Formant tracking adapted to acoustic-phonetic decoding

Y. Laprie

CRIN INRIA Lorraine
BP 239 54506 Vandœuvre-lès-Nancy France

Abstract

We describe a novel algorithm for formant extraction and tracking to be used in the context of acoustic-phonetic decoding. The construction of formant tracks involves both image processing techniques and geometric reasoning, which constitutes the originality and efficiency of the method.

The algorithm which works either on LPC or on cepstral data, comprises two distinct steps:

- first, formant segments are extracted from the signal by means of local processing,
- secondly geometric reasoning involving acoustic-phonetic knowledge is applied, in order to connect meaningful elementary tracks consistently and eliminate irrelevant segments.

1 Introduction

The importance of formant tracking in speech recognition as well as speech synthesis has led speech scientists to design automatic formant tracking algorithms. Actually two kinds of uses of formant tracking algorithm should be discerned:

Speech synthesis Assessment criteria of formant trajectories are the quality and intelligibility of the synthesized speech signal. Generally only the three first formants are required to perform synthesis.

Speech recognition Formant tracking is carried out to perform an acoustic-phonetic interpretation of speech signal. The phonetic decoder must have at its disposal formant trajectories for oral cavities and for nasal cavities. Moreover, all the acoustic coherent solutions should be available.

As a matter of fact most of the existing algorithms have been designed for speech synthesis and are not well suited to speech recognition. Besides some algorithms specially designed for the aim of speech decoding ([Laf 80]) most of the others ([McC 74], [Rab 69] . . . ) do not provide the acoustic-phonetic decoder with all the formant information it needs.

In the first section we will describe the data the algorithm works on. Insofar as the LPC method produces erroneous results for nasal sounds our algorithm can actually work either on LPC data or on cepstral data.

We consider that formant tracking is a two step process. The former is devoted to local tracking in order to generate elementary tracks, this does not require any knowledge. The latter is devoted to the propagation of elementary tracks, and takes into account different types of connection hypothesis. In the second section the way we build elementary formant tracks is described. In the third section the global formant tracking is outlined.

Finally we present and discuss results for a given phoneme sequence.

2 Formant frequency extraction

Formant frequencies are typically extracted using linear prediction spectra. The easiest way to perform this task is to pick the peaks of the linear prediction spectrum with the drawback that two merging formants might be represented by only one peak in the spectrum. Two closely spaced formant frequencies can be resolved by recomputing a spectrum inside of the radius 1 complex circle ([Mar 76] p 163). Actually root solving technique leads to all poles of the linear prediction spectrum and is currently used to extract formant frequencies. The major drawback of LPC is that it does not properly modelize the coupling between oral and nasal cavities. Thus formant frequency extraction with the LPC method can generate unrecoverable errors for nasal sounds.

In order to avoid this problem, our algorithm can process data stemming from cepstrally smoothed spectra. The cepstral smoothing does not involve any implicit speech production model and consequently yields correct formant frequencies even for nasal sounds. On the other hand, as the principle of cepstral smoothing is to separate the contribution of the vocal tract from the periodic excitation contribution, it is important to know the pitch period. Retaining too many cepstral coefficients compared to the pitch period length produces smoothed spectra with several peaks without any formant information. To get rid of these peaks without losing formant peaks we have optimized the parameters of the cepstral filtering \( l(nT) \) of [Rab 69].

\[
l(nT) = \begin{cases} 
1, & n \leq n_1 \\
1/2(1 - \cos(\pi(n - n_1)/(n_2 - n_1))), & n_1 \leq n < n_2 \\
0, & n \geq n_2
\end{cases}
\]

EUROSPEECH '89, Paris, France, September 1989
The following table gives the optimal parameters \( n_1 \) and \( n_2 \) for the range of normal male and female pitch frequencies.

<table>
<thead>
<tr>
<th>Pitch range</th>
<th>(&lt; 90) Hz</th>
<th>(&lt; 120) Hz</th>
<th>(&lt; 140) Hz</th>
<th>(&lt; 160) Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 )</td>
<td>18</td>
<td>17</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>( n_2 )</td>
<td>23</td>
<td>20</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

One notes that the higher the pitch frequency the smaller the \( n_1, n_2 \) parameters; this follows because for high values of pitch only a few cepstral coefficients are to be retained to keep the formant contribution. After adaptation to the pitch of the speaker studied, picking the visible (according to the spectrogram) peaks yields most formant frequencies.

As the following step in our algorithm uses image processing techniques a binary image (time x frequencies) is generated either with peaks of cepstrally smoothed spectrograms or with LPC roots. LPC and cepstrally smoothed spectra are computed every 2 ms with a 16 ms Hamming window.

## 3 Local Processing

The local formant tracking is designed to connect points which obviously belong to the same formant trajectory (without any knowledge on formant trajectories).

### 3.1 Local connection

Before building formant tracks the neighboring points are connected with dilatation, a basis transform of mathematical morphology which processes images as sets of pixels. Given a set of pixels \( B(x) \) centered on pixel \( x \) (here a rectangle of 7x5 pixels which corresponds to 14 ms \( \times \) 100 Hz), the dilatation of \( X \) by \( B(x) \) is the set of all the points \( z \) such that \( B(z) \) hits \( X \). Fig. 4 shows the dilatation of fig. 3.

### 3.2 Elimination of isolated points

Formant tracks are the largest connected components. Elimination of small components stemming from spurious peaks requires only a simple filtering according to the size of components.

It is very important that components have no hole, otherwise formant trajectories could give loops which are difficult to correct. It is with this aim in view that we extract the background connected component which hits the image border. The background pixels not belonging to this component are added to the connected component they make up an hole in.

