AUTOMATIC SLOPE MEASUREMENT ON FORMANT TRACKS

R. van Son

1. INTRODUCTION

In a research project on the spectral differences between utterances spoken with different speaking rates the need arose to measure the velocity of formant movements in running speech. A method for the automatic measuring of formant velocities is developed here.
With natural speech it is often very difficult to measure spectral changes in a speech signal. The spectrum of a constant or slowly changing signal can be determined almost to the theoretical limits. The measurements of a fastly changing signal, however, suffer from a lack of theoretical understanding and comprehensible representation. The central question is which changes are to be measured on a given set of spectra, measured on different points in time. Even the status of a spectrum, measured on a changing signal, is often unclear due to the implicit assumption of stationarity that underlies most spectral representations.
Interest in the spectral changes of speech signals is most often concentrated on the behaviour of spectral peaks. There are several ways to measure and represent spectral peaks.
One possibility is to transform the speech waveform into a spectrum, essentially making some type of time representation of bandpass filter outputs (this holds for normal Fourier Transforms). The problem is to identify peaks and follow their course in time and frequency. This is no trivial matter because it is difficult to decide what is a peak and what is not and which parts of the spectrum are instances of the same peak at different times.
Another possibility is to formulate a model of human speech production and measure the changes in the parameters of this model that affect the spectral contents of the speech signal. This last approach is followed in this paper with the use of Linear Predictive Coding (LPC). This LPC analysis can encode the spectral peaks of the speech signal in a fixed number of variable, second order, bandpass filters. The spectral parameters of interest are the centre frequency and the bandwidth of each spectral peak encoded this way. Here a method will be described for measuring spectral changes as used to study the course in time of the centre frequency. For this kind of study all spectral peaks have to be defined at all times. In normal LPC analysis, with the Levinson algorithm, sometimes a peak is "lost". To prevent this disturbance, a different algorithm is used here for the LPC analysis, the so called Splitt–Levinson algorithm. This algorithm was implemented by L. Willems (L.F.Willems, Robust formant analysis, IPO Annual Progress Report, 21, 1986, pp. 34–40).
Choosing an LPC representation has some advantages over pure spectral approaches. It is possible to manipulate all parameters of an LPC analysis and still resynthesize recognizable speech. Small changes in the parameters result in small changes in the resynthesized speech. In this way it is possible to test for clues for speech recognition or speech quality by changing the relevant parameters and, the other way around, to hear whether a change in parameters removes the quality of interest.

The spectral peaks that result from LPC analysis are often called formants. This is because the production model that forms the root of this approach, tries to model the effects of resonances in the speech organs. These resonances are, by definition, formants. The fit between the model and reality is, however, not good enough to ensure a perfect fit between the LPC spectral peaks and the formants. Sometimes there is a discrepancy between the measured peaks and the heard formants. Resynthesized speech however, mostly is of acceptable quality. In spite of the imperfect fit, these spectral peaks will be called formants hereafter.

2. MODELLING FORMANT TRACKS

If the object of measurement is to detemine the spectral change, i.e. the spectral velocity, then it is necessary to perform differentiations on the spectral data. Differentiation is an operation that is very sensitive to random measurement errors or noise. It amplifies those errors and noise in such a way that even small, local errors can completely corrupt velocity measurements. To deal with this phenomenon it is necessary to remove, at least part of, the noise from the data. To succesfully separate the desired signal and the noise, it is necessary to develop a model of the signal and the noise. If a good model for the speech production would have been available that could accurately describe the course of formants, the problem could be solved without major problems. But since such a model is not available yet, it is necessary to develop an accurate description of the signal without much reference to production.
It is very often possible to approximate a signal of unknown composition, a posteriori, to any desired accuracy by constructing a sum of standard functions. The remaining discrepancy between data and description is treated as noise and removed, only the modelled part is kept. It is important to choose the right class of functions to model the signal. An inappropriate model function will lead to a disturbed signal. Functions that can be made orthogonal should be prefered.
Choosing functions for modelling is always a guess. The guess made here is that an LPC formant track, f(t), on a given interval $[t_0,t_1>$ can be modelled a posteriori with any desired precision with a polynomial function that has the form:

$$\begin{aligned} f(t) &= a_0 + a_1 \cdot t + a_2 \cdot t^2 + ... \\ &= \sum_{j=0}^{\infty} a_j \cdot t^j \\ &= H^{\infty}(t) \end{aligned} \qquad (1)$$

with: $t \in [t_0,t_1>$

For any maximal given order J of the polynomials

$$H^J(t) = \sum_{j=0}^{J} a_j \cdot t^j \qquad (2)$$

is chosen such that it is the best approximation of f(t) for this order of polynomials on this interval. A better way to define $H^J(t)$ is to use a set of orthogonal polynomials like the (shifted) Legendre polynomials, especially if $J \geq 2$. Using a set of orthogonal functions has great methodological and computational advantages. A short description of shifted Legendre polynomials is given in Appendix A.

