A NOVEL METHOD OF FORMANT ANALYSIS AND GLOTTAL INVERSE FILTERING

Steve Pearson

Panasonic Technologies, Inc. / Speech Technology Lab
Santa Barbara, CA 93105

ABSTRACT

This paper presents a class of methods for automatically extracting formant parameters from speech. The methods rely on an iterative optimization algorithm. It was found that formant parameter data derived with these methods was less prone to discontinuity errors than conventional methods. Also, experiments were conducted that demonstrated that these methods are capable of better accuracy in formant estimation than LPC, especially for the first formant. In some cases, the analytic (non-iterative) solution has been derived, making real time applications feasible. The main target that we have been pursuing is text-to-speech (TTS) conversion. These methods are being used to automatically analyze a concatenation database, without the need for a tuning phase to fix errors. In addition, they are instrumental in realizing high quality pitch tracking, and pitch epoch marking.

1. BACKGROUND

The motivation behind this investigation is to implement a concatenation type TTS synthesizer based on a source-filter model of speech production. We favor a model that most closely represents the naturally occurring degree of separation between the human glottal source and the vocal tract filtering effect [1,2]. In addition, certain characteristics of the filter parameters and source waveform can be specified which are desirable for the synthesis process. These factors can lead to measurable criteria that can be applied to the filter parameters and source waveform. Our analysis method is essentially an iterative algorithm based on minimizing a cost function which expresses the above criteria.

In particular, we want to parameterize a concatenation unit waveform database automatically. Currently, our target is to decompose the data into a residual source waveform and an all-pole vocal tract model.

We can enumerate the key features required: (1) remove resonances from the residual to improve prosodic modification, (2) obtain smooth and accurate formant parameters for resynthesis, (3) develop a pitch synchronous analysis method, with pitch epoch marking, (4) make it totally automatic, although high speed analysis is not necessary.

This problem has previously been attacked in many ways, and clearly the primary tool has been LPC, however, despite trying out many of these methods, none were found that were adequate for our particular requirements. As is well known, it was found that LPC was often inaccurate in the low formant parameter estimates which translates into visible resonances in the residual. Also, LPC was found to give many discontinuities in the parameters. This basically implied hand correction of all the data. There are algorithms for fixing the formant discontinuities (e.g. [3], [4]), however it was decided to also look for an alternate solution.

A good alternate is analysis by synthesis which simultaneously optimizes a parametric voice source and vocal tract modeling filter ([5], [6], [7]), however, we wished to avoid the parametric source model assumption.

Hence, a slightly different approach was taken: optimize the filters to minimize a cost function applied to the residual, while imposing a constraint on the filters. Clearly LPC can be seen to be in this class of methods. The mean square linear prediction error is minimized while requiring that the inverse filter has a zero'th order coefficient of unity (i.e. the transfer function has unity gain at the infinity point). Translated into the frequency domain, LPC flattens (or whitens) the spectral magnitude of the residual. This may not be the best cost function.

The goal was to find some alternative cost functions and filter constraints that were more appropriate to our requirements. Several cost functions were found which can be expressed both in time domain or frequency domain. They have a commonality in that they derive from the arc-length of some curve. Thus, these will be referred to as arc length minimization (ALM) techniques. In general, iterative optimization methods were used, but in two cases, a non-iterative analytic solution was found. The methods will be described in detail, as well as some experiments to quantify their performance.

2. FORMANT ANALYSIS METHODS

All of the analysis methods described in this section follow the same general principle. First we assume a source-filter model of speech production. The filter is assumed to be controlled by a set of parameters which may have an additional constraint applied, and a cost function is defined on the source. In the analysis procedure, speech is filtered with the inverse of the given filter, and the residual is assumed to be an approximation of the source. The object is to minimize the cost function applied to the residual, while searching the filter parameter space. The set of minimizing filter parameters is the analysis output. The difference between the various methods is only the difference in cost function and filter constraint.

The analysis is repeated on subsequent frames of speech waveform data using a moving window scheme. All of the described methods were executed pitch synchronously during voicing, and with a fixed step size and window size during unvoiced speech. The initial estimate of the pitch epoch time points in the current frame is determined by autocorrelation and peak picking in the residual waveform. The estimate is improved iteratively as the inverse filter is improved.

Several analysis windows were tried: square window, Hamming window, and Hanning window. The square and Hamming window gave poor results, hence subsequently only the Hanning window was used. In particular, the Hanning window was modified to be asymmetric. It was centered on the current pitch epoch, and reached zero at adjacent pitch epochs, thus covering two pitch periods. An additional linear multiplicative component was included to compensate for increasing or decreasing amplitude in the voiced speech signal.

