An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra

STEPHANIE S. McCANDLESS

Abstract—**An algorithm is presented which finds the frequency and amplitude of the first three formants during all vowel-like segments of continuous speech. It uses as input the peaks of the linear prediction spectra and a segmentation parameter to indicate energy and voicing. Ideally, the first three peaks are the first three formants. Frequently, however, two peaks merge, or spurious peaks appear, and the difficult part is to recognize such situations and deal with them. The general method is to fill formant slots with the available peaks at each frame, based on frequency position relative to an educated guess. Then, if a peak is left over and/or a slot is unfilled, special routines are called to decide how to deal with them. Included is a formant enhancement technique, analogous to a similar technique which has been implemented via the chirp-z transform [8], which usually succeeds in separating two merged formants. Processing begins at the middle of each high volume voiced segment, where formants are most likely to be correct, and branches outward from there in both directions in time, using the most recently found formant frequencies as the educated guess for the current frame.**

The algorithm has been implemented at Lincoln Laboratory on the Univac 1219 and the Fast Digital Processor, a programmable processor [9], and has been tested on a large number of unrestricted sentences.

INTRODUCTION

THE SPEECH waveform can be modeled as the response of a resonator (the vocal tract) to a series of pulses (quasi-periodic glottal pulses during voiced sounds, or noise generated at a constriction during unvoiced sounds). The resonances of the vocal tract are called formants, and they are manifested in the spectral domain by energy maxima at the resonant frequencies.

The frequencies at which the formants occur are primarily dependent upon the shape of the vocal tract, which is determined by the positions of the articulators (tongue, lips, jaw, etc.). In continuous speech, the formant frequencies vary in time as the articulators change position.

The formant frequencies are an important cue in the characterization of speech sounds, and therefore, an automatic algorithm for reliably computing these frequencies would be useful for many aspects of speech research, such as speech synthesis, formant vocoder's, and speech recognition. Two basic approaches to the problem have been tried—analysis by synthesis and peak-picking from smoothed spectra.

In analysis by synthesis, an educated guess is made of the formant frequencies and bandwidths, and a spectrum is generated based on the educated guess. The formant frequencies for the synthesized spectrum are varied systematically until the differences between it and the actual spectrum are minimal, according to some criterion [2]. Olive [6] has recently worked out a method for varying all three formant frequencies, using a Newton–Raphson technique to find a least-squares fit.

In peak-picking, certain rules are applied to select the appropriate peaks from a smoothed spectrum at each frame to be the first three formants. The challenge is in recognizing which peaks are spurious and/or whether two formants have merged into one peak. Schaefer and Rabiner [8] found the best candidate peak in a specified region for each formant, using cepstrally smoothed spectra. Markel [5] used peak-picking from linear prediction spectra. If there were exactly 3 peaks under 3 kHz, he assumed these were the first three formants. Otherwise, continuity constraints were applied to delete a peak or insert a formant.

Analysis by synthesis has certain advantages in that it incorporates the entire spectral shape rather than simply the spectral peaks. Hence, a small spurious peak would not grossly change the general spectral shape caused by the formants, and therefore, would not alter the results drastically. Its disadvantages are that it requires a great deal of processing and that it depends on an accurate speech model. Hence, the method might have difficulties with sounds other than nonnasalized vowels.

Peak-picking is a more tractable problem with linear prediction spectra than with other forms of spectral analysis because spurious peaks are rare. However, peak mergers are common, as well as peak cancellations due to nasalization effects. The method presented here is a systematic, fully automatic algorithm, which is usually successful in solving these problems. It yields the frequencies of the first three formants and the spectral amplitudes at their frequency positions during all sonorant sounds in continuous unrestricted speech. No attempt was made to extract formant frequencies during obstruent sounds because the acoustic characteristics of these sounds are not well represented through formants. In addition, the spectral characteristics of the noise source tend to mask the vocal tract resonances.

LINEAR PREDICTION

The following is a brief discussion of linear prediction analysis. A more detailed treatment can be found elsewhere in the literature [1], [4], [5].

Manuscript received July 16, 1973; revised October 15, 1973. This work was sponsored by the Advanced Research Projects Agency of the Department of Defense.

The author is with Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Mass. 02173.

