A COOPERATIVE APPROACH TO FORMANT EXTRACTION

Alain Soquet
Institut des Langues Vivantes et de Phonétique
Université Libre de Bruxelles

ABSTRACT
In this paper, we investigate a way to improve formant extraction reliability by using a cooperative approach. The basic motivation is that extraction methods based on different principles should not fail simultaneously when evaluating the same quantity. Therefore, we propose to combine independent formant extractors. Each method provides a set of formant candidates. These candidates are then combined with a vote mechanism in order to keep those that most probably correspond to a formant and to reject the majority of spurious values.

INTRODUCTION
Problems in formant frequency estimation are due both to the difficulty of estimating the resonances of the vocal tract at any given time (formant extraction), as well as to the difficulty of obtaining reasonable contours (formant tracking). Most formant tracking algorithms use the output of the extraction algorithm for successive segments and try to detect and correct extractor mistakes by using speech knowledge expressed, for example, by heuristic rules [1] or statistical models [2].

People are faced with many difficulties when trying to estimate the formant frequencies. A first problem is related to the coupling between the excitation and the vocal tract. For a given vocal tract configuration, the complexity of the estimation of the formant frequencies depends on the acoustic excitation and on the position where this excitation takes place; the difficulties encountered for high pitched voices are well known. A second problem is related to the precision with which the formant frequencies can be determined. Lindblom [3] estimated the accuracy of spectrographic measurement to be approximately equal to a quarter of the fundamental frequency. Monson et al. [4] compared the accuracy of spectrographic techniques and of linear prediction analysis in measuring formant frequencies on synthetic speech tokens. They observed that, "for fundamental frequencies between 100 and 300 Hz, both methods are accurate to within approximately ±60 Hz for both first and second formants. The third formant frequency can be measured with the same degree of accuracy by linear prediction, but only to within ±110 Hz by spectrographic means. The accuracy of both methods decreases greatly when fundamental frequency is 350 Hz or greater." This study clearly illustrates the degree of accuracy that can be expected from a given extractor.

In order to improve the general performances of the extraction, we propose to combine formant candidates provided by different basic extractors.

BASIC EXTRACTORS
The speech signal was passed through a 5 kHz cutoff low-pass filter, and sampled at 10kHz. The signal was then preemphasized (1 - 0.95 z^1) before further processing.

We have used three well documented basic extractors. We have chosen the linear prediction [1], the cepstrum [6] and the group delay functions [7].

- Linear prediction (LPC): The LPC coefficients were computed with the autocorrelation method on a 25.6 ms frame multiplied by a Hamming window. The number of poles of the predictive filter was fixed to 12. The formant frequencies can be estimated from these coefficients by different means (see Christensen et al. [5] for a discussion).

- Cepstrum: The second method is based on cepstral smoothing [6]. The cepstral coefficients were computed from a 16 ms frame multiplied by a Hamming window. The parameters of the cepstral filtering have been chosen as suggested in [6] and [8]. In order to enhance the formant resolution in the smoothed spectra, we used the Chirp-Z transform [6]: this transform consists in evaluating the spectrum on a circle of radius α < 1.

- Group Delay Function (GDF): The group delay functions are the negative derivative of the Fourier transform phase. We used the method proposed in [7]. This method involves deriving a signal with the characteristics of a minimum phase signal. The peaks of the GDF derived from this phase function correspond to formants.

COOPERATIVE APPROACH
The basic motivation is that extraction methods based on different principles should not fail simultaneously when evaluating the same quantity.

