ROBUST ESTIMATION OF SPECTRAL CENTER-OF-GRAVITY TRAJECTORIES USING MIXTURE SPLINE MODELS

Don X. Sun

Department of Statistical Models and Methods Research
AT&T Bell Laboratories, Murray Hill, NJ 07974-0636, USA
dxsun@research.att.com

ABSTRACT

This paper presents a novel approach to the estimation of the trajectories of spectral center-of-gravity using robust statistical models with penalized weighted spline smoothers. Most of the existing methods for tracking speech formant trajectories are based on dynamic programming algorithms with certain continuity constraints on the formant frequencies ([6, 8, 9]). The objective functions (or loss functions) in these approaches are usually *ad hoc* and have very complex expressions that are difficult to optimize. Also, many existing methods rely on the accuracy of the LPC spectral peaks and are not very robust against possible missing or spurious peaks. In this paper, instead of using the peaks of the LPC spectral functions, we propose a new approach to the estimation of the "center-of-gravities" in spectrogram using mixture models of spline smoothers ([5, 10]). There are two major advantages in this new approach: 1) it is robust against missing peaks and spurious peaks that may occur in LPC peak finding algorithms; 2) trajectory smoothness is guaranteed by the properties of spline regression models.

1. INTRODUCTION

It is well known that significant amount of information of human speech can be obtained from the spectrogram of speech signal [11, 12]. In particular, spectral peaks or center-of-gravity trajectories play an important role in automatic speech recognition [7].

In addition to the traditional speech spectrograms that are based on Fourier transformation and linear prediction analysis, some other "spectrogram"-like speech representations have received great attentions among the research community in the recent years, such as the auditory model based analysis, perceptual linear predictive analysis [3], wavelet transformation ("waveletogram") [2] etc.

A common method of utilizing the information in the spectrogram is to extract the formant frequencies for the vowel-like sounds. Most of the formant estimation methods rely on the peaks of the spectra of the linear predictive coefficients (LPC) [1] or the complex roots of the polynomial of the z transformation of a linear predictor [9]. Also, dynamic programming methods are often used to obtain smoothed tracks of formant trajectories with some *ad hoc* objective functions [4, 6, 8, 9].

In this paper, we propose a new method of estimating center-of-gravity trajectories or "interesting" clusterings of any spectrogram-like speech representation. The new method utilizes the well known theory of spline curve smoothing and modeling techniques based on mixture distributions.

There are two major advantages of this method comparing to many other existing methods: 1) it does not require the estimation of formant candidates or peaks of the LPC spectra, hence it is robust against missing peaks and spurious peaks that may occur in peak identification procedure; 2) the trajectory smoothness is guaranteed by the properties of the spline models.

2. CURVE FITTING WITH PENALIZED LEAST SQUARES METHOD

We first review some fundamentals of the cubic spline curve fitting or smoothing method.

Suppose we have N observations $y_1, \cdots, y_N$ at N distinct knots $a = x_1 < x_2 < \cdots < x_N = b$, and we assume that the data can be represented by the following model

$$y_i = S(x_i) + \epsilon_i, \quad i = 1, \cdots, N \tag{1}$$

where $S(\cdot)$ is a deterministic function characterizing the major relationship between x's and y's, and the ϵ_i's are independent random variables assumed to have Gaussian distributions.

The objective is to estimate the function $S(\cdot)$ such that it is as close to the data path as possible, yet, on the other hand, as smooth as possible. This idea can be formulated through the following objective function:

$$L(S) = \eta \int_a^b [S''(x)]^2 \, dx + \sum_{i=1}^{N} \lambda_i (y_i - S(x_i))^2 \tag{2}$$

where η is the smoothness parameter and the λ_i's are the weights associated with y_i's, representing the relative contribution of the ith observation to the model estimation. It is easy to see that a larger η leads to a smoother curve and a larger λ_i causes the estimated value of $S(x_i)$ to be closer to the observation y_i.

