Formant-Contour Parameterisation of Vocalic Sounds by Temporally-Constrained Spectral Matching

Frantz CLERMONT

Department of Computer Science,
Australian Defence Force Academy,
University of New South Wales,
Canberra, ACT 2601, Australia

ABSTRACT: A method is described for estimating the three lowest, resonance or *formant* frequencies (F_1, F_2 and F_3) of a vocalic sound, and for tracking the temporal course of each formant through the duration of the sound. The problem of estimating these frequencies at a short-time frame of the speech signal is approached by spectral matching, i.e., by analysis-by-synthesis of hypothesised spectra. The problem of tracking such spectra over consecutive frames is then recast as an optimum path search, with temporal constraints defined in a Dynamic Programming framework. Both estimation and tracking algorithms hinge on the formant-enhancement and the formant-sensitivity properties of the negative derivative, Linear-Prediction phase spectrum. Using a moderately large dataset of ten English vowels produced randomly by a male speaker three times in CVd and VC (C = /b, d, g/) contexts, the method presented here is shown to yield formant-contours (the F_1 and F_2 in particular) which are very similar to those tracked manually by an expert phonetician. The strength of the correlations is found to be 0.99, 0.98 and 0.71 for F_1, F_2 and F_3, respectively.

INTRODUCTION

An important area of research in speech science and technology concerns the phonetic interpretation (Broad, 1972) of the acoustic, speech signal produced through the human vocal tract. Toward this end, the (three) lowest resonance frequencies of the vocal tract, also known as the formants, are particularly useful parameters because of the tangible acoustic link, which they provide between the phonetic identity of a speech sound and its associated articulatory gestures. However, whilst the speech signal is more readily observed than the articulators of the vocal tract, unsupervised formant-frequency estimation (especially through the speech stream) is fraught with appreciable difficulties, which are not fully resolved and thus make it difficult to automate acoustic-phonetic processing of speech.

The time-honoured method of formant estimation is based on hand-drawn tracings through regions of spectrograms, which display the highest concentrations of spectral energy. While formant measurements using the sound spectrograph are expected to be accurate, it is clear that preparation of spectrogram-traced formant data can easily become very laborious and even prohibitive.

Perhaps the most generally used, algorithmic approach to formant estimation consists of selecting, on a frame-by-frame basis, the three to four prominent peaks of the Linear-Prediction (LP) magnitude spectrum. Formant-*contours* are then constructed by applying some heuristic rules in order to ensure temporal continuity and proper ordering of the individual peaks. While this approach remains attractive owing to algorithmic simplicity and computational efficiency, it is not always free from post-estimation human intervention, nor does it overcome the problem of merged peaks, nor may it be expected to automatically secure consistent continuity along the time course of each individual formant.

We endeavour therefore in this paper to present a more robust method (Clermont, 1991; Clermont, 1988a): (1) which can be applied to determine in some error sense the hypothesised (or candidate) spectrum that best matches the original spectrum of a given frame; and (2) which embodies certain constraints aimed at objectively restricting choice of consecutive candidate spectra.

METHOD OF FORMANT-CONTOUR PARAMETERISATION

APPROACH:

The problem of formant-contour parameterisation (or formant tracking as better known in the literature) of a vocalic sound is recast here as that of finding, in some error sense, the sequence of candidate spectra which best match the original spectra through the duration of the sound. The process of generating "candidate" or "trial" spectra per time-frame is of course reminiscent of pioneering research by Bell et al. (1961) on the Analysis-by-Synthesis (AbS) process and the underlying concept of an error measure associated with every candidate spectrum. If the AbS-process is carried for a sequence of time-frames as would be the case for the problem at hand, then the resulting error measures might be interpreted as per-frame (or local) costs. By combining local costs with some frame-to-frame transition costs, the formant-tracking problem may then be re-defined as that of finding the best set of time-sequenced candidate spectra such that some total cost is minimised over all time frames. This is equivalent to solving a sequential optimisation problem, to which the technique of Dynamic Programming (DP) is well suited (Bellman, 1957; Lee, 1989). Hence, the acronym DPTRAK sometimes used hereafter to refer to our formant-tracking method.

ANALYSIS-BY-SYNTHESIS and the LINEAR-PREDICTION PHASE SPECTRUM:

One elegant property of the LP-phase spectrum studied by Yegnanarayana (1978) concerns the mutual interference of adjacent poles. The interference is attenuated in the Negative Derivative of the linear-prediction Phase Spectrum (NDPS) which, in contrast to the LP-magnitude spectrum, consists of additive rather than multiplicative sets of resonance curves. The NDPS may therefore be expected to behave like a filter which de-emphasises spurious peaks and enhances true formant peaks. This behaviour is re-interpreted as the the formant-enhancement property of the NDPS.

