP* Center, Rutgers University, Piscataway, NJ 08855, USA and Unités *TELE* and *FYMA*, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Trevor Hastie was partially supported by grant DMS-9504495 from the National Science Foundation, and grant ROI-CA-72028-01 from the National Institutes of Health. Part of this work was performed while Dr. Hastie was a member of the research staff at AT&T Bell Laboratories, Murray Hill, New Jersey.

References

[1] K. Assaleh. *Robust features for speaker identification.* PhD thesis, CAIP Center - Rutgers University, The State Univeristy of New Jersey, New

Brunswick, NJ, 1993.

[2] V. Caselles. A geometric model for active contours in image processing. Technical report, Ceremade, centre de recherce des mathématiques de la décision, 1987. *Cahier de mathématiques de la décision*, 9210.

[3] I. Daubechies and S. Maes. Wavelets in medicine and biology. In A. Alroubi and M. Unser, editors, *A nonlinear squeezing of the continuous wavelet analysis based on auditory nerve models.* CRC Press, April 1996.

[4] Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39:1–38, 1977.

[5] J. L. Flanagan. *Speech analysis synthesis and perception.* Springer-Verlag, 1972.

[6] W. T. Freeman. *Steerable filters and local analysis of image structure.* PhD thesis, M.I.T, Cambridge, MA, June 1992.

[7] J.F. Kaiser. Some useful properties of the Teager's energy operators. In *Proc. IEEE ICASSP*, pages II.149–II.152, 1993.

[8] S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2 (part2)):pp. 129–136, 1982.

[9] S. Maes. *The wavelet tranform in signal processing, with application to the extraction of the speech modulation model features.* PhD thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, June 1994.

[10] S. Maes. Fast quasi-continuous wavelet algorithms for analysis and synthesis of 1-D signals. *to be published in SIAM J. Applied Math. (Nov. 1997)*, April 1995.

[11] S. Maes. Robust speech and speaker recognition using instantaneous frequencies and amplitudes obtained with wavelet-derived synchrosqueezing measures. In *Program on Spline Functions and the Theory of Wavelets*, Montreal, Canada, March 1996. Centre de Recherches Mathématiques, Université de Montréal. invited paper.

[12] S. Maes. Signal analysis and synthesis with 1-D quasi-continuous wavelet transform. In *Proc. 12th. International Conference on analysis and optimization of systems*, Paris, June 1996. IRSIA.

[13] P. Maragos, J. F. Kaiser, and T.F. Quatieri. Energy separation in signal modulation with application to speech analysis. *IEEE Trans. Signal Proc.*, 41(10):pp. 3024–3051, October 1993.

[14] D. Marr and E. Hildreth. Theory of edge detection. *Proc. of Royal Soc. of London*, pages pp. 187–217, 1980.

[15] J. Max. Quantizing for minimum distorsion. *IRE Trans. Inform. Theory*, 6 - 1(3):pp. 7–12, 1960.

[16] O. Rioul and M. Vetterli. Wavelet and signal processing. *IEEE Sign. Proc. Magazine*, pages 14–38, October 1991.

[17] C. Ronse. On idempotence and related requirements in edge detection. *preprint*, 1993.

[18] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithm. In *Modélisations mathématiques pour le traitement d'image*, Problèmes non linéaires appliqués, Rocquencourt, France, 1992. Ecoles CEA - EDF -INRIA, March 16-18.

[19] A. Sashua. Structural saliency: The detection of globally salient structures using a locally connected network. Revised version, Dept. Appl. Math. and Comp. Science, The Weizmann Inst. of Science, Rehovot, Israel, 1990.

[20] A. Sashua and S. Ullman. Structural saliency: The detection of globally salient structuresusing a locally connected network. In *Proc. 2nd Int. Conf. Computer Vision*, pages 321–327, December 1988.

[21] A. Sashua and S. Ullman. Grouping contours by iterated pairing network. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Neural Information Processing Systems*, volume 3. Morgan Kauffman Publ., San Mateo, CA, 1991.

[22] J. Serra and L. Vincent. An overview of morphological filtering. *Circuits Systems and Signal Process.*, 11(1):pp. 47–108, 1992.

[23] S. Venkatesh and R. Owens. Implementation detail of a feature detection algorithm. Technical Report 89/12, Univ. Western Australia, Dpt. Comp. Sc., 1989.