Approximate the sampled speech waveform $s(n)$ by another sequence $\hat{s}(n)$, by linearly predicting from the past p samples of $s(n)$:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k). \tag{1}$$

The unknowns a_k in (1), can be determined by minimizing the mean squared difference, E, between $s(n)$ and $\hat{s}(n)$ over N samples of $s(n)$:

$$E = \frac{1}{N}\sum_{n=0}^{N-1}[s(n) - \hat{s}(n)]^2$$

$$= \frac{1}{N}\sum_{n}[s(n) - \sum_{k} a_k s(n-k)]^2. \tag{2}$$

By setting $\partial E/\partial a_j$ to 0 for $j = 1,2,\cdots,p$, and simplifying, one obtains:

$$\sum_{k} a_k \phi_{jk} = x_j \qquad \begin{matrix} j = 1,2,\cdots p \\ k = 1,2,\cdots p \end{matrix} \tag{3}$$

where

$$\phi_{jk} = \sum_{n} s(n-j)s(n-k)$$

$$x_j = \phi_{j0}.$$

In the study discussed here, the speech waveform was sampled at 10 kHz after 6 dB/octave preemphasis. A new set of $p = 14$ coefficients a_k was computed every 5.0 ms on 25.6 ms of Hanning windowed speech, using the digital inverse filtering technique [5].

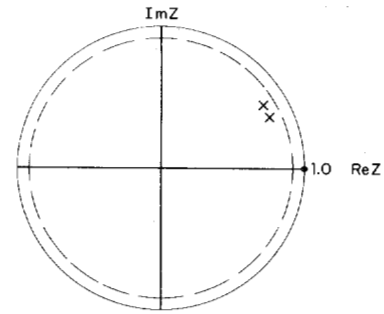
The choice of 14 predictor coefficients was arrived at through experimentation. It was found that, with fewer coefficients, the merging formants of certain sounds, such as ɜ and ɔ, were represented by only one complex pole-pair. With 14 coefficients the problem of spurious peaks is more common. However, the algorithm is capable of handling spurious peaks, whereas a missing pole-pair is an irrecoverable problem.

Once the coefficients a_k are available, it is an easy matter to obtain the approximated spectrum of $s(n)$. One simply evaluates the magnitude of the transfer function $H(z)$ of the filter represented by the coefficients a_k, at N equally spaced samples along the unit circle in the z-plane:

$$H(z) = a_0/1 - \sum_{k=1}^{p} a_k z^{-k} \tag{4}$$

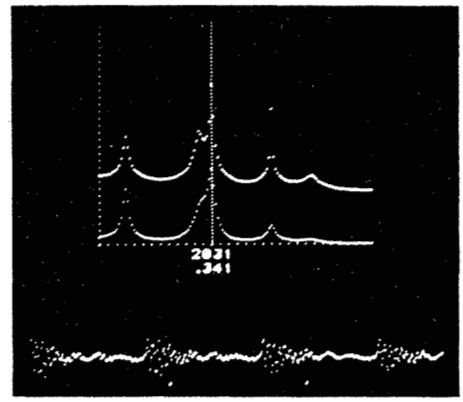
where (4) is evaluated at $z = \exp[j(2\pi n/N)]$ for $n = 0,1,\cdots,N-1$.

N can be chosen arbitrarily large to increase frequency resolution, at the expense of computation time. For our purposes, we have chosen $N = 256$, resulting in approximately 40 Hz spectral resolution. This means that 1) formant values are accurate to within 20 Hz and 2) any two spectral peaks which are within 80 Hz of each other cannot be resolved. The accuracy is sufficient for our needs, and we found that the second constraint rarely causes problems.



Evaluate $H(z) = \dfrac{a_0}{1 - \sum_{k=1}^{p} a_k z^{-k}}$ at equally spaced samples along a circle of radius <1, to enhance the peaks of the 2 poles that are close together.

(a)



(b)

Fig. 1. (a) Formant enhancement. (b) Linear prediction spectrum in | r | of "there" before and after enhancement.

Two closely spaced formants frequently merge into one spectral peak, and cannot be resolved on the unit circle even with infinite resolution. However, they can often be separated by simply recomputing the spectrum on a circle of radius less than 1. This amounts to reevaluating $H(z)$ at $z = re^{j(2\pi n/N)}$, $r < 1$. Because the contour comes in closer to the two poles, their peaks are enhanced, and a separation is effected (Fig. 1).

The method also frequently works to bring out a peak whose bandwidth was too wide due to pole zero interplay (as in nasalized vowels). Essentially, moving the circle inward is equivalent to moving the poles outward to a larger radius, and, therefore, to a narrower bandwidth.

