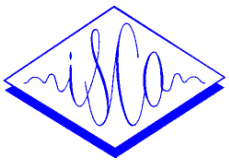

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Active Models for Regularizing Formant Trajectories

Yves Laprie and Marie-Odile Berger
CRIN CNRS & INRIA Lorraine
B.P. 239. 54506 Vandoeuvre-les-Nancy. France

10.21437/ICSLP.1992-264

Abstract

Formant tracking is one of the most important issues in speech recognition. Nevertheless. most algorithms are based on local methods and take into account global formant properties with difficulty. We therefore propose a formant tracking algorithm which provides a global point of view on formant tracking. The salient idea is to combine local tracking to generate elementary formant tracking hypotheses and *an active method* to regularize global formant trajectories in the following way: the formant trajectory is the closest curve to that of the hypothesis maximizing energy incorporated by the formant and which is sufficiently smooth. The main qualities of this algorithm are robustness and ability to detect accurate and regular formant trajectories.

1 Introduction

The success of global approaches in automatic speech recognition. especially Hidden Markov Models. is explained by the dearth of reliable and robust automatic acoustic attribute detectors. It is thus much easier to feed spectral vectors rather than acoustic attributes into a speech recognition system.

Although formant trajectories probably constitute the most important acoustic attribute. there is not formant tracker that can determine formant frequencies reliably. especially in regions where formant transitions provide important information about the place of articulation for consonants [6].

Formant tracking is an arduous task because resonance frequencies are not directly observed. Actually. source harmonics whose intensity has been increased due to their proximity to a resonance frequency are monitored.

Signal processing techniques allow only imperfect separation of the vocal tract and source contribution. Furthermore the higher the pitch frequency and the faster it changes. the greater the difficulty of locating peaks of the vocal tract filter function. Hence. formant trajectories appear as discontinuous and erratic peak tracks on LPC or cepstrally smoothed spectrograms although vocal tract evolution. and by implication formant evolution is continuous. The formant tracking algorithm must thus label tracks in terms of formants as well as possible and deal with the following problems:

- a track may be spurious and does not fit any formant.
- a track may represent two merged formants which have not been separated by signal processing.
- a formant may not have been detected either because it is too weak or because the speech signal is noisy.

The classical formant tracking algorithms relying on a local point of view are doomed to failure as soon as a formant trajectory becomes noisy or difficult to track (because two formants are too close together to be distinguished. voicing is not regular enough. . .).

In order to tackle the above-mentioned problems we propose a global approach which consists in tracking formant trajectories as curves on the whole tracking duration and which permits overriding of local spectrogram irregularities.

More precisely we first define a local assessment criterion which ensures that peaks fitting F1. F2 and F3 at every instant are realistic from an acoustic point of view (section 2). This criterion operates on F1. F2. F3 simultaneously and yields a result all the better when F1. F2. F3 incorporates more energy. F1-F2 and F2-F3 frequency constraints are satisfied and F1. F2. F3 levels are consistent with acoustic theory.

We then search formant trajectories for the whole duration of tracking as curves sufficiently smooth and satisfying the local assessment criterion as well as possible at every instant. Our algorithm is made up of four steps:

1. **local processing** Elementary tracks are built by local tracking on cepstrally smoothed spectra and are lines of peaks which may fit pieces of formant trajectories. They are built by a simple algorithm assuming that for a small time interval the frequency of a formant evolves little:

2. **elementary track labelling** Each elementary track is then labelled in terms of formant trajectories (F1. F2. F3) and contributes to a total rough labelling (total. because the labelling is for F1. F2 and F3 together and for the whole speech segment studied. rough because accurate location and formant label are not well defined) of the elementary track set (Fig. 1.b). Actually. in order to take into account possible errors the most likely total rough labellings are generated. Every rough labelling consists of the set of tracks labelled as F1. F2 or F3 and results from a labelling algorithm which can make the following hypotheses to take possible errors into account: this track is spurious. there is no track for this formant. this track represents two merged formants. The track labelling is boostraped by a peak-formant assignment procedure called throughout of the tracking.

3. **generation of initial formant trajectories** Elementary tracks representing the same formant are connected end to end to form an initial formant trajectory. The curve thus built is a rough approximation of the formant trajectory be-

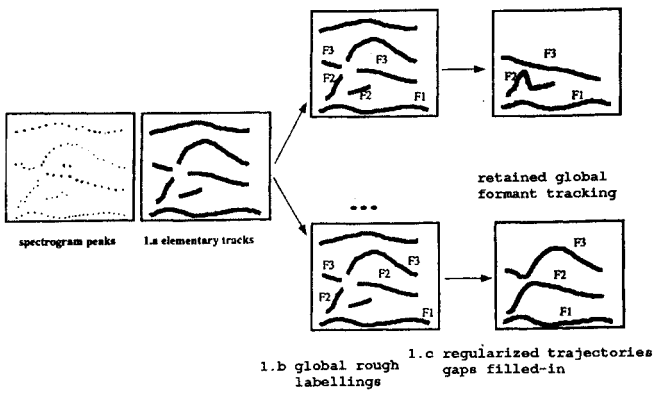

Figure 1: algorithm layout

cause this curve contains the most significant points of the trajectory and is highly approximative between two consecutive elementary tracks, which moreover may be jagged.