A *corollary* of the formant-enhancement property concerns the Euclidean distance between a pair of NDPS, which was shown (Yegnanarayana and Reddy, 1979) to be sensitive to deviations localised only around formant-peak regions, and may therefore be interpreted as a Euclidean distance between a paired set of actual formant peaks. In addition, the NDPS distance is functionally equivalent to an index-weighted, Euclidean distance between a paired set of LP-derived cepstral coefficients. The NDPS may therefore be said to also embody a formant-sensitivity property.

In light of the properties reviewed above, the NDPS distance could then be used to indirectly estimate formant-related similarity between entire speech sounds, as well as between selected spectral regions of a single frame of a given sound. This further suggests the possibility of determining which subsets of spectral peaks, or equivalently which spectra synthesised from these subsets of peaks, best match the original spectrum of a given frame, subject to minimal influences by non-considered peaks. From this speculation arises the concept of per-frame "simplified spectra", which may be synthesised from combinations of the available candidate peaks.

Two concepts are therefore put forward to achieve spectral analysis-by-synthesis on a frame-by-frame basis. Given a set of candidate peak frequencies and bandwidths, one generates certain combinations of three- or four-peak sequences which are, in turn, synthesised to produce our so-called simplified spectra. A Spectral Distance Matrix (SDM) is then constructed, in which columns correspond to time frames, and rows contain NDPS distances between simplified and original spectra.

DYNAMIC PROGRAMMING and TEMPORAL CONSTRAINTS:

The decision process inherent to a DP-optimisation is interpreted in our context to have as many stages as there are time frames, and there will be per stage as many states as there are available simplified spectra. The objective function at any given stage (later than the first stage) is defined as the sum of contributions from that stage and previous ones. A contribution is expressed as the minimum of linear combinations of the local (or state occupancy) costs at a given stage (later than the first one) and of the (transition) costs in associating every state of that stage to those of the previous stage. For every contribution a DP-pointer is retained which links a current to a previous state.

If occupancy and transition costs are quantified in terms of a suitable similarity measure (such as the NDPS distance discussed earlier), then the DP-decision process may be said to minimise the cumulated cost through an NDPS distance matrix. The optimum sequence of states over time is retrieved by first finding the state of the last stage at which the cumulated cost (or distance) is minimum, and by sequentially recalling the DP-pointers to trace back the optimum path from the last to the first stage. The formant-frequency contours are then constructed by recursive selection of the simplified spectra pertaining to the optimum DP-path.

ALGORITHMIC IMPLEMENTATION of the DPTRAK METHOD

SPECTRAL ANALYSIS-BY-SYNTHESIS and DISTANCE MATRIX (PART 1):

Part 1 of the algorithmic implementation of the DPTRAK method first consists of traditional pre-processing of the input speech signal. All-pole, Linear Predictive Coding analysis is performed at every frame, and the pole frequencies and bandwidths resulting from the polynomial root-solving form our set of candidate formant frequencies and bandwidths (or candidate peaks in short). This preliminary analysis also involves deriving LP-cepstral coefficients, which are used to represent our so-called *original* spectrum at every frame.

The pairs of candidate frequencies and bandwidths of a given set per frame are then combined, in turn, to form *four*-peak sequences which are first converted into LP-autoregressive coefficients and then into LP-cepstral coefficients (Markel and Gray, 1976: see conversion algorithms on pp. 95-98). For a given frame, there are as many sets of the latter cepstral coefficients as there are four-peak sequences (or equivalently simplified spectra).

Note that only *crescendo* sequences are considered in order to preserve the constraint that formant frequencies are to be in increasing order on the resonance scale. Note further that, although our aim is to estimate the first three formant peaks, the inclusion of a fourth peak automatically enlarges our choice of simplified spectra, and may thus be expected to improve the likelihood of finding at least the first three peaks which best match those of the original spectrum at a given frame.

Thus, Part 1 of the algorithmic implementation yields the Spectral Distance Matrix (SDM), the rows of which contain index-weighted cepstral (or NDPS) distances between simplified spectra and original spectra at individual frames. The matrix has as many as columns as there are time-frames, and there are per column as many row entries as there are simplified spectra at every frame.