PEAKS VS. POLES

Clearly, an obvious method for extracting formants from linear prediction would be to solve for the poles of the filter by setting the denominator in (6) to zero and solving for the roots of the resulting pth order polynomial in z. Some or none of the roots would be real, and the rest would be complex pole pairs which might or might not be

formants. Out of those pole pairs, one would have to select three on the basis of frequency location, sufficiently narrow bandwidth, and some kind of formant continuity criterion, to be the first three formants.

Another technique, requiring much less computation, would be to simply pick the first three peaks in the spectrum and call those the first three formants, making the assumption that a pole strong enough to show up as a peak is necessarily a formant. Such a method works very well most of the time, but mistakes will occur during the following situations.

1) Often two poles show up as only one peak because they are close together in frequency.

2) Occasionally a pole due to frequency shaping will appear as a small peak, which would be incorrectly interpreted as a formant.

It was decided to use peak-picking rather than root extraction, and to develop an algorithm to solve cases 1) and 2) above.

NASALS AND NASALIZATIONS

Nasals present a special problem to any formant tracking algorithm because there are zeros in the transfer function in addition to the poles. In a nasal, the poles are resonances of the nasal tract and the oral tract is a closed side branch, which causes zeros. Frequently, F_2 is greatly reduced in amplitude, because of a nearby zero; and, in fact, often there is no peak corresponding to F_2.

Nasalization of a vowel is a problem of similar nature. In this case, the nasal cavity is an open side branch, causing extra zeros and extra poles. In a nasalized front vowel, typically, there is an extra small peak slightly above F_1 in frequency. In a nasalized back vowel, the apparent bandwidth of F_1 becomes quite wide, because of a nearby zero, and sometimes there is no peak for F_1.

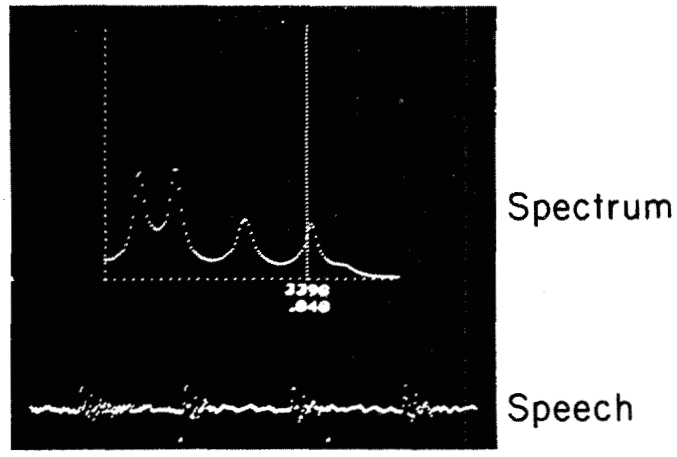
Fig. 2 shows some examples of the problems discussed above.

HIGHLIGHTS OF THE ALGORITHM

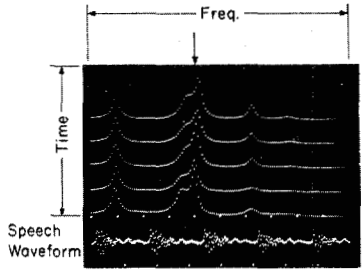
Continuity is one of the strongest constraints that one can rely on in tracking formants. In general, one would expect the frequencies of F_1, F_2, and F_3 in the current frame to be near where they were in the previous frame, because the articulators cannot move too much in 5 ms. However, it is dangerous to rely too heavily on continuity for two reasons: 1) formant frequencies can change considerably within 5 ms, as at the boundary between a nasal and a vowel, and 2) if there is a single bad frame, for instance, a frame mistakenly labeled voiced, it might cause bad decisions in all of the following frames. Therefore, the algorithm uses continuity in an initial decision, but later applies another approach in situations where continuity has failed to yield reasonable results.

In order to lessen the possibility of getting off on the wrong track, it was decided to begin in a region where formants are most likely to be correct, and to branch from there towards the less certain areas. The approach is, therefore, to find an "anchor point" in the middle of each vowel, and to branch from there in both directions in time, using the most recently found formant frequencies as a guide in determining the formants in the current frame.