4. **global formant tracking by an active method** By using the *active contour model* [3] the initial formant trajectories move under the action of spectrogram energy (section 3) towards the nearest formant trajectory. The three regularized formant trajectories form a global tracking solution, one for every total rough labelling. The one which gives rise to the highest assessment criterion value is the most likely to represent the real formant trajectories.

Steps 3 and 4 are the key steps of the algorithm, because the fact that formant trajectories may include gaps, renders the calculation of the assessment criterion simultaneously for F1, F2 and F3 impossible. Let us consider the example of a formant trajectory (suppose that is F2) which has not been detected between t_1 and t_2. The quality criterion is thus difficult to calculate and the quality of this global formant tracking solution (F1, F2 and F3 trajectories together) is not readily comparable with the quality of other formant tracking solutions. Step 4 allows a regularized and complete (especially during the gaps) formant trajectory to be generated according to the spectrogram energy. That permits an accurate calculation of the assessment criterion for the whole tracking duration. Furthermore step 4 regularizes the formant trajectories and removes irregularities of elementary tracks.

2 The assessment criterion

An assessment criterion is necessary to evaluate the quality of the formant trajectories $(trajF1, trajF2, trajF3)$ built by the formant tracker between t_i and t_f. We have thus defined an instantaneous local criterion a which allows the global quality Q of trajectories to be evaluated by:

$$Q(trajF1, trajF2, trajF3) = \sum_{t=t_i}^{t=t_f} a(trajF1(t), trajF2(t), trajF3(t))$$

There are general sufficient easy-to-use constraints which can be directly used in the assessment criterion, this is the case for energy and frequency constraints. Conversely, formant level information has not been included in the assessment criterion but in the strategy of peak-formant assignment.

Thus, the criterion involves the following two aspects:

Energy Let us consider combinations of three peaks, each peak representing one formant. The combination which incorporates the most energy is the most likely to represent F1, F2 and F3.

Frequency constraints Frequencies of F1 (resp. F2) and F2 (resp. F3) are not independent. Fig. 2 shows the frequency domains that F1 and F2 (resp. F2 and F3) may belong to. These domains are used to define constraints placed upon formants. The level of satisfaction equals 1 if the point representing F1 and F2 (resp. F2 and F3) is within the F1-F2 (resp. F2-F3) domain and continuously decreases towards 0 when the point moves away from the F1-F2 (resp. F2-F3) domain.

Let us consider e_1, e_2 and e_3; the energies of peaks fitting F1, F2, F3, s_{12} and s_{23}; the satisfaction levels of frequency constraints, the criterion value is:

$$a = e_1 s_{12} + e_2(s_{12} + s_{23})/2 + e_3 s_{23}$$

Taking into account the level of formants [2] could have given rise to one more factor in the assessment criterion. Theoretically, knowledge of frequencies and bandwidths of the three first formants allows the calculation of their levels in the spectrum. Levels thus calculated and levels of peaks associated to F1, F2 and F3 could be compared to assess the peak-formant assignment built by the formant tracker. This assessment scheme comes up against the following two problems:

- The frequency response of the microphone used to record speech is generally not known and is often not perfectly flat anyway. Differences of levels less than 5 dB are thus not significant.
- Formant bandwidths cannot be assumed constant (they vary from 40 Hz to 200 Hz for F3) [5] and cannot be readily extracted from spectra, even by LPC.

The formant levels measured by Pols [4] may be used to assess the peak-formant assignment. The experiments we have performed, have brought to the fore that formant levels can be used only to prevent the algorithm from confusing a back vowel whose F1 and F2 are close together with a front vowel. Actually, frequencies of F3 for a back vowel and F2 for a front vowel are both in [2,000 Hz, 3,000 Hz] but the F3 level is about 15 dB below the F2 level. That is not used in the assessment criterion but in the peak-formant assignment strategy.