It is well known that the problem of minimizing the objective function in (2) has a unique explicit solution [5], which is indeed the natural cubic spline function. The estimation procedure is briefly described as follows. Let

$$(\mathbf{Y})_{N\times 1} \quad = \quad (y_1, \cdots, y_N)^T$$

$$\begin{aligned}(\mathbf{F})_{N\times 1} &= (f_1,\cdots,f_N)^T = (S(x_1),\cdots,S(x_N))^T \\ (\mathbf{A})_{(N-2)\times 1} &= (A_2,\cdots,A_{N-1})^T = (S''(x_2),\cdots,S''(x_{N-1}))^T\end{aligned}$$

denote the observation vector, function values of $S()$ and the second derivatives. Let $h_i = x_i - x_{i-1}$ denote the spacing of the x variable. The estimate of $S()$ is obtained through the estimation of f_i's and A_i's.

Let

$$\mathbf{B} = \begin{bmatrix} \frac{(h_2+h_3)}{3} & \frac{h_3}{6} & 0 & \cdots & 0 \\ \frac{h_3}{6} & \frac{(h_3+h_4)}{3} & \frac{h_4}{6} & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \frac{h_{N-1}}{6} \\ 0 & \cdots & 0 & \frac{h_{N-1}}{6} & \frac{(h_{N-1}+h_N)}{3} \end{bmatrix} \quad (3)$$

$$\mathbf{D} = \begin{bmatrix} \frac{1}{h_2} & -\frac{1}{h_2}-\frac{1}{h_3} & \frac{1}{h_3} & & \cdots & 0 \\ 0 & \frac{1}{h_3} & -\frac{1}{h_3}-\frac{1}{h_3} & \frac{1}{h_4} & & \vdots \\ \vdots & \ddots & & \ddots & & \\ 0 & \cdots & & & & \frac{1}{h_N} \end{bmatrix} \quad (4)$$

$$\mathbf{\Lambda} = \text{diag } (\lambda_1,\cdots,\lambda_N) \quad (5)$$

where $\mathbf{B}$ is an $(N-2)\times(N-2)$ matrix and $\mathbf{D}$ is an $(N-2)\times N$ matrix. Then, the solution to the minimization problem is

$$\begin{aligned}\mathbf{A} &= (\eta\mathbf{D}\mathbf{\Lambda}^{-1}\mathbf{D}^T + \mathbf{B})^{-1}\mathbf{D}\mathbf{Y} \quad (6) \\ \mathbf{F} &= \mathbf{Y} - \eta\mathbf{\Lambda}^{-1}\mathbf{D}^T\mathbf{A}. \quad (7)\end{aligned}$$

To evaluate the function $S()$ at any given point x within $[a,b]$, say between x_{i-1} and x_i, we can use the following formula.

$$S(x) = \frac{1}{6h_j}\Big[A_{i-1}(x_i-x)^3 + A_i(x-x_{i-1})^3 + (6f_{i-1} - A_{i-1}h_i^2)(x_i - x) + (6f_i - A_ih_i^2)(x-x_{i-1})\Big]$$

where A_1 and A_N are set to zero by assumption.

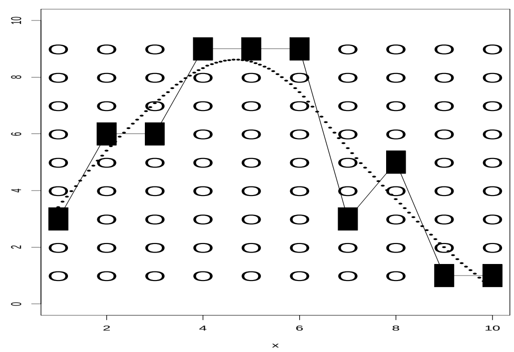
3. SPLINE ESTIMATION FOR MULTIPLE OBSERVATIONS

The previous formulation is only for the case where all the observations y_i's have distinct knots x_i's, since identical x_i's will imply zero h_i's, which are not invertible. However, we can formulate this case in a slightly different way to avoid the singularity problem. Let $a = x_1 < x_1 < \cdots < x_N = b$ be the sequence of distinct knots among all the x_i's, and y_{ij} be the jth observation at knots x_i with weight λ_{ij}, $i = 1,\cdots,N; j = 1,\cdots,n_i$.