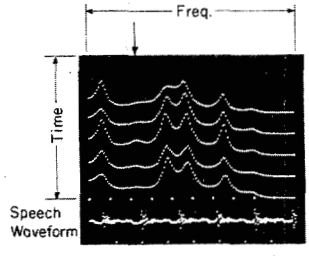
Another feature of the algorithm is that, in the processing



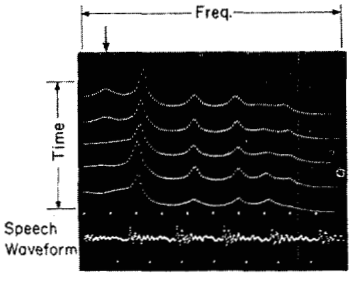
(a)



(b)



(c)



(d)

Fig. 2. (a) Typical linear prediction spectrum (in vowel | a |). (b) Five sequential spectral cross sections in | r | of "there" showing merger of F_2 with F_3. (c) Five sequential spectral cross sections in | ae | of "man" showing small extra peak between F_1 and F_2 due to nasalization. (d) Five sequential spectral cross sections in | ɔ | of "on" showing nasalization of F_1.

at each frame, the formants are dealt with in parallel rather than in series. A typical peak-picking formant tracker would make a decision on F_1 and then eliminate that peak as a possible candidate for F_2. Then, after a peak was selected for F_2, that peak could not be later called F_3. The approach here is to fill available formant slots with available peaks, initially allowing each peak to fill more than one slot. Then in a later step, such duplicates could be dealt with, again in a symmetric way. Thus, if a peak met the criteria for F_1, but met even better the criteria for F_2, it would eventually be called F_2.

SEGMENTATION

Since the formants are tracked only in vowel-like sounds, and processing is begun in the middle of each vowel, it is necessary to have some form of segmentation of the speech waveform to determine which frames are voiced, and where to mark the anchor points.

The first step is to eliminate frames whose total spectral energy is below a threshold for silence. Then, the Gold–Rabiner pitch detector [3] and a ratio of the low frequency energy to the high frequency energy are used in combination to determine whether a frame is voiced. Each resulting voiced region is separated into "vowel" and "not vowel" on the basis of the total energy in the spectrum and the energy in the region from 640 to 2880 Hz. If either of these energy functions has a sufficiently deep valley between two peaks, then a boundary is marked between the valley and each surrounding peak at the place when the slope of the energy function is a maximum, thus dividing up the vowel-like regions into high energy voicing (vowels) and low energy voicing (intervocal is voiced consonants).

Fig. 3 shows a flow chart of the anchor point scheme. Processing of the backward branch is begun at the next anchor point, and continued until an unvoiced frame is encountered, or until a frame is encountered which had already been processed by the previous forward branch. Then the forward branch from the same anchor is begun, and continued until an unvoiced frame is encountered, or until a new vowel segment boundary is reached. At that point, processing jumps to the next anchor point, begins again with a backward branch, and so forth, until the sentence is complete.

THE PROCESSING OF EACH FRAME

At each frame one begins with four vacant formant slots, four estimates for the frequencies of the formants, and one, two, three, or four peaks. The task is to fill the slots with the peaks, based on the estimate frequencies, in such a way that spurious peaks and missing peaks can be recognized as such and dealt with. (No special attempt is made to fill the F_4 slot. It only exists to prevent F_4, when it exists, from competing with F_3 for the F_3 slot.)

The six steps of the peak mapping algorithm are listed in Fig. 4 and explained in greater detail below.

Step 1: *Fetch Peaks.* Find the frequencies and amplitude of up to four peaks in the region from 150 to 3400 Hz.

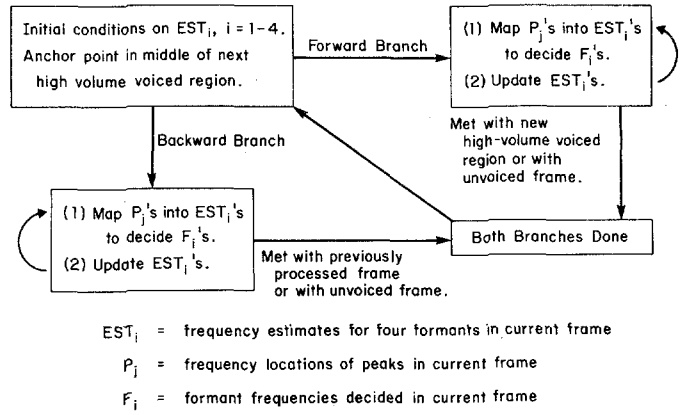


Fig. 3. Flow chart of anchor point scheme.

