Global active method for automatic formant tracking guided by local processing

Marie-Odile BERGER and Yves LAPRIE
CRIN CNRS & INRIA Lorraine
B.P 239
F54506 Vandœuvre-les-Nancy

Abstract

In a formant tracking algorithm, combining both local viewpoint and global aspect of formant trajectories is difficult. We present an approach which incorporates local processing to build elementary tracks and an active method *which generates rough formant track hypotheses and makes the resulting trajectories move towards the formant trajectories.*

1 Introduction

An important issue in speech study is the knowledge of formant trajectories, and thus formant tracking, because it permits approximation of the vocal tract shape and is thus useful to get an idea of the uttered phones.

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 track evolution, and thus formant evolution is continuous. The formant tracking algorithm must thus interpret tracks in terms of formant as well as possible and deal with the following problems:

- a track may be a spurious one which 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.

Most formant tracking algorithms rely on local methods by connecting peaks of contiguous spectra and hence cannot give any appropriate response to the above-mentioned problems. Despite this, local tracking algorithms can be used to bootstrap a global and regularizing formant tracking algorithm which is the purpose of this work.

2 Principle

The proposed formant tracking algorithm includes local processing to build elementary tracks and optimization techniques to extend and to regularize elementary tracks. It operates as follows (Fig. 1):

1. Elementary tracks are built by local tracking on cepstrally smoothed spectra (Fig. 1.a).
2. Each elementary track is then interpreted in terms of formant trajectories (F1, F2, F3) and contributes to a total (for F1, F2 and F3 together and for the whole speech segment studied) and rough (accurate location and formant label are not well defined) interpretation of the elementary track set (Fig. 1.b). Actually in order to take into account possible errors the most likely total rough interpretations are generated. Every interpretation 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 a spurious one, there is no track for this formant, this track represents two merged formants.
3. For every rough interpretation, tracks are regularized, gaps are filled-in and formant trajectories

0-8186-2920-7/92 $3.00 © 1992 IEEE

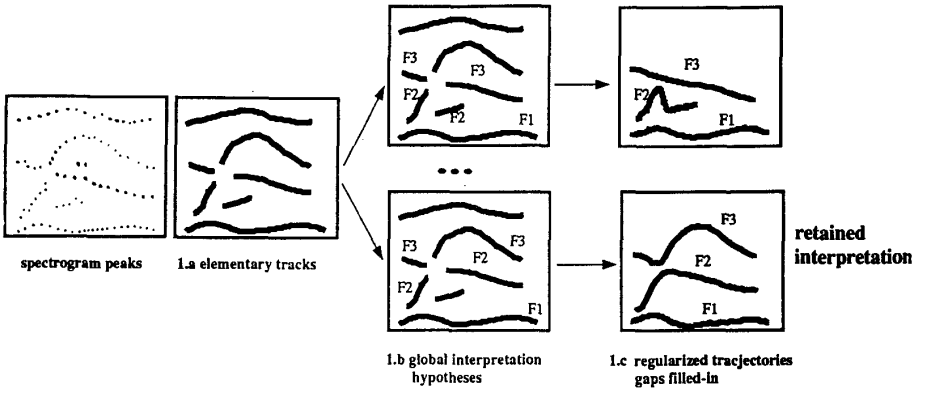

Figure 1: algorithm layout

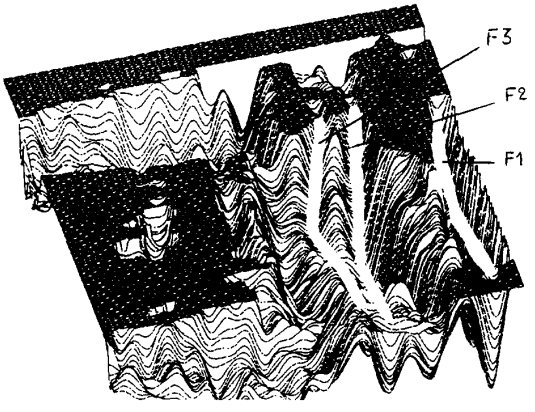

Figure 2: opposite spectrogram viewed as a potential field and formant trajectories

obtained are assessed (Fig. 1.c). With this aim in view the interpretation incorporating as much energy as possible in the neighborhood of the initial interpretation is located: the opposite of the cepstrally smoothed spectrogram is regarded as a potential field whose minima represent formants (Fig. 2). From a physical point of view a curve placed in the vicinity of a minimum is thus attracted to the nearest potential valley, i.e. the nearest formant trajectory; that is realized by minimizing an energy functionnal. This allows the gaps in formant tracks to be filled-in, tracks to be smoothed and energy incorporated by formant tracks to be calculated.