After the calculations of $H^J(t)$ the original formant track is replaced by

$$f(t) = H^J(t) + \varepsilon(t) \tag{3}$$

in which $\varepsilon(t)$ is an error term. For high orders of J it will be difficult to determine the best intervals $[t_i,t_{i+1}>$ to fit $H^J(t)$ on f(t). The order of the model function should therefore be as low as possible. For measuring formant slopes (=velocity) an order of 1 will do, for measuring formant acceleration an order of 2 is necessary. In the discussion below an order of 1 will suffice, the order indication of the model functions $H^1(t)$ will be omitted hereafter.
For this first order polynomial model to make a good fit it is important to choose appropriate intervals. The formant track is modelled as a succession of simple straight line segments. If the boundaries between succesive line segments are chosen wrong, the resulting modelled track will have hardly any resemblance to the originally measured formant track. In this model therefore the original formant track f(t) is divided in intervals $T_i = [t_i,t_{i+1}>$ that do not overlap. In every interval T_i the formant is modelled with:

$$f(t) = H_i(t) + \varepsilon(t) = a_i \cdot t + b_i + \varepsilon_i(t) \tag{4}$$

$t \in T_i = [t_i,t_{i+1}>$
$H_i(t)$: a straight line on T_i
$\varepsilon_i(t)$: the error term on T_i, defined by $\varepsilon_i(t) = f(t) - H_i(t)$

Next $\varepsilon_i(t)$ can be modelled by a Gaussian distributed noise term $e_i(t)$ with expected value $E(e_i(t)) = 0$ and variance $E(e_i(t)^2) = \sigma_i^2$. $H_i(t)$ becomes the straight line that minimizes σ_i^2. In this model the value of the formant at time $t \in T_i$ is $H_i(t)$ and the slope is a_i.
The assumption that $\varepsilon_i(t)$ can be modelled by a Gaussian distributed noise term is made for convenience. It is possible to use other distributions but calculating the best fit becomes time consuming and for the simple example described here there is no point in using any other distribution. The minimizing criterion for the best fitting function can be altered to emphasize the errors in special parts of the interval, e.g. the centre of the interval, by using a weighting function.

The preceding argument can be summarized as follows:
With LPC analysis it is possible to extract formant frequencies from a speech signal. These formants form tracks in time. Each of these tracks, represented by the function f(t), can be modelled by dividing the track in non–overlapping intervals T_i and replacing the measured track f(t) with:

$$f(t) \approx H_i(t) + e(t) = a_i \cdot t + b_i + e_i(t) \tag{5}$$

$t \in T_i = [t_i,t_{i+1}>$
$H_i(t)$: a straight line on T_i
$e_i(t)$: a Gaussian noise term on T_i, defined by
$E(e_i(t)) = 0$
$E(e_i(t)^2) = \sigma_i^2$ (i.e. independent of t)

In equation 5 the best guess for $H_i(t)$ is the linear regression line on T_i.

3. SEGMENTATION

In the preceding model, segmenting the tracks in independent intervals is crucial for a good fit of the model on the tracks. Such intervals are called line segments here. A line segment is defined as the largest interval in which the formant track can be modelled by a straight line according to equation 5. The segmentation can be done in an automatic way if there is a smallest interval length τ for which there is no smaller line segment. If there is such a smallest length of a line segment, then it is possible to find all the boundaries between the segments. This is done by deciding whether a test segment of the track (called Δ_0) with a length smaller than or equal to the smallest interval length (i.e. $|\Delta_0| \leq \tau$) contains a boundary between line segments. If it is concluded that the test segment does contain a boundary between line segments, then the best point to place this boundary can be found. This test segment is shifted over the track until all possible boundaries are found.
The decision whether or not the test segment contains a boundary between line segments is made by trying to find a subdivision of Δ_0 in two subsegments Δ_1 and Δ_2 that have a lower expected value for the remaining variance of their regression lines (called $E(v_1^2)$ and $E(v_2^2)$) than the undivided test segment (called $E(v_0^2)$). If there is no boundary present in Δ_0, that is, Δ_0 is completely confined in a segment (T_i) of the track with only one straight line segment, then all subdivisions of Δ_0 will have the same expected values for the remaining variance of there regression lines as Δ_0 itself. Or, for all subdivisions Δ_1 and Δ_2 of Δ_0 lying in segment T_i :