(1) FETCH PEAKS P_j IN NEXT FRAME.
(2) FILL FOUR FORMANT SLOTS S_i WITH PEAKS USING ESTIMATES EST_i AS GUIDE.
(3) REMOVE DUPLICATE PEAKS.
(4) DEAL WITH UNASSIGNED PEAKS.
(5) DEAL WITH UNFILLED SLOTS.
(6) RECORD ANSWERS AS FORMANTS F_i FOR THIS FRAME AND AS ESTIMATES EST_i FOR NEXT FRAME.

Fig. 4. Six steps to decide formants in each voiced frame.

Step 2: *Fill Slots.* Fill each formant slot S_i, $i = 1$ to 4, with the best candidate peak p_j, by the following rule. The peak p_j closest in frequency to estimate EST_i goes into slot S_i.[1]

The important thing to note is that every slot gets filled in Step 2. If there was only one peak, for instance, it would be put into all 4 slots.

Step 3. *Remove Duplicates.* If the same peak p_j fills more than one slot S_i keep it only in the slot S_k which corresponds to the estimate EST_k that it is closest to in frequency, and remove it from any other slots.

Step 4: *Deal with Unassigned Peaks.* If there are no unassigned peaks p_j, go to Step 5. Otherwise, try to fill empty slots with peaks not assigned in Step 2 as follows.

a) If there is a peak $p_{j=k}$ unassigned, and an $S_{i=k}$ unfilled, fill the slot with the peak and go to Step 5. Or if there is a peak $p_{j=k}$ unassigned, but slot $S_{i=k}$ is already filled, check the amplitude of p_k as follows: if amp $(p_k) < \frac{1}{2}$ amp (peak already assigned to S_k) throw p_k away and go to Step 5. Otherwise, go to (b).

b) If p_k is still unassigned, but $S_{i=k+1}$ is unfilled, move the peak in $S_{i=k}$ to $S_{i=k+1}$, and put p_k in S_k. Go to Step 5.

[1] The EST_i at the anchor point are set to initial conditions as follows:
Male Voices—EST_1 = 320 Hz, EST_2 = 1440 Hz, EST_3 = 2760 Hz, EST_4 = 3200 Hz.
Female Voices—EST_1 = 480 Hz, EST_2 = 1760 Hz, EST_3 = 3200 Hz, EST_4 = 3520 Hz.
These settings were determined empirically, and were found to be reasonable for most speakers. Some experimentation was done with speaker adaptation to determine these initial settings, and in most cases the resulting improvements were minimal.

c) If p_k is still unassigned, but $S_{i=k-1}$ is unfilled, move the peak in $S_{i=k}$ to $S_{i=k-1}$, and put p_k in S_k. Go to Step 5. If a), b), and c) all fail, throw p_k away.

Step 5: *Deal with Unfilled Slots.* If S_1, S_2 and S_3 are all filled, go to Step 6. (F_4 may or may not be filled.) *Otherwise*: Recompute the spectrum on a circle with radius less than one to enhance the formants and hopefully separate two merged peaks. Go to Step 1.

The enhanced spectrum is computed initially with $r = 0.98$. If the spectrum fails to yield a peak to fill the empty slot, then Steps 1–5 are repeated again with $r = r - 0.004$. The radius is shrunk repeatedly in this manner until a peak is finally found or until $r = 0.88$; at which point it is assumed that no peak exists to fill the empty slot.

Finally, whether or not enhancement has succeeded, the amplitudes of the peaks are reset to the amplitudes in the original spectrum. In addition, if the empty slot was S_3, and enhancement failed to yield a peak, then the peak in S_4 is moved down to S_3, assuming that F_3 was mistakenly called F_4.

Step 6: *Record Answers.* Accept formant slot contents as answers for this frame. Also, use formant slot contents as estimates for next frame. (If a slot is empty, keep the original formant estimate for that formant.)