4. The global interpretation is the one which maximizes energy incorporated by formants while satisfying acoustic constraints imposed on formants.

3 Building elementary tracks

Tracks are built from peaks which have been detected on cepstrally smoothed spectra. Smoothing parameters are adapted to the speaker so that most peaks correspond to a formant; the higher the pitch the stronger the smoothing. In order to detect all the peak tracks which may fit a formant the time interval between two spectra is set to 1 or 2 ms. The local tracking algorithm connects peaks of contiguous spectra which should belong to the same track. It eliminates tracks that are too short and can detect a track in spite of the absence of some peaks. For every track built, all the peaks, their bandwidth and level are known.

4 Rough interpretation of peak tracks in terms of formants

The aim of this step is to interpret peak tracks as well as possible (which have been constructed at the previous step) in terms of formants. The algorithm used (steming from the one proposed in [4]) gives the n most likely interpretations; it relies on the following optimality criterion: the best track interpretation (for F1, F2 and F3 together) is the one which incorporates more energy in terms of formant than any other and which satisfies the constraints on formants. These constraints (defined for male and female speakers) are as follows:

- the frequency track must belong to the frequency domain of the formant it fits;
- the frequencies of tracks fitting F1 and F2 (resp. F2 and F3) must satisfy the frequency constraint

on F1-F2 (resp. F2-F3). For F1-F2 it is the well known vowel chart in the plane F1-F2.

The algorithm operates in two steps; first of all the best interpretation is (at any time) investigated locally according to the above mentioned optimality criterion; the set of instantaneous interpretations obtained is then used to bootstrap the global interpretation.

During the first step, the algorithm may assume that a formant is too weak to be detected (which prevents the interpretation from failing because there is no track for a formant), or that an elementary track fits two merged formants when its bandwidth is large enough.

During the second step, the n best interpretations are investigated by a recursive exploration which branches out in case there are two tracks in conflict to represent a sole formant or a track possibly fits two formants.

To ensure the hypothesis that two formants are merged in a sole track is relevant we take advantage of the redundancy of formant frequencies and formant levels [2]; this allows analytical verification that the track levels are compatible with this hypothesis.

5 Global formant tracking by an active method

Each formant hypothesis yields a sequence of elementary tracks labelled by F_1, F_2 or F_3. The elementary tracks labelled with the same formant are bound in chronological order by filling in the spaces with straight lines (Fig.3.b). This gives rise to curves which are a rough approximation of formant tracks. Those curves are called initialization curves in the remainder of this paper.

We have therefore to find the curve in the neighborhood of 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 curve minimizes the global energy

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

where $s : [0,1] \rightarrow I\!R^2$, $s \rightarrow v(s) = (x(s), y(s))$ is a parametrization of the curve. $E_{Formant}$ is all the smaller when the energy on the curve is great and E_{smooth} is all the smaller when the curve is regular. The parameter λ controls the compromise between the degree of regularization and its closeness to the elementary tracks. Equation (1) is solved variationally using the Euler equations:

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

Since this equation is ill-conditioned and since we are only interested in solutions which lie in the vicinity of the initialization, we are solving the associated evolution equation.

$$\frac{\partial v}{\partial t} - v'' + v^{iv} - \frac{\partial |E_{spectro}(v(s)|}{\partial v} = 0 \quad (3)$$

Starting from the initialization, the curve deforms 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.*

The problem is discretized and the curve is represented by a set of equidistant points. Equation (3) is then solved iteratively using a finite difference scheme which depends on the boundary conditions imposed on the curve to ensure the problem has a unique solution: since the formants are nearly parallel to the spectrogram x axis, we expect a track existing in the same time interval as the initial hypothesis; we therefore impose $x(0) = x_0(0)$ and $x(1) = x_0(1)$ (where $v_0 = (x_0, y_0)$ is the parametrization of the initialization) and only regularizing boundary conditions are imposed on y: $y''(0) = y'''(0) = y''(1) = y'''(1) = 0$. More explanations can be found in [1].