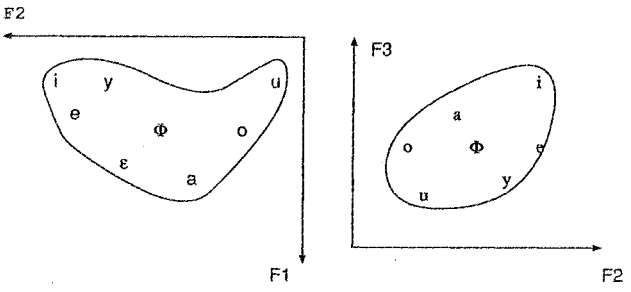

Figure 2: Frequency domains of F1-F2 and F2-F3

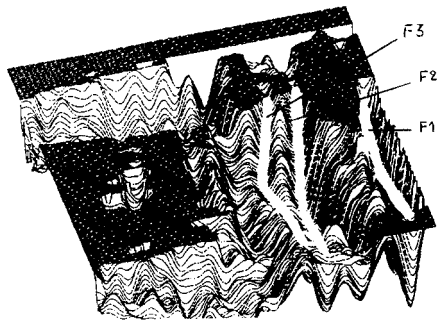

Figure 3: opposite of the spectrogram viewed as a potential field and formant trajectories

3 Global formant tracking by an active contour method

The initial formant trajectories (the elementary tracks labelled with the same formant connected in chronological order by filling in the spaces with straight lines) are a rough approximation of formant trajectories. These curves are called initialization curves in this section.

We have therefore to find the curve closest to the initialization curve which maximizes energy incorporated by formants and which is smooth enough. In order to tackle this problem, we use the *active contour* concept introduced in [3]: the required trajectory is a curve $s : [0,1] \rightarrow I\!R^2,\ s \rightarrow v(s) = (t(s), F(s))$ lying in the time $\times$ frequence space which minimizes a global energy E simultaneously incorporating constraints on the energy level and on the smoothness of the track:

$$\begin{aligned} E &= E_{Formant} + \lambda E_{smooth} \\ &= -\int_0^1 E_{spectro}(v(s))ds + \lambda \int_0^1 \alpha|v'|^2 + \beta|v''|^2 ds \end{aligned} \quad (1)$$

The first term $E_{Formant}$ is all the smaller when the energy on the curve is great and the second term E_{smooth} is usually used in spline theory and is all the smaller when the curve is regular (α influences the curve length whereas β influences its curvature). The parameter λ controls the compromise between the degree of regularization and its closeness to the elementary tracks.

We can easily give a physical interpretation to this energy: let us consider the opposite of the cepstrally smoothed spectrogram as a potential field (Fig. 3); the potential valleys thus represent the formants. A curve placed in the vicinity of a minimum is thus attracted to the nearest potential valley, i.e. the nearest formant trajectory;

Equation (1) is solved variationally using the Euler equations namely the first derivative of E:

$$-\alpha v'' + \beta v^{iv} - \frac{\partial E_{spectro}(v(s)}{\partial v} = 0 \quad (2)$$

Starting from the initialization, the curve is deformed and moves until the nearest minimum of E, i.e. the nearest formant, has been reached; the formant model we use is hence an *active model.*

An iterative scheme is used to solve (2) from the initialization. The curve is discretized and is represented by a set of equidistant points $(v_i = (t_i, F_i))_{(0<i<N)}$. Using traditional discretization of first and second derivatives

$$v_i' = v_i - v_{i-1} \ , v_i'' = v_{i+1} - 2v_i + v_{i-1},$$

equation (1) can be written in matrix form as:

$$Av = \frac{\partial E_{spectro}(v(s))}{\partial v},$$

where A is a pentadiagonal matrix depending on the boundary conditions imposed on the curve to ensure the problem has a unique solution. Since we expect a track in the same time interval as the initial hypothesis, we therefore impose $t(0) = t_i$ and $t(1) = t_f$ (where $v_0 = (t_0, F_0)$ is the parametrization of the initialization) and only regularizing boundary conditions are imposed on F: $F''(0) = F'''(0) = F''(1) = F'''(1) = 0$. More explanations can be found in [1].

This method is a powerful tool to find the curves incorporating as much energy as possible from a rough initialization. Note that this algorithm does not consist in a simple smoothing of the initialization curve; the method is above all an efficient way to simultaneously perform

- dynamic evolution of the track which is attracted to lines of the spectrogram on which the energy is maximal (formants) from a rough and incomplete initial formant trajectory
- track smoothing

During analysis of a speech segment limited by t_i and t_f, one of the initial formant trajectories may be defined only on a subinterval of $[t_i, t_f]$ but the trajectory can be extended in the following way [1]: the trajectory is slightly extended in the direction of its tangent; the curve thus formed is used as an initialization curve which is regularized and optimized according to the spectrogram energy by the active method. This process may be iterated if necessary.

Once the formant trajectories have been completed and regularized the assessment criterion can be calculated simultaneously for F1, F2 and F3 for the whole tracking duration; that allows the best global tracking solution to be retained.

4 Results and concluding remarks

The experiments we have carried out have confirmed that taking into account the global aspect of formant tracking through an active contour method improves the results appreciably.