Fig. 5(a) shows how the algorithm would work in a typical situation, as in a vowel. In Fig. 5(b), three pathological cases are illustrated. The frequency locations of the peaks and of the estimates are indicated by an x in the diagrams. An arrow from a peak to an estimate indicates that that peak would fill the formant slot corresponding to that estimate in Step 2. The double line through the arrow indicates that the peak was removed from the corresponding slot in Step 3. In the "Formant Merger" example, enhancement would be called upon in Step 5 to yield a new peak. In the "Rapid Formant Motion" example, Step 4 (b) would move peak 3 up into the vacant slot 3, and would put peak 2 into the now vacant slot 2. In the "Spurious Peak" example, even if peak 2 passed the amplitude test in Step 4 (a), it would still be thrown away, as no slot was available for it.

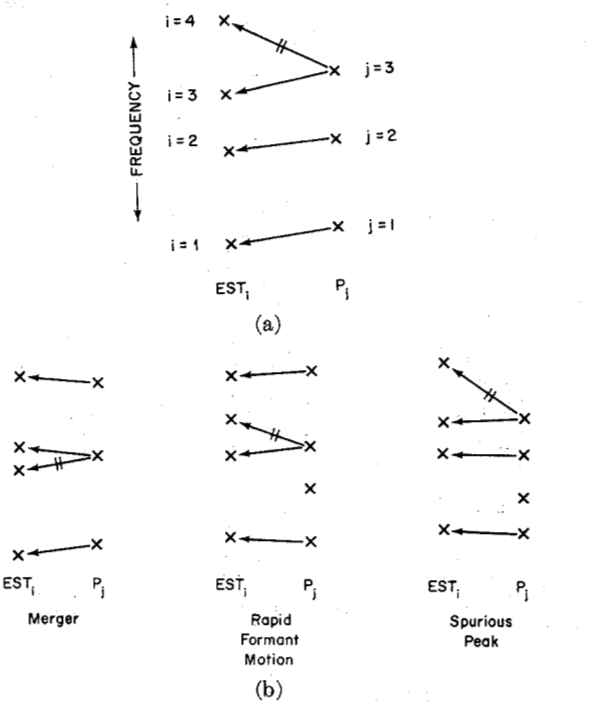


Fig. 5. (a) How the algorithm would work in a typical case. (b) Three pathological cases.

FINAL SMOOTHING

After Steps 1–6 above have been applied at each voiced frame in the sentence, to yield three formant tracks, each formant track should be smoothed separately in some way. The approach is to first correct any obvious gross errors, and to then send each track through a simple zero phase filter. However, each of these steps is done with caution because it is undesirable to 1) attempt to make smooth tracks in a region where they are very bumpy, as these tracks would convey false information (for instance, in a segment which was mistakenly labeled voiced), 2) smooth out sudden shifts in formant frequencies (for instance, at the junction of a vowel and a nasal), or 3) grossly alter good data by smoothing it with bad data that happen to be adjacent to it (for instance, at the boundary between a vowel and a burst, if the onset of voicing had not been accurately determined).

Therefore, strong constraints are required for both the gross corrections and the smoothing filter. Unaligned frames are sligned by interpolation only in regions where the 4 frames surrounding the faulty frame(s) are relatively smooth. In addition, the output of the zero phase filter is not written over the original formant frequency in any frame when the output frequency is sufficiently different from the input frequency. The result is that the formant tracks will become very smooth where they were already fairly smooth; but sudden changes in formant frequency will be retained; and any region where formant tracks were too bumpy will remain untouched.

The following is a detailed discussion of the final smoothing algorithm.

1) If a formant is missing in a single frame, fill in its frequency and its amplitude with the average of the values in the previous and the following frames.

2) If a formant is grossly out of line or missing in one, two, or three frames, but well aligned in the two previous and two following frames, correct the misaligned frames by interpolation as follows.

Let the frequency location of formant F_i in the nth frame be L_n.

Define $D_{a,b} = L_a - L_b$, a measure of the alignment of a particular formant. θ = the threshold = 240 Hz. If $D_{n,n-1} < \theta$, frame n is considered smooth. If $D_{n,n-1} > \theta$, an attempt is made to smooth frame n, but only if either a), b) or c) is true.

a) If $D_{n-1,n-2} < \theta$, $D_{n+1,n-1} < \theta$, and $D_{n+2,n+1} < \theta$, then replace L_n with $(L_{n+1} + L_{n-1})/2$, and move to frame $n + 1$. (One frame out of line.)

b) If $D_{n-1,n-2} < \theta$, $D_{n+2,n-1} < \theta$, and $D_{n+3,n+2} < \theta$, then replace L_n with $(L_{n+2} + L_{n-1})/2$, and move to frame $n + 1$. (Two frames out of line.)

c) If $D_{n-1,n-2} < \theta$, $D_{n+3,n-1} < \theta$, and $D_{n+4,n+3} < \theta$, then replace L_n with $(L_{n+3} + L_{n-1})/2$, and move to frame $n + 1$. (Three frames out of line.) The new L_n is used in evaluating frame $n + 1$.