(Fig. 3.d) exhibits the global regularized solution from the initial curves (Fig. 3.c). A first assessment of the tracks obtained is based on the energetic density on the track.

This method is then 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 track assumption
- smoothing of the track

6 Results and discussion

Robustness and ability to detect accurate and regular formant trajectories are the main qualities of this algorithm. Fig. 3 and Fig. 4 show fine results both

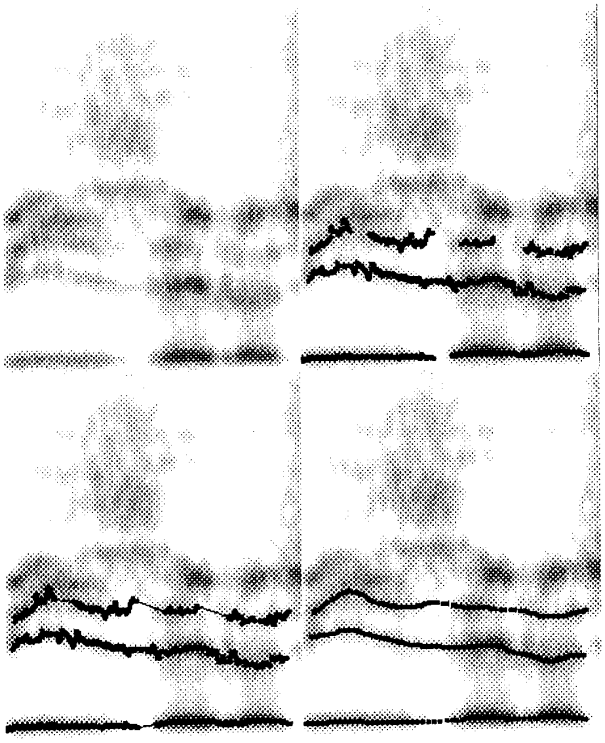

Figure 3: "bise et le" uttered by a male speaker
(a) spectrogram, (b) best hypothesis for the total interpretation, (c) inatializing curves (d) global regularized solution

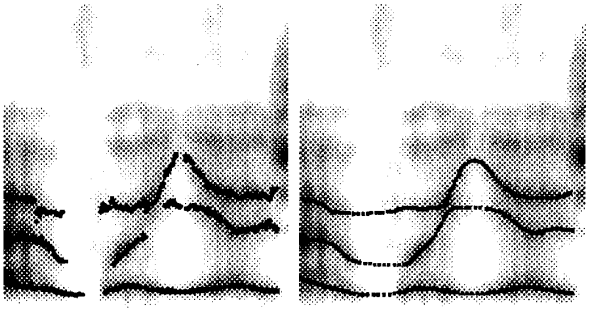

Figure 4: "un voyageur" uttered by a female speaker
(a) best rough total interpretation, (b) final result

for a poorly defined F3 and for voices with high pitch. This allows the use of a very sensitive local tracker, with a very short time step, to perform the elementary track detection. Even if elementary tracks show some small irregularities and errors, using a global regularization method allows the local anomalies to be overcome.

Moreover experimentation shows that only a few total interpretations are produced by the second step of our algorithm. The investigation of the generated total rough interpretations is therefore not time consuming.

Furthermore the implicit numerical formant tracking evaluation resulting from the minimization process enforces the strength of this approach. A minor weakness has been observed for voiced fricative sounds, where formant trajectories are attenuated by friction noise that makes the active curve drift towards noise peaks. Nevertheless this problem is encountered whatever the formant tracking algorithm.

7 Conclusion and perspectives

Most standard formant tracking algorithms are devoted to local formant tracking without dealing with the global regularity of formant trajectories, for the whole speech segment. The active method allows both smoothing and completion of the elementary tracks and formant trajectories to be reached progressively. We are currently improving our algorithm to distinguish two formants fitting a sole track. We are going to use the same method with a less smoothed spectrogram in the neihgborhood of this tracks exploiting the chirp z-transform [5].

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*, 1:321–331, 1988.

[4] Y. Laprie. Optimum spectral peak track interpretation in terms of formants. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 1261–1264, Kobe, Japan, November, 1990.

[5] L.R. Rabiner, R.W. Schafer, and C.M. Rader. The chirp z-transform algorithm and its application. *The Bell System Technical Journal*, 48:1249–1292, 1969.