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 preamphasized ($1 - 0{,}95\ z^{-1}$) 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 $\alpha < 1$.
- **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 principle. Let $C^i(f)$ be the set of candidates provided by the i^{th} method. $C^i(f)$ 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 $F_{min} = 0$.
2. Search for the first candidate such as $f > F_{min}$. 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 ΔF 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** $N \geq N_{min}$ **then** a candidate of frequency F is proposed by the cooperative approach:

$$F = \frac{\sum_{i=1}^{M} f_i c_i}{\sum_{i=1}^{M} c_i} \qquad (1)$$

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

$$C^{vote}(F) = \sum_{i=1}^{M} c_i \qquad (2)$$

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

If $N < N_{min}$ **then** let $F_{min} = f$.

6. Back to step 2.

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

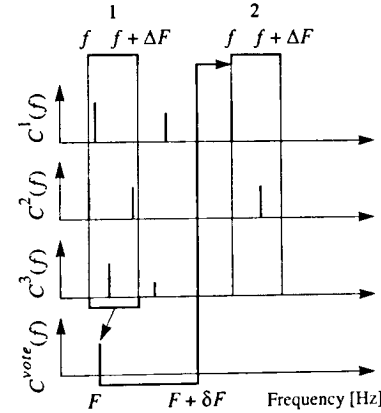

Figure 1: Example of two iterations of the combination mechanism for formant frequency estimation.

The parameters of the cooperative approach have been chosen as follows. The window length ΔF is related to the precision of the evaluation by the different methods; we fixed $\Delta F = 200Hz$. δF correspond to the minimal distance between two consecutive formant frequencies; we chose $\delta F = 100Hz$. The choice of N_{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:

- $C^{lpc}(f)$ is created by finding the poles of the LPC transfert function and by taking the corresponding amplitudes of the LPC spectrum.
- $C^{cepstre}(f)$ is obtained by picking peaks of the Chirp-Z transform and by taking the corresponding amplitudes of the smoothed spectrum.
- $C^{gdf}(f)$ 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, œ, 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 formantic 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 basis of different representations of the speech signal. We focussed on two sets of rough errors: the insertion errors (see table 1) and the omission errors (see table 2).

Table 1: Notations used for insertion errors.

Location of the insertion	Notation
before F1	$\times F1$
between F1 and F2	$F1\times F2$
between F2 and F3	$F2\times F3$
between F3 and F4	$F3\times F4$

Table 2: Notations used for omission errors.

Omission of one formant	Notation
F1	$\not F1$
F2	$\not F2$
F3	$\not F3$
F4	$\not F4$

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 $\alpha = 0.8$.

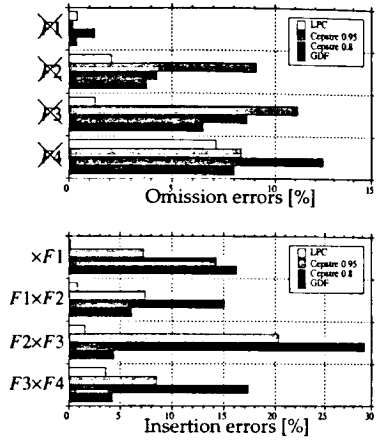

Figure 2: Omission and insertion errors for the different basic methods LPC, cepstrum ($\alpha = 0.95$ and $\alpha = 0.95$) and GDF.

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 $\times F1$). 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 ($N_{min} = 3$).

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

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

Method	$\not F1$	$\not F2$	$\not F3$	$\not F4$
LPC	46	235	146	810
Cepstre 0,95	21	1032	1261	949
Cepstre 0,8	142	487	982	1411
GDF	42	432	741	911
VOTE 1	76	1050	1220	1054
VOTE 2	170	589	878	1051

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

Method	$\times F1$	$F1\times F2$	$F2\times F3$	$F3\times F4$
LPC	14	95	177	407
Cepstre 0,95	825	845	2315	970
Cepstre 0,8	1626	1712	3328	1974
GDF	1845	697	497	481
VOTE 1	6	4	22	40
VOTE 2	2	3	32	67

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

[1] J. D. Markel, and A. H. Gray, "Linear prediction of speech," Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1976.

[2] G. E. Kopec, "Formant tracking using hidden markov models and vector quantization," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, n°4, pages 709-729, 1986.

[3] B. Lindblom, "Accuracy and limitations of sonagraph measurements," Proceedings of the 4th International Congress of Phonetic Sciences, Helsinki,The Hague, 1962.

[4] R. B. Monsen, and A. M. Engebretson, "The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction," Journal of Speech and Hearing Research, vol. 26, pages 89-97, 1983.

[5] R. L. Christensen, W. J. Strong, and E. P. Palmer, "A comparison of three methods of extracting resonance information from predictor-coefficient ceded speech," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, n°1, pages 8-14, 1976.

[6] R. W. Schafer, and L. R. Rabiner, "System for automatic formant analysis of voiced speech," J. Acoust. Soc. Am., vol. 47, n°2, pages 634-648, 1970.

[7] H. Murthy, and B. Yegnanarayana, "Formant extraction from Group Delay Function," Speech Communication, vol. 10, pages 209-221, 1991.

[8] Y. Laprie, "Formant tracking adapted to acoustic-phonetic decoding," Eurospeech, pages 669-672, 1989.