Let

$$\begin{aligned}\mathbf{\Lambda}^* &= \text{diag } (\sum_{j=1}^{n_1}\lambda_{1j}, \sum_{j=1}^{n_2}\lambda_{2j},\cdots,\sum_{j=1}^{n_N}\lambda_{Nj}) \\ \mathbf{Y}^* &= (\sum_{j=1}^{n_1}\lambda_{1j}y_{1j}, \sum_{j=1}^{n_2}\lambda_{2j}y_{2j},\cdots,\sum_{j=1}^{n_N}\lambda_{Nj}y_{Nj})^T\end{aligned}$$

Figure 1. An alternative view of spline fitting



then the solution is almost identical to that in (6) and (7), i.e.,

$$\begin{aligned}\mathbf{A} &= (\eta\mathbf{D}(\mathbf{\Lambda}^*)^{-1}\mathbf{D}^T + \mathbf{B})^{-1}\mathbf{D}(\mathbf{\Lambda}^*)^{-1}\mathbf{Y}^* \quad (8) \\ \mathbf{F} &= (\mathbf{\Lambda}^*)^{-1}(\mathbf{Y}^* - \eta\mathbf{D}^T\mathbf{A}). \quad (9)\end{aligned}$$

4. MIXTURE MODEL OF MULTIPLE SPLINES

In order to illustrate the idea of mixture spline models, we first introduce a different view of the spline fitting method discussed in the previous sections. The equivalence of the two ways of formulating the spline smoothing problem will enable us to estimate the center-of-gravity trajectories of spectrogram-like speech representations.

Suppose we observe $y_1,\cdots,y_N$ at N distinct knots $a = x_1 < x_2 < \cdots < x_N = b$, with weights $\lambda_1,\cdots,\lambda_N$. Let $(z_1,\cdots,z_J)$ be an ordered set of all the distinct values in $(y_1,\cdots,y_N)$. Now we can represent the original data path by a rectangle array of values, (u_{ij}), or more precisely, (x_i, z_j, u_{ij}), where

$$u_{ij} = \begin{cases} 1, & \text{if } z_j = y_i; \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$L^*(S) = \eta\int_a^b [S''(x)]^2\,dx + \sum_{i=1}^N\sum_{j=1}^J u_{ij}\lambda_{ij}^*(z_j - S(x_i))^2 \quad (10)$$

where $\lambda_{ij}^* = u_{ij}\lambda_i$. It can be easily verified that (10) is equivalent to (2). This formulation is illustrated in Figure 1, where the hollow circles represent the points with $u_{ij} = 0$ and the solid boxes represent the points with $u_{ij} = 1$. The dotted line in the figure indicated the fitted spline model.

Now, we are ready to estimate mixture of multiple splines for the center-of-gravity trajectories of a spectrogram-like speech representation. Let $(x_1,\cdots,x_N)$ represent the time axis of a spectrogram, and $(z_1,\cdots,z_J)$ represent the frequency axis of a spectrogram. We use u_{ij} to represent the intensity or magnitude in a spectrogram associated with time frame x_i and frequency component z_j.

Suppose we want to identify M spline models corresponding to the center-of-gravity trajectories in a spectrogram: $S_m(x)$, $m = 1, \cdots, M$. To quantify the "closeness" of an observation to the mth spline, let w_{mij} denote the probability of observation (x_i, z_j, u_{ij}) belonging to the mth spline, and let σ_m^2 denote the variance of the error distribution associated with the mth spline mixture. To estimate $S_m()$, w_{mij}, and σ_m^2, we propose the following algorithm.