3) Smooth each formant track twice using the following filter:

$$F_i'(n) = \tfrac{1}{4}F_i(n-1) + \tfrac{1}{2}F_i(n) + \tfrac{1}{4}F_i(n+1),$$

but only at those frames where $|F_i'(n) - F_i(n)| < 100$ Hz.

RESULTS

Fig. 6 shows the results of the formant tracking algorithm for the phrase "average uranium lead ratio" spoken by a woman. In Fig. 6(a) the first three peaks are written over the spectrogram and also shown as three separate functions above the spectrogram. In Fig. 6(b), the first three formants as determined by the algorithm, are shown in a similar way for comparison. The three functions above are shown on a greatly constricted scale, and are useful mainly for determining which formant number was associated with each peak, and for identifying gross errors. It can be noted that in the | r | of "uranium" and in the | i | of "ratio," F_3 had merged with F_2, and was found by enhancement.

Fig. 7 shows a similar comparison for the sentence, "the box was thrown beside···," spoken by a man. F_1 had merged with F_2 in the | a | of "box," showing up as a discontinuity in the P_1 and P_2 curves, and a vacancy in the P_3 curve. The algorithm was able to recognize that it was F_1 that was missing and to find it by enhancement. Similarly, in the | n | of "thrown," F_2 was missing and was recovered by the algorithm through enhancement.

The algorithm is very successful in recognizing peak merger problems and the missing formants are nearly always recovered by enhancement. However, the results for nasalized vowels and nasals are not always so predictable. In some nasalized back vowels, F_1 had to be recovered through enhancement. In those cases, F_1's frequency was often observed to be abnormally high. Occasionally in nasalized back vowels, an extra peak appeared between F_2 and F_3, which was sometimes tagged as F_3. In nasalized front vowels there was frequently an extra peak slightly above F_1. Sometimes this extra peak was called F_2 by the algorithm, and the actual second formant was called F_3. In nasals, enhancement had to be called upon quite often to recover F_2. In some nasals, enhancement failed to find a peak for F_2, but F_1 and F_3 were nearly always correctly identified.

A fully automatic implementation of the algorithm runs in 2 or 3 times real-time on a computer facility at Lincoln Laboratory consisting of a Univac 1219 and the Fast Digital Processor (FDP), a programmable processor [9] designed for real-time digital signal processing. The

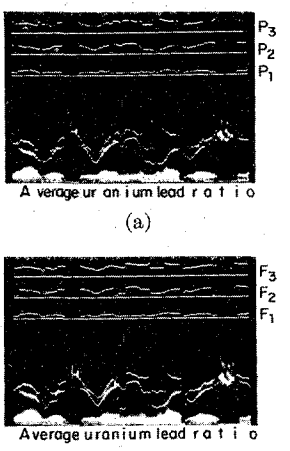


Fig. 6. (a) First three peaks in each voiced frame written over spectrogram and shown above it as three separate functions. (b) First three formants, as computed by algorithm, shown as in Fig. 6(a) for comparison.

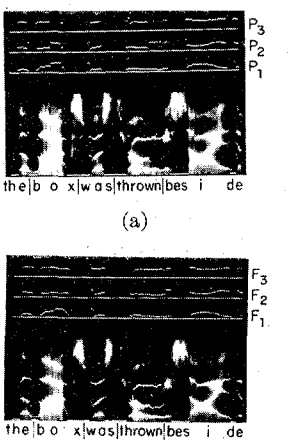


Fig. 7. (a) First three peaks in each voiced frame written over spectrogram and shown above it as three separate functions. (b) First three formants, as computed by algorithm, shown as in Fig. 7(a) for comparison.

linear prediction analysis is done entirely on the FDP in less than real-time. The time consuming part of the formant tracking algorithm is the iterative computation of enhanced spectra to find missing formants. We have used a conservative approach in shrinking the radius by very small increments, because the FDP is fast enough that time is not an issue. One could, to save time, use enhancement only at a few different radii, for instance, 0.96, 0.92, and 0.88, with the results being not as good as the results here, but still much better than no enhancement at all.