DP-SEARCH FOR OPTIMUM PATH (PART 2):

Part 2 consists of finding an optimum path through the SDM constructed at the previous stage, by minimising a DP-cost function which combines NDPS distances local to every frame with NDPS distances between simplified spectra of consecutive frames. The *local component* of the DP-cost function is given by the SDM row entries, and the *temporal-continuity component* is defined in terms of frame-to-frame NDPS distances. The further use of the NDPS distance for estimating spectral continuity between consecutive frames is well motivated by the formant-sensitivity property described earlier. Accordingly, the total DP-cost function may be expected to have a strong influence on the temporal alignment of the respective, individual peaks of consecutive spectra. The total cost function may be described as a cumulative sum of local and frame-to-frame NDPS distances, and is expressed mathematically as follows:

$$D_{j,n} = d_{j,n} + \min_{j=1}^{N_c} \{ D_{j,n-1} + d[(C_k)^s_{j,n} , (C_k)^s_{j,n-1}] \}, \quad \text{for } 2 \leq n \leq N_t \qquad (1)$$

where the $(C_k)^s_{j,n}$ and $(C_k)^s_{j,n-1}$ are the LP-cepstral coefficients for simplified spectra of current and previous frames, respectively, and where:

- $j \equiv$ row index of SDM.
- $n \equiv$ column index of SDM (also time-frame index).
- $N_t \equiv$ number of time frames.
- $N_c \equiv$ per-frame number of combinations of N_p-peaks taken q=4 at a time $(N_p!/[q!(N_p - q)!])$.
- $d_{j,n} \equiv$ local cost (NDPS distances of SDM).
- $d[(C_k)^s_{j,n} , (C_k)^s_{j,n-1}] \equiv$ continuity cost (Frame-to-frame NDPS distances).
- $D \equiv$ cumulative cost.
- NDPS distance $d = \sum_{k=1}^{M} k^2 [C_k - C'_k]^2$, where C_k and C'_k are a paired set of LP-cepstral coefficients, and M is a finite number of coefficients.

Note that the total number (N_c) of four-peak combinations may be different for every frame, depending on the number of candidate peaks available from the all-pole LP-analysis. For presentation's sake, however, N_c is assumed to be the same for every frame in our mathematical formulation (Equation 1) of the DP-cost function.

PERFORMANCE EVALUATION of the DPTRAK METHOD

APPROACH:

The stumbling question which arises in evaluating a formant "tracker" is deciding how good the estimated or tracked values are. Although the method presented here uses objective measures to determine the best matching, consecutive spectra through a vocalic sound, it will perform only as well as can be expected within the limits of all-pole, LP-modelling of voiced speech sounds. It is therefore desirable to seek independent judgements of the formant values obtained automatically. One approach considered in this study is to compare, for the same speech material, the F_1, F_2 and F_3 yielded by the DPTRAK method to a *reference* formant set hand-edited by a human "expert" (an acoustic phonetician in this case). The degree of proximity between the two sets may then be interpreted as a measure of goodness of the estimated formants.

Since an important aspect of our approach to formant tracking is the use of temporal constraints to find the best matching contours, formant-data comparison was also carried to assess global effects of our DP-based constraints. Toward this end, two distinct sets of computer-generated formant data were created. *DP0-contours* refer to temporally-unconstrained, formant-contours which were constructed by selecting, for every frame, the simplified spectrum closest to the original spectrum in an NDPS sense, and then juxtaposing the so-selected simplified spectra frame-by-frame. In contrast, *DP1-contours* refer to temporally-constrained, formant-contours which were obtained recursively along the optimum DP-path of the SDM.

SPEECH MATERIAL and REFERENCE FORMANT DATA:

The reference formant data used for comparison were measured in a previous study (Broad and Clermont, 1987) of coarticulation in English syllables. The speech material comprises ten American English vowels in CVd and VC contexts (where C = /b, d, g/, and V is as in "beet", "bit", "bet", "bat", "but", "hot", "bought", "foot", "boot", "bird") produced by one adult male, native speaker of American English. There are three repetitions of each CVd and VC syllable; the waveforms were quantised to 12 bits and sampled at 10 kHz. The vowel boundaries were determined by visual inspection of the waveforms, followed by auditory confirmation.