1. Obtain initial estimates of $S_m(x)$, and σ_m^2. One way to do this is to divide the frequency axis into M groups, and fit a single spline model for the observations within each group. Another way to do this is to estimate the peak values of the LPC spectra, and take the average value of each formant over all the time frames as the initial value of $S_m()$.
2. For given initial values of $f_{mi} = S_m(x_i)$, $i = 1, \cdots, N$; $m = 1, \cdots, M$, the probability of observation (x_i, z_j, u_{ij}) belonging to the mth mixing component is:

$$w_{mij} = \frac{\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_m} \exp\left\{-\frac{1}{2\sigma_m^2}(z_j - S_m(x_i))^2\right\}}{\sum_{m=1}^{M} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_m} \exp\left\{-\frac{1}{2\sigma_m^2}(z_j - S_m(x_i))^2\right\}} \tag{11}$$

3. For given w_{mij}'s , we can estimate each individual spline separately by minimizing the following objective function

$$L(S_m) = \eta \int_a^b [S_m''(x)]^2 \, dx + \sum_{i=1}^{N} \sum_{j=1}^{J} w_{mij} \, u_{ij}(z_j - S_m(x_i))^2. \tag{12}$$

   The solution to this problem is identical to that in (8) and (9) by replacing λ_{ij}^* with $w_{mij}u_{ij}$.
4. The error variance σ_m^2 can be estimated by

$$\sigma_m^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} w_{mij} \, u_{ij}(z_j - S_m(x_i))^2}{\sum_{i=1}^{N} \sum_{j=1}^{J} w_{mij} \, u_{ij}.} \tag{13}$$

5. Go to step 1 until convergence occurs, say, when the decrease of the objective function is smaller than a prespecified small constant.

5. EXAMPLES

In this section, we illustrate the proposed method by the example shown in Figure 2–7.

Figure 2 shows the initial estimates of the three spline models (three straight horizontal lines). Figure 3–5 show the three separate mixing components in the first iteration and Figure 6 shows the overall fitting when te first iteration finished. Similarly, Figure 7 shows the overall fitting when te fifth iteration finished. The improvement along the progress of the algorithm can be easily noticed.

6. CONCLUSIONS

Significant amount of information of human speech is contained in the spectral clustering and center-of-gravity trajectories. In this paper, we proposed a new method for estimating such interesting features in a spectrogram using

Figure 2. Iteration: 0 (Initial estimates)

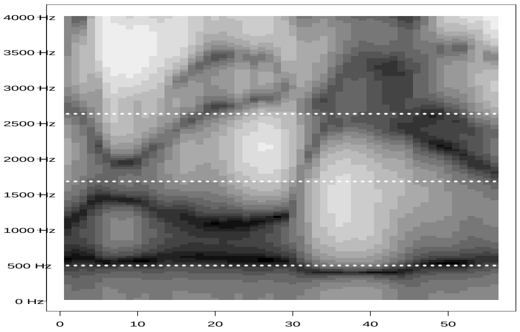


Figure 3. Iteration: 1 (1st mixing component)

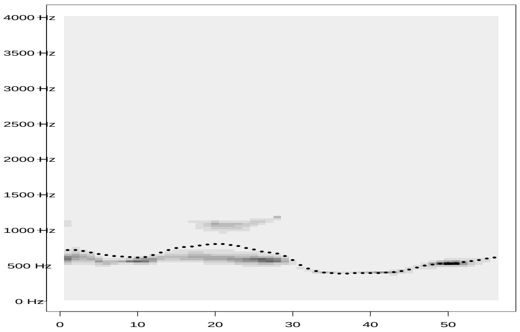


Figure 4. Iteration: 1 (2nd mixing component)

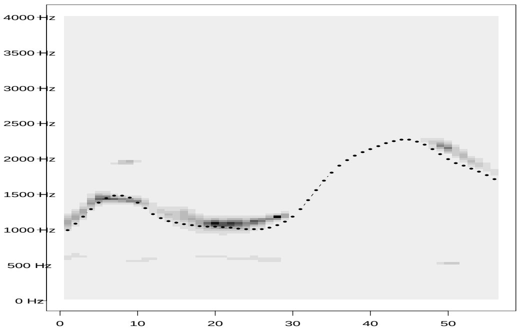


Figure 5. Iteration: 1 (3rd mixing component)