$$E(v_0^2) = E(v_1^2) = E(v_2^2) = \sigma_i^2 \tag{6}$$

with:
$E(v_0^2)$, $E(v_1^2)$, $E(v_2^2)$: the expected values of the remaining variance of the regression lines in the segments Δ_0, Δ_1 and Δ_2
v_0^2, v_1^2 and v_2^2: the estimated or calculated values of the remaining variance of the regression lines in the segments Δ_0, Δ_1 and Δ_2
σ_i^2 : the variance of the model noise term in segment T_i (cf. equation 5)

If, however, the test segment Δ_0 contains a boundary between two segments, T_i and T_{i+1}, with different model lines (not only different noise terms), then there exists at least one subdivision of Δ_0 in two segments Δ_1 and Δ_2 that has a lower expected value of the remaining variance than the test segment itself. Or

$$|\Delta_0| \cdot E(v_0^2) > |\Delta_1| \cdot E(v_1^2) + |\Delta_2| \cdot E(v_2^2) \tag{7}$$

with: $|\Delta_0| = |\Delta_1| + |\Delta_2|$ the lengths of the segments

The subdivision with the lowest remaining variance, $|\Delta_1| \cdot E(v_1^2) + |\Delta_2| \cdot E(v_2^2)$, has expected values of the remaining variance of the regression lines that are equal to the variances of the noise terms in T_i and T_{i+1}. That is:

$$E(v_1^2) = \sigma_i^2 \qquad E(v_2^2) = \sigma_{i+1}^2 \tag{8}$$

and

$$|\Delta_1| \cdot E(v_1^2) + |\Delta_2| \cdot E(v_2^2) = |\Delta_1| \cdot \sigma_i^2 + |\Delta_2| \cdot \sigma_{i+1}^2$$