Fig. 4 shows a fine result for a poorly defined F3. The result giving the highest assessment value is as expected. Note that there are two close sets of initial formant trajectories which lead to similar and correct results. There is still a slight "bump" at the beginning of the F3 trajectory which seems to be inevitable considering the spectrogram, except by increasing the smoothness constraint, i.e. by increasing β.

Fig. 5 demonstrates one of the major improvements provided by our approach, that is the reliability of formant transitions (especially for F2 and F3). That should help acoustic attribute based speech recognition systems to improve identification of place of articulation for consonants.

The fact that F1 and F2 are merged in the middle of /too/ is due to the fact that there is a sole peak for F1 and F2 on the smoothed spectrogram. Actually, the active method is almost stabilized in Fig. 5.d, but it finally merges F1 and F2. Our future work will concern the incorporation of acoustic constraints between formants in the active method in order to track simultaneously F1, F2 and F3.

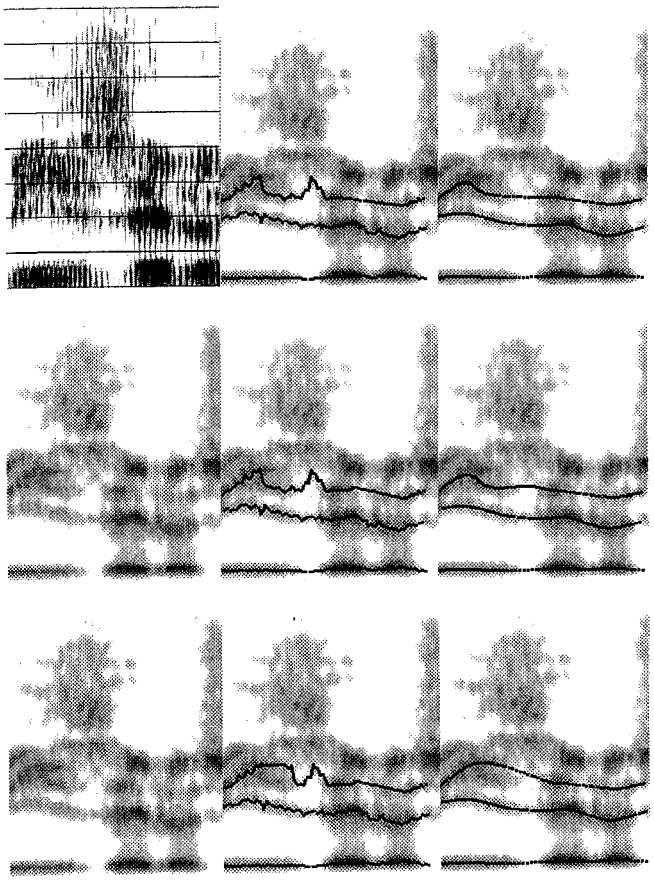

Figure 4: "bise et le" uttered by a male speaker, smoothed or wide band spectrogram, initial formant trajectories and regularized formant trajectories four three total rough labellings from the best to the worst

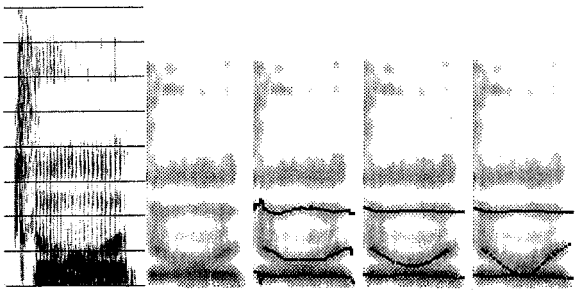

Figure 5: /too/ (extracted from "manteau") uttered by a male speaker
(a) spectrogram, (b) smoothed spectrogram, (c) initial formant trajectories, (d) and (e) regularized formant trajectories

References

[1] M.O. Berger and R. Mohr. Towards Autonomy in Active Contour Models. In *Proceedings of 10th International Conference on Pattern Recognition, Atlantic City, NJ (USA)*, pages 847–851. IEEE, June 1990.

[2] G. Fant. *Analytical constraints on the composition of speech spectra.* The Hague: Mouton & Co., 1970.

[3] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988.

[4] H. R. C. Tromp L. C. W. Pols and R. Plomp. Frequency analysis of dutch vowels from 50 male speakers. *J. Acoust. Soc. Amer.*, 53(4):1093–1101, 1973.

[5] M. Mrayati and B. Guerin. Etude des caractéristiques des voyelles orales françaises par simulation du conduit vocal avec pertes. *Revue d'Acoustique*, (36):18–32, 1976.

[6] V.W. Zue. From signals to symbols to meaning: on machine understanding of spoken language. In *Proc. of the XIIth International Congress of Phonetic Sciences*, pages 74–83, Aix-en-Provence, August 1991.