We adopted a majority vote among M methods based on the following principles. Let \( C(f) \) be the set of candidates provided by the \( i \)th method. \( C(f) \) is different from zero only for the values of \( f \) corresponding to a candidate of the \( i \)th method. The procedure consists in six stages:

1. Let \( F_{\text{min}} = 0 \).
2. Search for the first candidate such as \( f > F_{\text{min}} \). The procedure stops if no candidate is found.
3. Let \( N = 0 \).
4. For each method \( i \), look between the frequencies \( f \) and \( f + \Delta F \) (with \( \Delta F \) the length of the search interval) for the candidate with the lowest frequency. If it does exist, put its frequency and amplitude respectively in \( f_i \) and \( c_i \) and increment \( N \). Otherwise let \( f_i = 0 \) and \( c_i = 0 \).
5. If \( N > N_{\text{min}} \) then a candidate of frequency \( F \) is proposed by the cooperative approach:

\[
F = \frac{M}{\sum_{i=1}^{M} f_i c_i} \sum_{i=1}^{M} c_i
\]

and its amplitude \( C^{\text{vote}}(F) \) is given by:

\[
C^{\text{vote}}(F) = \sum_{i=1}^{M} c_i
\]

Let \( F_{\text{min}} = F + \delta F \), with \( \delta F \) the minimal frequency difference between two successive formants.

If \( N < N_{\text{min}} \) then let \( F_{\text{min}} = f \).

Two iterations of this procedure are presented graphically in figure 1 for \( M = 3 \) and in the case of an unanimous vote (\( N_{\text{min}} = 3 \)).

![Figure 1: Example of two iterations of the combination mechanism for formant frequency estimation.](image)

The parameters of the cooperative approach have been chosen as follows. The window length \( \Delta F \) is related to the precision of the evaluation by the different methods; we fixed \( \Delta F = 200Hz \). \( \delta F \) correspond to the minimal distance between two consecutive formant frequencies; we chose \( \delta F = 100Hz \). The choice of \( N_{\text{min}} \) depends on the performances of the basic extractors: if the extractors tend to propose too many candidates with the formants among them, an unanimous vote could be a good choice. On the contrary, if the extractors tend to miss formants, a vote at the absolute majority could be more adequate. We will thus come back on the choice of this parameter during the evaluation.

We have chosen \( M = 3 \) and used the three basic extractors described above. The different set of candidates are
derived from the basic methods as follows:

- $C_{\text{LPC}}(f)$ is created by finding the poles of the LPC transfer function and by taking the corresponding amplitudes of the LPC spectrum.

- $C_{\text{cepstr}}(f)$ is obtained by picking peaks of the Chirp-Z transform and by taking the corresponding amplitudes of the smoothed spectrum.

- $C_{\text{str}}(f)$ is obtained by picking peaks of the group delay function.

**EVALUATION**

The basic extractors and the cooperative approach have been evaluated on a corpus of VCV logatomes, with V a vowel among [a, o, e, i, u, y] and C a plosive among [p, t, k, b, d, g]. The corpus has been produced by three male speakers and segmented manually, in order to locate the segments with formant-like structure. A measurement of the first four formants has then been made every 10 msec giving a total of 11385 measurements. The reference has been obtained manually on the value of the Chirp-Z transform having been produced by three male speakers giving a total of 11385 measurements.

The reference has been obtained manually on the value of the Chirp-Z transform giving a total of 11385 measurements.

**Location of the insertion**

<table>
<thead>
<tr>
<th>Location of the insertion</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>before F1</td>
<td>xF1</td>
</tr>
<tr>
<td>between F1 and F2</td>
<td>F1xF2</td>
</tr>
<tr>
<td>between F2 and F3</td>
<td>F2xF3</td>
</tr>
<tr>
<td>between F3 and F4</td>
<td>F3xF4</td>
</tr>
</tbody>
</table>

**Table 1: Notations used for insertion errors.**

**Table 2: Notations used for omission errors.**

<table>
<thead>
<tr>
<th>Omission of one formant</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>x</td>
</tr>
<tr>
<td>F2</td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td></td>
</tr>
<tr>
<td>F4</td>
<td></td>
</tr>
</tbody>
</table>

The experiment has been conducted for two versions of the cepstrum depending on the value of the Chirp-Z transform coefficient: $\alpha = 0.95$ and $\alpha = 0.8$.