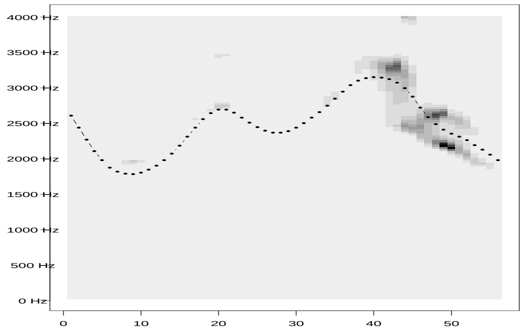


Figure 6. Iteration: 1 (overall fitting)

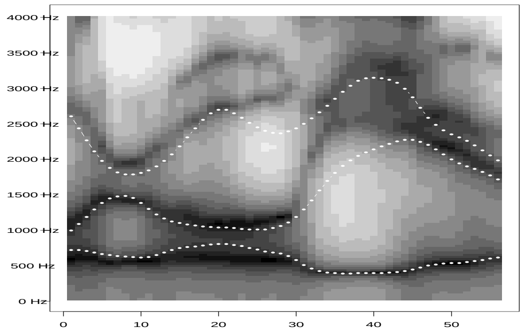
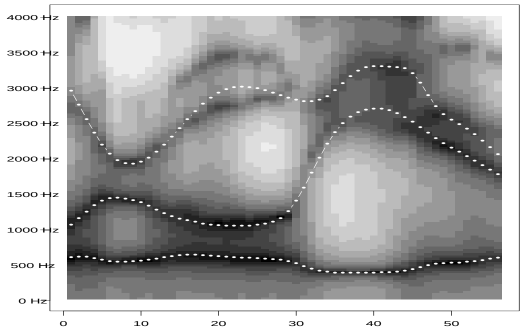


Figure 7. Iteration: 5 (overall fitting)



mixture models of multiple splines. The proposed method is easy to implement via an *EM*-type iterative algorithm. There is no need for dynamic programming and evaluation of complex objective functions. The main advantages of the proposed method are the robustness against missing and spurious LPC spectra peaks, the smoothness of the resulting estimates of the trajectories, the ease of implementation, and the flexibility in applications to other spectrogram-like speech representations, where "spectral peaks" may not be well defined.

Acknowledgements

This work was partially supported by the National Science Foundation Grant No. DMS-9057429, while the author was at the Department of Applied Mathematics and Statistics at the State University of New York at Stony Brook. Also, the author would like to thank Prof. Trevor Hastie at Stanford University for his stimulating discussions.

REFERENCES

[1] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):63–655, 1971.

[2] B. Gidas and A. Murua. Classification and clustering of stop consonants via nonparametric transformations and wavelets. In *ICASSP*, pages 872–875, Detroit, Michigan, 1995.

[3] H. Hermansky. Perceptual linear predictive analysis of speech. *Journal of the Acoustical Society of America*, 1990.

[4] M. J. Hunt. A robust formant-based speech spectrum comparison measure. In *Proceedings ICASSP*, pages 1117–1120, 1995.

[5] P. Lancaster and K. Salkauskas. *Curve and Surface Fitting: An Introduction.* Academic Press, 1986.

[6] H. Ney. A dynamic programming technique for nonlinear smoothing. In *Proceedings ICASSP*, pages 62–65, 1981.

[7] M. Phillips, J. Glass, and V. Zue. Automatic discovery of acoustic measurements for phonetic classification. In *Proc. Int'l Conf. Spoken Language Processing*, pages 795–798, 1992.

[8] B. G. Secrest and G. R. Doddington. An integrated pitch tracking algorithm for speech systems. In *Proceedings ICASSP*, pages 172–157, 1992.

[9] D. Talkin. Speech formant trajectory estimation using dynamic programming with modulated transition costs. Manuscript, Entropic Inc., 1993.

[10] G. Wahba. *Spline Models for Observational Data.* SIAM, Philadelphia, PA, 1990.

[11] V. Zue. The use of speech knowledge in automatic speech recognition. *Proc. of the IEEE*, 3:1602–1550, 1985.

[12] V. Zue. Notes on speech spectrogram reading. unpublished course notes, M.I.T., Cambridge, MA., 1991.