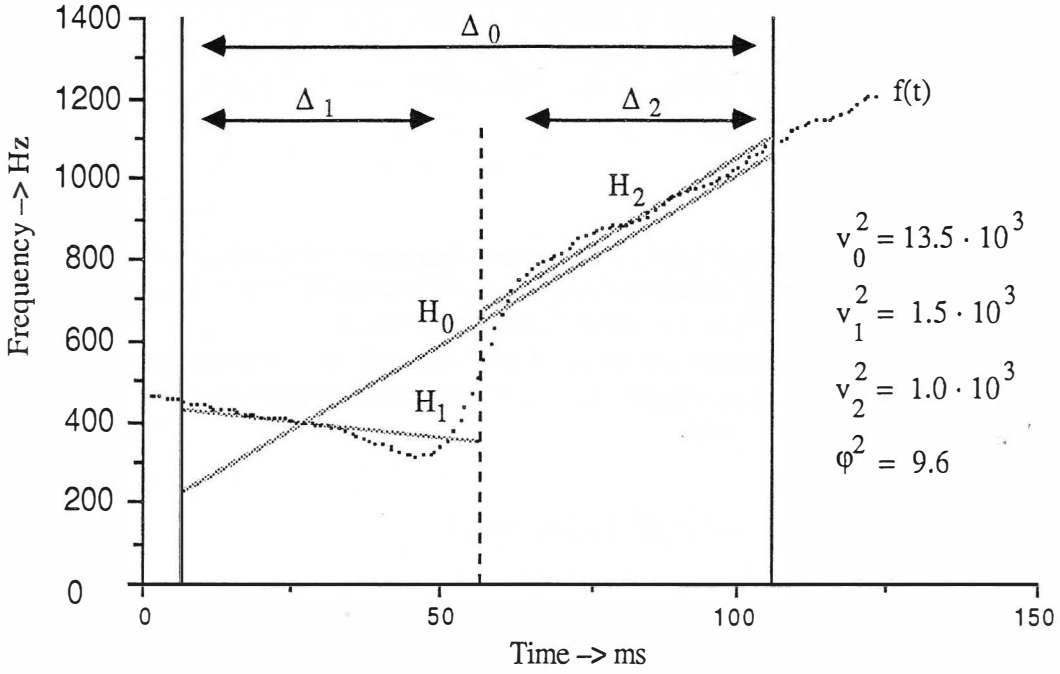

Fig. 1: An example of a formant track f(t) and the calculated values of the parameters of equation 10 on a test segment Δ_0 (see text for explanation). The division used is indicated by a dashed line and is the one with the highest value of φ^2.
Δ_0 is a test segment with $n_1+n_2=100$ points. Δ_1 and Δ_2 are the two neighbouring sub–segments of Δ_0, each containing $n_1=n_2=50$ points.
H_0, H_1 and H_2 are the regression lines on these three segments.
It can be seen that the test segment, Δ_0, is chosen too large. Three line segments are actual present inside the test segment Δ_0, which results in a total of two boundaries. But inside a test segment only one boundary between line segments can be found with the method described here. As is shown in this figure.

These equations are valid with a continuous formant track and an unlimited number of realisations of the SAME formant track. If only one realisation of the track is available and only a limited number of measuring points in a segment are known then it is necessary to work with the estimated values alone. Equation 7 will become:

$$n_0 \cdot v_0^2 > n_1 \cdot v_1^2 + n_2 \cdot v_2^2 \qquad (9)$$

with: $n_0 = n_1 + n_2$ the number of measured points in the segments Δ_0, Δ_1 and Δ_2

If a subdivision is found for which this inequality holds, then there is a segment boundary in Δ_0. The best guess for the position of this boundary is the point that separates the subsegments Δ_1 and Δ_2 with the lowest value of $n_1 \cdot v_1^2 + n_2 \cdot v_2^2$. If this value is not equal to zero then take this subdivision and rewrite equation 9 to:

$$\varphi^2 = \{ (n_1+n_2) \cdot v_0^2 - (n_1 \cdot v_1^2 + n_2 \cdot v_2^2) \} / \{ n_1 \cdot v_1^2 + n_2 \cdot v_2^2 \} > 0 \qquad (10)$$

φ^2 Is the largest value possible for the quotient on this test segment (see fig. 1). If both sides of equation 9 are equal to zero, there is no boundary in the test segment. If only the right hand side of equation 9 is equal to zero, then there is a boundary in the test segment. Because of definition and the fact that v_0^2 is calculated from the same points

as $n_1 \cdot v_1^2 + n_2 \cdot v_2^2$ is calculated, φ cannot be smaller than zero. It is however easily seen that $\varphi > 0$ is possible with no boundary present.This erroneous boundary detection results from stochastic errors in the estimators v_0^2 , v_1^2 and v_2^2. For this reason equation 10 should be changed to:

$$\varphi^2 > \delta^2 \tag{11}$$

for detection of a boundary. δ Is a dimensionless number which gives a threshold for detection in numbers of standard deviations difference between $(n_1+n_2) \cdot v_0^2$ and the smallest possible $n_1 \cdot v_1^2 + n_2 \cdot v_2^2$ value in the test segment.
Because there are different numbers of points involved in calculating the different estimated variances, it is important to use unbiased estimators. Here the following unbiased estimators are used:

$$v_0^2 = \{ \sum_{i=1}^{n1+n2} (f(t_i) - H_0(t_i))^2 \} / \{n_1+n_2-2\} \tag{12}$$

and

$$v_{12}^2 = \{ \sum_{i=1}^{n1} (f(t_i) - H_1(t_i))^2 + \sum_{j=1}^{n2} (f(t_j) - H_2(t_j))^2 \} / \{ n_1+n_2-4\}$$

with: $v_{12}^2 = \{ n_1 \cdot v_1^2 + n_2 \cdot v_2^2 \} / \{ n_1+n_2 \}$
$t_i \in \Delta_1$
$t_j \in \Delta_2$
$H_0(t)$, $H_1(t)$ and $H_2(t)$ the regression lines in the segments Δ_0, Δ_1 and Δ_2

In this notation φ^2 will become:

$$\varphi^2 = \{ v_0^2 - v_{12}^2 \} / v_{12}^2 > \delta^2 \tag{13}$$

for boundary detection.
Two assumptions are critical to the fit of the model track on the formant track. First there is no more than one segment boundary in any part of the track with a length $\leq \tau$, with τ being defined as some minimal length greater than or equal to the length of the test segment. Second the formant tracks consist of straight line segments with additive Gaussian noise. If the first assumption does not hold and a test segment contains two or more segment boundaries, then the behaviour of φ^2 will become dependent on where the boundaries are inside the test segment. The detection and assignment of boundaries between line segments becomes very irregular. If the second assumption does not hold and the formant tracks are curved, then boundaries will be placed in such a way that the regression lines will fit the curve with more or less constant variance.