**Table 3: Comparison of omission errors for LPC, cepstrum ($\alpha = 0.95$ and $\alpha = 0.8$), GDF, VOTE 1 and VOTE 2 on the whole corpus.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Omission errors [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>46 235 146 810</td>
</tr>
<tr>
<td>Cepstr 0.95</td>
<td>21 1032 1261 949</td>
</tr>
<tr>
<td>Cepstr 0.8</td>
<td>142 487 982 1411</td>
</tr>
<tr>
<td>GDF</td>
<td>42 432 741 911</td>
</tr>
<tr>
<td>VOTE 1</td>
<td>76 1050 1220 1054</td>
</tr>
<tr>
<td>VOTE 2</td>
<td>170 589 878 1051</td>
</tr>
</tbody>
</table>

**Table 4: Comparison of insertion errors for LPC, cepstrum ($\alpha = 0.95$ and $\alpha = 0.8$), GDF, VOTE 1 and VOTE 2 on the whole corpus.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Insertion errors [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>14 95 177 407</td>
</tr>
<tr>
<td>Cepstr 0.95</td>
<td>825 845 2315 970</td>
</tr>
<tr>
<td>Cepstr 0.8</td>
<td>1626 1712 3328 1974</td>
</tr>
<tr>
<td>GDF</td>
<td>1845 697 497 481</td>
</tr>
<tr>
<td>VOTE 1</td>
<td>6 4 22 40</td>
</tr>
<tr>
<td>VOTE 2</td>
<td>2 3 32 67</td>
</tr>
</tbody>
</table>

The figure 2 shows the results for basic extractors: LPC, cepstrum ($\alpha = 0.95$ and $\alpha = 0.8$) and GDF. It can be seen that LPC is the most reliable of the four methods, in every error category. GDF turns out to give very satisfactory results with reasonable error rates both for omissions and insertions (except for the category $xF1$). For cepstrum, the lowering of the omission error rates related to the use of $\alpha = 0.8$ instead of $\alpha = 0.95$ is quite clear, but causes an important increase of the number of insertions. The reliability of this method remains quite low.

Given the important amount of insertion errors caused by cepstrum and GDF, we chose a unanimous vote ($N_{\text{min}} = 3$).

The method VOTE 1 is obtained by combining the candidates of the extractors LPC, cepstrum with $\alpha = 0.95$ and GDF. VOTE 2 is obtained by combining the candidates of the extractors LPC, cepstrum with $\alpha = 0.8$ and GDF.

The results for the individual extractors and the cooperative approach are presented on table 3 and table 4 respectively for the omission and the insertion errors.

It can be seen that the insertion error rates of the cooperative approaches are extremely low in comparison with the individual extractors. The lowering varies from a factor 10 with the LPC to 100 for cepstrum with $\alpha = 0.8$. This result clearly confirms the main hypothesis of the cooperative approach.

Unfortunately, the use of a unanimous vote has an important drawback: the omission error rates are comparable with those of the less performant extractor participating to the vote. The gain in performance of VOTE 2 compared to VOTE 1 directly reflect the lowering of omission errors of the cepstrum with $\alpha = 0.8$ instead of $\alpha = 0.95$.

**CONCLUSION**

The results show that the use of a cooperative approach allows the suppression of most of the insertion errors. Indeed, the number of insertion errors is reduced by a factor 10 to 100, depending on the individual method chosen as reference. This clearly illustrates the basic advantage of cooperative approach: the candidates proposed by the vote mechanism are likely to correspond to formant values. However, we have noted that the results for omission errors are comparable to those obtained by the least successful method used for the vote. Therefore, the individual methods have to be chosen so as to have a low omission error rate even if the insertion error rate is relatively high, since most of the insertion errors are eliminated by the vote mechanism.

**ACKNOWLEDGMENTS**

This work was partially supported by the "Communauté Française de Belgique" in the framework of the ARC 93/98—168 project.

**REFERENCES**