### 3.3 Skeletonization of connected components

The connected components we have built are thick lines and are not well suited for the more abstract processing which follows. Thus they are skelletized from the image distance [Arc 85]. The main advantage of this algorithm is that it stores the end points and junction points of the skeleton that allows us to prune the meaningless short branches.

### 3.4 Elementary track building

In order to provide global formant tracking with elementary tracks the skeleton is explored to build lines which are coherent with the concept of formant. Thus the following rules have to be respected to obtain suitable formant candidates:

- a branch with two points for one instant cannot represent any formant and must at least be divided in two parts,
- when an input branch is divided into two or more branches the branch is extended with the output branch whose direction is the nearest to the input direction,
- the previous rule cannot be applied if one of the two output branches is much shorter than the other. In this case the input branch is extended by the longer one.

The way to generate the formant candidates is to choose one end point and then to explore the branch beginning with this end point. If a junction is encountered the branch is extended according to the rules we have just described. The construction is finished when each one of the end points and junctions has been visited.

### 3.5 Approximation of elementary tracks

The elementary tracks are then approximated with segments for the following reasons:

- it is simpler for the acoustic-phonetic decoder to work on formants represented by segments than on formants represented by a list of points,
- the next step of the algorithm is to connect the elementary tracks and it uses knowledge which is expressed with segments.

We will only outline the polygonal approximation algorithm. When approximating a branch \( t(A, B) \) \( (A \) and \( B \) are the end points) the point \( C \) whose distance to the straight line \( AB \) is maximal is determined. If the distance is small enough the track \( t(A, B) \) is approximated by the straight line \( AB \) otherwise the algorithm processes the subtracks \( t(A, C) \) and \( t(C, B) \) recursively. During approximation, tracks are updated if they do not satisfy the first rule.

## 4 Global reasoning

Until now we have built elementary formants tracks without taking into account the other tracks. In order to interpret these tracks in terms of formants we must:

- connect neighbouring elementary tracks that obviously correspond to the same formant,
- search for the possible formant crossings,
- score and if necessary update the tracks according
4.3 Track reliability

Connection hypothesis can be regarded as mutual enhancement of the confidence of the two corresponding tracks because the hypotheses are consistent with acoustic constraints which are imposed to formant trajectories. The larger the formant track, possibly extended by connection hypotheses, the higher its explicative potential and its reliability.

So that this approach will get complete short elementary tracks to be visited, possibly to be penalized.

If a short track is isolated or not consistent with the neighbour tracks it gets a penalization mark. If a short track has been connected with a larger one the neighbour tracks of the shortest will be revisited and possibly reorganized; that means that a subtrack which is no longer consistent with the global formant tracking can be penalized and isolated. Since no track has been deleted the phonetic decoder is provided with at least potentially all possible interpretations.

5 Results and discussion

In order to assess our algorithm we have tested it on a given phoneme sequence /waja/ for male and female speakers. This sequence has been chosen for the following reasons:

- F2 and F3 exchange their affiliation cavity, F2 is first bound to the front cavity (/w/) and then to the back cavity (/j/),
- formant transitions are large enough, so that continuity is insufficient for formant tracking,
- formant tracking is difficult for voices whose pitch frequency is high (> 200 Hz).

Fig. 5 shows the results for a female speaker whose pitch frequency is 250 Hz. The elementary tracks are represented by thick lines; there is one connection hypothesis at maximum of F3 and a formant crossing one which is represented by a thin line.

5.1 Comparaison of LPC and cepstral data

The performance of LPC and cepstral methods are almost equivalent for male speakers. On the other hand speaker adaptation with pitch frequency for cepstral smoothing gave much better results for all female speakers (fig. 2 and 3).

Furthermore the cavity affiliation exchanges appear more clearly with cepstral smoothing than with LPC. This difference is again more appreciable for female speakers than male speakers.

5.2 Local processing

The dilatation transform is a very simple and efficient way to connect neighboring points because it does not imply any knowledge (e.g. local directions) on the tracks which are built. The weak point of local processing is the skeletonization algorithm whose behaviour is difficult to control at end points. For two connected components which obviously should be connected to each other, the thinning algorithm may generate two skeletons whose end directions do not correspond exactly to those of the connected components. In the near future, we will implement the algorithm of [Bar 88] specially designed for line thinning which is the
case of connected components.

With regard to skeleton pruning, results could be slightly improved by taking into account the relative position of the branch to be pruned in elementary tracks. If a short branch is in the middle of a long track it can be pruned without any difficulty (fig. 6a). In case a branch is near end points (fig. 6b) it may be deleted only if its direction is consistent neither with that of the track that it belongs to nor with the neighbouring tracks.

5.3 Global formant tracking

Experiments show the effectiveness of the connections hypotheses we have proposed. Each of these hypotheses has been triggered for the situation it was designed for. The acoustic knowledge we have added to this algorithm represents only a part of what a human expert can use. Besides the existing connection hypothesis there is some common formant configurations the expert can take advantage of to connect the elementary tracks which may match a given configuration.

6 Conclusion

This algorithm takes the opposite view of the formant trackers based on Markov models. Formant trackers described by Kopec [Kop 86] do not include any "ad hoc" knowledge. Conversely our algorithm makes use of ad hoc knowledge to ensure that the formant tracks are consistent with the acoustic constraints which are imposed.

The main quality of this algorithm is that it pays attention to the acoustic events which generally escape the common formant tracking algorithms. This way of tracking formants which has already provided good results is a convenient way to enable formant extraction techniques (local processing) cooperate with acoustic knowledge stored as connection hypotheses and constraints.

References