In an actual implementation of the described boundary detector one shifts the test segment one point at a time and accepts only subdivisions with lowest v_{12}^2 which divide the test segment in two parts of equal length. This secures the use of the most accurate estimation of v_{12}^2 for boundary detection. Every boundary is shifted in the centre of the test segment only once and so can be detected only once.
To calculate two regression lines in a test segment, this segment must contain at least 6 points (three points for each regression line). This constraint determines the minimal time resolution needed for the formant measurements.

4. SEGMENTATION OF SEVERAL TRACKS SIMULTANEOUSLY

If more than one formant track is used simultaneously to detect synchronous segment boundaries, a total v_0^2 and a total v_{12}^2 are calculated by summing the individual v_0^2 values for all tracks and by summing the individual v_{12}^2 values for every subdivision of the test segment for all tracks. The equation (13) for boundary detection will not change, but total values will be used for the estimated variances instead of individual values. This is the equivalent of treating the frequency values of different tracks as independent dimensions and stating that each segment contains a multidimensional straight line.

5. OTHER PARAMETERS FOR DETECTING BOUNDARIES

The method to detect boundaries in formant tracks described above is purely statistical. It is possible to use other clues to find those segment boundaries. For example, a change in the voicing of speech (voiced to unvoiced or the reverse) signifies an important change in speech that is likely to have an important effect on formant tracks. It is also possible to use threshold values for the energy of the speech signal or other threshold values to find important changes in the signal. Of all possible parameters that could be used to detect segment boundaries, only the voicing transition is currently implemented, complementary to the formant tracks themselves, of course.

6. COMPARING STRAIGHT LINES

After the segmentation, the formant track is divided into a large number of segments. The regression lines of many of these segments will not differ markedly from that of their neighbours. It is possible to remove a significant number of those segment boundaries and merge segments by comparing the regression lines of neighbouring segments.

Comparing straight lines is done by calculating a distance between lines in a shared interval. The distance of the straight lines in two neighbouring segments T_i and T_{i+1} is defined here as the Root Mean Square difference between the two lines in the total interval $(T_i \cup T_{i+1})$. The difference between the lines is measured perpendicular to some standard line. This standard line can be the time axis, a regression line through the combined interval, the bisector line that divides the arc between the lines evenly in two, or it can be some other line. Using the bisector line as the standard line for distance measurement results in the shortest distance between lines and is currently implemented (see appendix B for the actual calculation of the distance).

The mean line distance, defined as above, depends on the total interval length and tends to infinite large values if the interval length becomes infinite. So this distance is not a quality of the two lines but of the two lines in an interval and depends on the interval. Long intervals must resemble each other more than short intervals in order to be merged into one interval. This distance can be calculated over several formants simultaneously by treating each formant as an independent dimension and the lines as multidimensional straight lines. The total squared distance is calculated by summing the individual squared distances.

Using the line distance to remove unwanted segment boundaries gives the opportunity to segment with high sensitivity and to remove excess boundaries afterwards. This is important because while the segmentation stage has only a narrow, local, scope, the

comparing stage has a scope that can be infinite in principle. A local scope is noise sensitive and error prone.

7. CONCLUSIONS

An implementation of the theory discribed above was made on a μVAX II mini–computer. Some minor changes were introduced. First, the condition that there should be no more than ONE segment boundary in the test segment was relaxed. Instead of this strict condition, only a minimal distance between segment borders was demanded. This proved to work well. Second, it appeared that the condition of dividing the test segment into two equal sized sub–segments to signal a segment boundary sufficed to select only few excess boundaries. There was no need for an additional threshold for boundary detection (δ^2 in equation 11). When a minimal RMS line distance is used to decide whether a boundary separates distinct parts of the formant track, then it is possible to eliminate these excess segment boundaries as well as some others that do not separate distinct parts of the formant track.
An example formant track was segmented and modelled with this program. The results are displayed in fig. 2a and 2b. The modelled track of fig. 2b is used to measure the slope of the original track.

ACKNOWLEDGMENT

This project was supported by the Foundation for Speech Technology, which is funded by the Dutch National Program for the Advancement of Information Technology (SPIN).