The formant-frequencies F_1, F_2 and F_3 were estimated at 11 equally spaced frames, through the vowel interval of each syllable for a total of 1980 frames. The formant-estimation method used by the acoustic phonetician proceeded in two steps: (1) all-pole, root solving of 14th-order (autocorrelation) LP-analysis on Hamming-windowed frames of 25.6-msec duration, followed by (2) hand editing of the resulting formant candidates on the basis of the estimated bandwidths, expected formant ranges, and temporal continuity. When a formant could not be selected using these criteria, a value was interpolated between adjacent frames.

RMS DIFFERENCES and CORRELATIONS:

The results obtained by comparing computer-generated with hand-tracked vowel formant-contours are described in Table 1, in terms of correlations and Root-Mean-Square (RMS) differences (or errors).

In particular, the DP0- and DP1-results are summarised in the middle and rightmost columns of Table 1, respectively. Observation of these two columns first indicates that the RMS-error is relatively much smaller for F_1 and F_2 than it is for F_3. The correlations shown in parentheses confirm that our method fares rather well in tracking the lowest two formants with correlations near unity for F_1 (*0.99*) and F_2 (*0.98*), while DP-tracking of the F_3 appears to be relatively less successfull with a correlation near only *0.7*. Furthermore, cross-examination of the DP0- and the DP1-results reveals an appreciable decrease in RMS-error for F_1 (18%) and F_2 (28%) and a much smaller improvement for F_3 (3%). These results generally suggest that the use of NDPS distances combined with DP-based temporal constraints does yield a better choice of time-sequenced simplified spectra.

An interesting question now arises as to whether the RMS-errors reported above can be re-interpreted in order to shed more light on the "goodness" of the DP-tracked formant estimates. In other words, is the intuitively small, RMS-error of 21 Hz for DP1-tracked F_1-contours, for example, small enough? Or, is the relatively large, RMS-error of 342 Hz for DP1-tracked F_3-contours within acceptable limits?

(American English) Vowel Formant-Contours in CVC context	DP0-Results (without temporal constraints)	DP1-Results (with temporal constraints)
F_1	25 Hz (0.98)	21 Hz (0.99)
F_2	149 Hz (0.95)	107 Hz (0.98)
F_3	351 Hz (0.70)	342 Hz (0.71)

Table 1: Summary of DPTRAK performance in estimating F_1-, F_2- and F_3-contours of American English Vowels (Broad and Clermont, 1987) produced in CVC context by one adult male, native speaker. Results are expressed in terms of *RMS-errors* (Hz) and *correlations* (shown in parentheses) between contours yielded by the DPTRAK method and those tracked by an acoustic phonetician. Input candidate (formant) peaks for both machine and human "trackers" are the same roots of 14th-order, all-pole LP-filtering. The DP0-results (*middle column*) concern temporally-*unconstrained* contours. The DP1-results (*rightmost column*) concern temporally-*constrained* contours.

(American English) Vowel Formant-Contours in CVC context	Inter-Repetition Dispersion	DP1-Results (within IR-dispersion)
F_1	25 Hz	2 Hz (99%)
F_2	52 Hz	4 Hz (98%)
F_3	75 Hz	4 Hz (80%)

Table 2: Summary of DPTRAK (DP1-process) performance (rightmost column) in relation to inter-repetition dispersions (middle column). Adjacent to RMS values (Hz) are shown in parentheses the corresponding percentages of differences (between DP1-contours and hand-tracked formant-contours) which lie within the respective inter-repetition dispersions for F_1, F_2 and F_3.

ERROR ANALYSIS within INTER-REPETITION DISPERSION:

One approach to re-interpreting the RMS-errors discussed thus far is to examine them in relation to the dispersion across the three repetitions of the hand-tracked formant data, for one can hardly expect to overcome the limits of measurement noise and random variations amongst repetitions of the same utterance. The overall Inter-Repetition (IR) dispersion is defined, for each formant, as the square root of the pooled inter-repetition variance for a given vowel. If the DPTRAK method was exact for the population-mean contours, then the expected RMS-error against the hand-tracked data would be of the order of the IR-dispersion. The following questions can then be addressed with the IR-dispersion as a baseline measure: (1) how many of the differences between DP-tracked and hand-tracked values of the F_1, F_2 and F_3 lie within the spread of their respective IR-dispersion?; and (2) what is the global RMS-error for that subset of the DP-tracked formant values which differ from the hand-tracked ones by less than the IR-dispersion?