Statistics were collected for some 50 sentences on how often the various correction measures were necessary. The statistics showed that it is much more common in linear prediction spectra for a peak to be missing than for a spurious peak to exist. Although statistics varied considerably from sentence to sentence, on the average enhancement was tried in about 15 percent of the voiced frames. Out of these, a peak was found through enhancement in 9 out of 10 cases. In the remaining cases, either

the frame was mistakenly labelled voiced, or the formant was too strongly cancelled by a nearby zero (in nasals and nasalized vowels); or, rarely, a peak merger was not resolved.

In 1 percent of the voiced frames, F_3 was mistakenly callee F_4 initially, and was later moved to the F_3 slot after enhancement failed to yield a peak. In another 3 percent continuity constraints had failed in the initial slot-filling steps, and peaks had to be moved to new slots in Step 4 to accommodate a peak about to be thrown away. An equal number of peaks *were* thrown away in Step 4 either because they failed to pass the amplitude test or because there was no slot available for them. These extra peaks were usually due to nasalization effects.

About 3 percent of the time the output of the smoothing filter was not written over the original formant value, either because of a sharp shift in formant frequency or because of an irregular trajectory, as in a voiced fricative.

Second pass corrections of gross errors were rare, occurring only about 1.5 percent of the time.

SUMMARY

A completely automatic algorithm has been developed which yields the first three formants during all voiced sounds in continuous unrestricted speech. It uses the peaks of the linear prediction spectra as input, and resolves peak mergers and spurious peak problems in a systematic way. It has been tested on a large number of sentences spoken by several different speakers. The algorithm was found to be extremely successful with nonnasalized sounds. The results in nasals and nasalized vowels were not as predictable, but were still quite good most of the time.

Work is now beginning on the application of the output of the formant tracking algorithm to both a speech understanding project and a formant vocoder.

REFERENCES

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.
[2] C. G. Bell *et al.*, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1725–1736, 1961.
[3] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, 1969.
[4] J. Makhoul and J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt, Baranek, and Newman, Inc., Cambridge, Mass., Rep. 2304, 1972.
[5] J. D. Markel, "Digital inverse filtering—a new tool for formant trajectory estimation," Speech Com. Res. Lab., Santa Barbara, Calif., Monograph 7, Oct. 1971.
[6] J. P. Olive, "Automatic formant tracking by a Newton–Raphson technique," *J. Acoust. Soc. Amer.*, vol. 50, pp. 661–670, 1971.
[7] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–184, 1952.
[8] R. W. Schaefer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 637–648, 1970.
[9] B. Gold *et al.*, "The FDP, a fast programmable signal processor," *IEEE Trans. Comput.*, vol. C-20, pp. 33–39, Jan. 1971.

Lumped Parameter Electromechanics of Electret Transducers

THOMAS B. JONES, MEMBER, IEEE

Abstract—**A lumped parameter formulation for the electromechanics of a class of electret transducers is developed. An energy method is used to calculate the force of electrical origin and the technique is employed in the solution of two important electret transducer geometries. The results are checked by a surface integration of the Maxwell stress tensor.**

I. INTRODUCTION

SINCE the initial discovery of the electret, attributed principally to Eguchi [1], numerous applications have been found for these permanently electrified dielectrics. Prominent among these applications are electret condenser microphones [2], [3], and electrostatic voltmeters and motors proposed by Jefimenko [4]–[6]. Basically, an electret is a solid dielectric upon which a permanent electrification has been impressed. Two distinct types of electrification have been identified, though both are apt to coexist in a given electret. One, the so-called *homo-charge* electrification, is due to the transfer of charge to the electret surface from the charging electrode. The other type, referred to as *hetero-charge* electrification, is a volume effect (*remanent polarization*, analogous to *remanent magnetization* in a permanent magnet).

Calculations of electrostatic forces in electrets have been performed before [5], [7]. However, a rigorous calculation of these forces using the standard techniques of lumped parameter electromechanics is not found in the literature. The purpose of this paper is, then, to develop a unified formulation of the lumped parameter electromechanics of simple yet practical electret transducers. The formulation relies on a modification of classical energy

Manuscript received June 11, 1973; revised October 17, 1973.
The author is with the Department of Electrical Engineering, Colorado State University, Fort Collins, Colo. 80521.