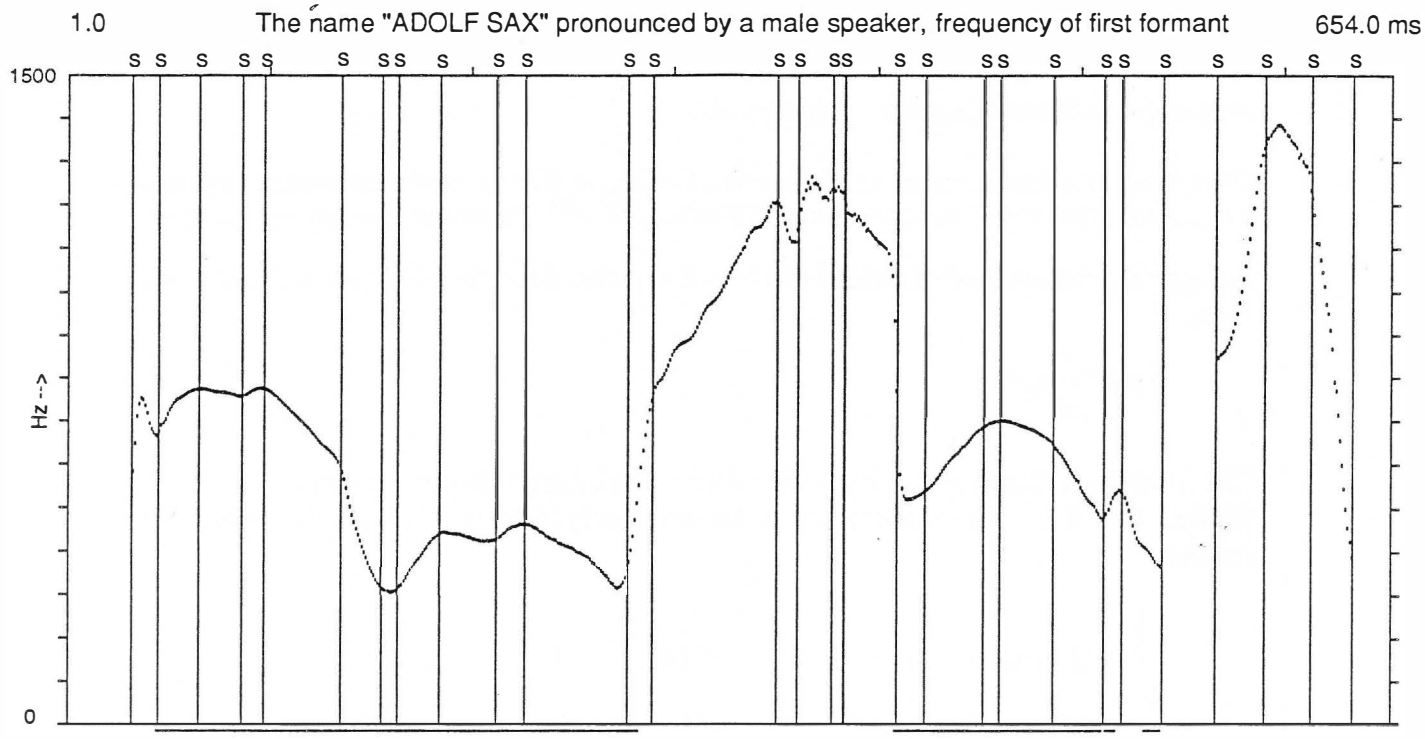

Fig. 2a: The results of a segmentation of the track of the first formant, voicing is indicated by a black line below the frame. Testinterval: 20 points, minimal distance between borders: 4 points, threshold for detection of a border: none, borders on voicing change, minimal RMS line distance: 30 Hz.