The results obtained within respective IR-dispersions of F_1, F_2 and F_3 are summarised in the rightmost column of Table 2. One can observe that nearly all the differences between DP1-tracked and hand-tracked F_1- (99%) and F_2-contours (98%) lie within the spread of their respective IR-dispersions. In addition, the corresponding RMS-errors (2 and 4 Hz, respectively) clearly indicate that the DPTRAK method is in nearly perfect agreement with the human "expert". In contrast, only 80% of the differences for F_3 do not exceed the limits of the IR-dispersion, thus confirming that there is a definite discrepancy between the two tracking methods as far as the third formants are concerned. Until further investigation is conducted in a future study, the DPTRAK method is retained to be in error 20% of the time in estimating F_3.

CONCLUDING DISCUSSION

A method has been herein described for automatically tracking the temporal course of the three lowest, resonance frequencies measured through the duration of vocalic sounds. The distinctive aspect of the method is perhaps better portrayed in terms of the judicious use of the formant-enhancement and formant-sensitivity properties of the LP-phase spectrum, which appears to have received less attention than the now traditional, yet more limited LP-magnitude spectrum.

Although the formant-tracking problem has been quite independently (Talkin, 1987) recast in a DP-framework similar to that described here, our approach differs markedly as far as the local and transition costs associated with the DP-search are concerned. In this regard, the NDPS embodies very desirable properties, which make it possible to indirectly estimate the proximity between a paired set of spectral peaks, both on a per-frame and on a frame-to-frame basis. Thus, while the DP-cost function in our method is simply, yet adequately expressed in terms of index-weighted cepstral (or NDPS) distances, Talkin's local component of the cost function, for example, is defined as a weighted sum of the very bandwidths and frequencies of the candidate peaks. The method presented here would seem therefore to be more attractive and perhaps more robust, as deviations between paired sets of spectral peaks are automatically monitored in our DP-framework which integrates the properties of a more enhanced LP-spectral representation.

Our evaluation of the formant-tracking method described in this paper attests to a satisfying performance on a moderately large amount of vowel data. Comparison of DP-tracked with hand-tracked formant-contours indicates that the method may be expected to fare well in estimating especially the F_1- and F_2-contours of vocalic sounds. The relatively less successful results obtained for F_3 are not yet fully understood, although preliminary examination of the errors seems to indicate that it is the back vowels which are most affected. Further study is therefore warranted in order to overcome the apparent limitation of the DPTRAK method in tracking F_3-contours of certain vocalic sounds. Notwithstanding this limitation, the method represents a step forward as it may be expected to automatically yield reliable contours of F_1 and F_2, which are generally considered "the most characteristic two features of vowels" (Pols *et al.*, 1969).

REFERENCES

Bell, C.G., Fujisaki, H., Heinz, J.M., Stevens, K.N. and House, A.S. (1961), "Reduction of speech spectra by analysis-by-synthesis techniques", J. Acoust. Soc. Am. *33*: 1725-1736.

Bellman, R. (1957), *Dynamic Programming* (Princeton University Press, Princeton, New Jersey).

Broad, D.J. (1972), "Formants in automatic speech recognition", J. Int. Man-Machine Studies, *4*: 411-424.

Broad, D.J. and Clermont, F. (1987), "A methodology for modeling vowel formant contours in CVC context", J. Acoust. Soc. Am, *81*: 155-165.

Clermont, F. (1991), "Formant-Contour Models of Diphthongs: A Study in Acoustic Phonetics and Computer Modelling of Speech", *Doctoral Thesis*, The Australian National University: Research School of Physical Sciences and Engineering, Computer Sciences Laboratory.

Clermont, F. (1988a), "Formant contour extraction by a temporally-constrained search of the spectral resonance space", J. Acoust. Soc. Am., *84*: S21-22.

Lee, C-H. (1989), "Applications of dynamic programming to speech and language processing", *AT&T* Technical Journal *68*: 114-130.

Markel, J.D. and Gray, A.H. (1976), *Linear Prediction of Speech* (Springer-Verlag, Berlin).

Pols, L.C.W., van der Kamp, L.J.T. and Plomp, R. (1969), "Perceptual and physical space of vowel sounds", J. Acoust. Soc. Am. *46*: 458-467.

Talkin, D. (1987), "Speech formant trajectory estimation using dynamic programming with modulated transition costs", J. Acoust. Soc. Am. *82*: S55.

Yegnanarayana, B. (1978), "Formant extraction from linear-prediction phase spectra", J. Acoust. Soc. Am. *63*: 1638-1640.

Yegnanarayana, B. and Reddy, R. (1979), "A distance measure derived from the first derivative of the linear prediction phase spectra", IEEE Int. Conf. on Acoust., Speech and Sig. Proc., Conf. Record: 744-747.