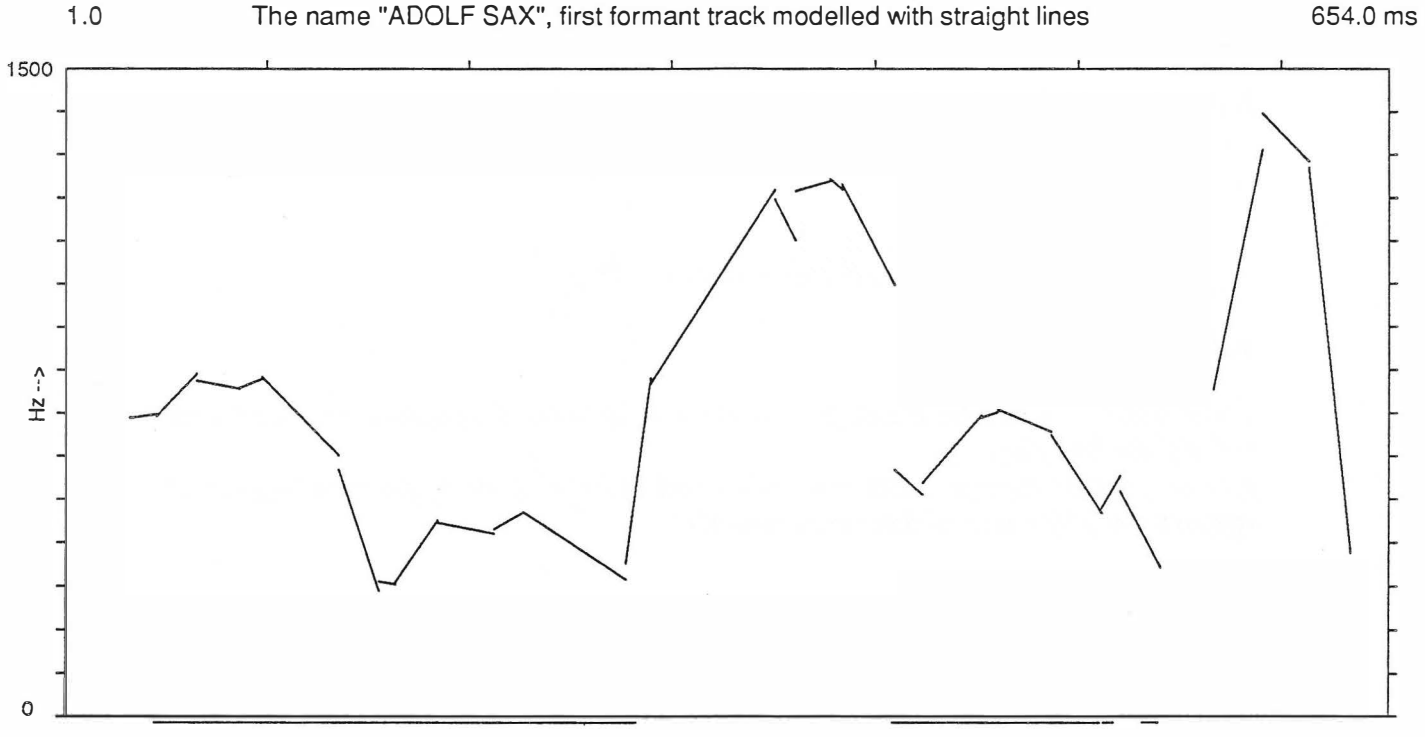

Fig. 2b: The above formant track modelled with straight lines. Remaining standard deviation: 22 Hz.

Appendix A: Shifted Legendre polynomials

(This appendix is adapted from: M.Abramowitz, I.A.Stegun, Handbook of mathematical functions, Dover publications 1965[9], National Bureau of Standards 1964[10], The section on orthogonal functions)

A Legendre polynomial of order J is a function defined for $t \in [-1,1]$ or $t \in [0,1]$ of the form:

$$L_J(t)=\sum_{j=0}^{J} a_{Jj}\cdot t^j$$

The functions defined on $t \in [0,1]$ are called shifted Legendre polynomials.
Shifted Legendre polynomials are orthogonal polynomials. That is, they obey the relation:

$$\int_0^1 L_I(t)\cdot L_J(t)\,dt = \begin{cases} 0 & \text{if } I\neq J \\ h_J \neq 0 & \text{if } I=J \end{cases}$$

and for the Shifted Legendre polynomials: $h_J = 1/\{2\cdot J+1\}$

The first four polynomial functions are:

$$L_0(t) = 1$$
$$L_1(t) = 2\cdot t - 1$$
$$L_2(t) = 6\cdot t^2 - 6\cdot t + 1$$
$$L_3(t) = (40\cdot t^3 - 60\cdot t^2 + 24\cdot t - 2)/2$$

If the interval is $t \in [0,k]$ then the first four funtions change into:

$$L_0(t) = 1$$
$$L_1(t) = 2\cdot t / k - 1$$
$$L_2(t) = 6\cdot t^2 / k^2 - 6\cdot t / k + 1$$
$$L_3(t) = (40\cdot t^3 / k^3 - 60\cdot t^2 / k^2 + 24\cdot t / k - 2)/2$$

and $h_J = k / \{2\cdot J+1\}$

These functions can be translated to another interval, $t' \in [k_1,k_2]$, by substituting $t=t'-k_1$ and $k=k_2-k_1$.
Any continuous function, f(t), that exists and is finite in every point of [0,k] can be approximated by a sum of these polynomials

$$f(t) = \sum_{j=0}^{\infty} A_j\cdot L_j(t)$$

Because of orthogonality it is possible to calculate the factors A_j independent of one another with the following relation:

$$A_j = [\int_0^k f(t) \cdot L_j(t)\, dt] / h_j$$

With this relation it is possible to calculate the factors A_j in a very efficient way.

Any straight line on the interval [0,k] can be written as:

$$f(t) = a \cdot t + b$$
$$= A_0 + A_1 \cdot L_1(t)$$

$t \in [0,k]$
and:
$A_0 = b + a \cdot k / 2$
$A_1 = a \cdot k / 2$

Appendix B: Calculation of line distance

Define two straight lines:

$g(t) = a \cdot t + b$
$h(t) = c \cdot t + d$

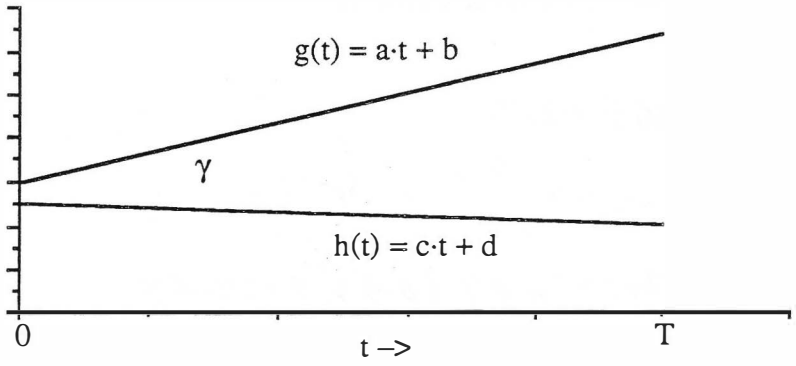

Fig. 2: Two straight lines, g(t) and h(t), with an angle of γ inbetween.

The distance between these two lines is defined here in the interval [0,T >. Any other interval can be transformed to this interval easily. The distance is defined perpendicular to the bisector line. The bisector line between g(t) and h(t), i.e. the line that divides the angle γ into two equal halves, is calculated as follows.

Define the bisector line as:

$$b(t) = e \cdot t + f$$

The angle between g(t) and h(t) is called γ and is:

$$\gamma = \text{arctangent}\ (\ (a-c)\ /\ (1+a\cdot c)\ )$$

define:

$$\Gamma = \text{tangent}\ (\ \gamma / 2\ )$$

then the parameters of the bisector line become:

$$e = (\ c+\Gamma\ )\ /\ (1-c\cdot\Gamma)$$
$$f = \{\ (1+c\cdot e)\cdot b + (1+a\cdot e)\cdot d\ \}\ /\ \{\ 2 + (c+a)\cdot e\ \}$$

To calculate the distance perpendicular to b(t) all lines are rotated and translated such that b(t) lies on the horizontal axis. In this reference frame the new lines g'(t) and h'(t) are:

$$g'(t) = a'\cdot t + b'$$
$$h'(t) = c'\cdot t + d'$$

and
$$a' = \{\ a-e\ \}\ /\ \{\ 1+a\cdot e\ \}$$
$$b' = \{\ b-f\ \}\cdot\sqrt{[\ \{\ a'^2+1\ \}\ /\ \{\ a^2+1\ \}\ ]}$$

$$c' = \{\ c-e\ \}\ /\ \{\ 1+c\cdot e\ \}$$
$$d' = \{\ d-f\ \}\cdot\sqrt{[\ \{\ c'^2+1\ \}\ /\ \{\ c^2+1\ \}\ ]}$$

The distance D is defined in this reference frame as:

$$D^2 = [\ \int_0^T \{\ g'(t)-h'(t)\ \}^2\ dt\ ]\ /\ T$$

This can be simplified to:

$$D^2 = (\ a'-c'\ )^2\cdot T^2\ /\ 3 + (\ a'-c'\ )\cdot(\ b'-d'\ )\cdot T + (\ b'-d'\ )^2$$

The mean distance is D.