In the current implementation, an exhaustive search is approximated by a steepest descent search algorithm. This brings up the problems of starting points and local minima. As pointed out by Olive [8], using the previous frames solution as a starting point can encourage continuity, but also may give a sub-optimal solution. For the experiments described in this paper, several starting points were tried, including the previous frame and neutral values, and the best result was saved.

In some of the methods it was possible to introduce heuristic smoothing which eliminated some of the local minima (this will be described in more detail later). Also, it was possible to make a rough qualitative evaluation of each method in terms of the extent to which they were plagued by local minima.

All of the methods are further related by a common element. In each, the core of the cost function is an arc-length. We use the standard arc-length

$$arc-length = \sum_{n=1}^{N} \sqrt{(x_n - x_{n-1})^2 + (y_n - y_{n-1})^2}$$

but also the related sum of square lengths

$$square-length = \sum_{n=1}^{N} \left\{ (x_n - x_{n-1})^2 + (y_n - y_{n-1})^2 \right\}$$

where (x_n,y_n) is a sequence of points.

Now we can list the cost functions that were investigated.

1. arc-length of windowed residual waveform vs. time
2. square length of windowed residual waveform vs. time
3. arc-length of log spectral magnitude of windowed residual vs. mel frequency
4. arc-length in z-plane of complex spectrum of windowed residual parametrized by frequency
5. square length in z-plane of complex spectrum of windowed residual parametrized by frequency
6. arc-length in z-plane of complex log of the complex spectrum of windowed residual parametrized by frequency

The last four are computed in the frequency domain using an FFT of adequate size to compute the spectrum.

For example, for (6), if $Y_n = R_n*\exp(j*\theta_n)$ is the FFT of size N,

$$cost = \sum_{n=1}^{N} \sqrt{\log^2 (\frac{R_n}{R_{n-1}}) + (\theta_n - \theta_{n-1})^2}$$

The filter used in this investigation is a cascade of second order poles and zeros which are parameterized by frequency and bandwidth. The inverse filter is the same structure with poles and zeros reversed. It was found that some of the methods were able to track zeros in a speech signal, but with more difficulty. For simplicity, in this paper we assume that the inverse filter is a cascade of zeroes, whose transfer function has the polynomial expansion with coefficients A_0, A_1, ... A_M.

The filter constraints that were investigated were the following:

A. the filter has unity gain at DC (or zero frequency) equivalently, sum A_i = 1

B. the transfer function has unity value at the infinity point equivalently, (as in LPC) A_0 = 1

C. the filter output is normalized so that the maximum magnitude is constant

In the cost functions that included the log magnitude spectrum, it was found that smoothing could eliminate some problems with local minima by eliminating the effects of harmonics or sharp zeros. The smoothing functions considered were 3, 5, and 7 point FIR, LPC and Cepstral smoothing, and a heuristic smoothing that removed dips. The latter was implemented as follows: in 3, 5, or 7 point windows in the log magnitude spectrum, low values were replaced by the average of two surrounding higher points, or if the higher points did not exist the target point was left unchanged.

3. PITCH TRACKING AND EPOCH MARKING

These methods are inherently pitch synchronous, hence an initial estimate of pitch epochs is required. But in addition, since our target is TTS synthesis, it is desirable, for the purpose of prosodic modification, to have very accurate pitch epoch marking. It turns out that the methods being described work fairly well to improve on pitch extraction and epoch marking.

The pitch tracking works best by applying the above cost function (1) and constraint (C). This smoothes out the residual waveform, but maintains the size of the pitch peak. The autocorrelation can then be applied, and is less likely to suffer from higher harmonics.

The residual waveform peak is sometimes a consistent approximation to the pitch epoch, however often this peak is noisy or rough causing inaccuracies, hence we need a better method. It was observed that, when the inverse filter was successful in canceling the formants, the phase of the residual approached a linear phase (at least in the lower frequencies). If the origin of the FFT analysis is centered on the approximate epoch time, the phase becomes nearly flat.

This suggests a method where the epoch point becomes one of the parameters in the minimization space when the cost function includes phase. The cost functions (3), (4), and (5) above include phase, hence in these cases the epoch time was included as a parameter in the optimization. The results were very consistent epoch marking, except when the speech signal was low. It also turns out that the accuracy of estimating formant values for the frequency domain cost functions is greatly improved by the simultaneous optimization of the pitch epoch point and corresponding alignment of the analysis window.

4. PERFORMANCE EVALUATION

Two aspects of analysis performance are of most interest to us: the frequency of occurrence of discontinuity errors, and the accuracy of formant frequency and bandwidth estimation. The former is actually more important since this impacts our ability to rely on totally automated methods. Discontinuity errors show up as very noticeable artifacts in resynthesis. Unfortunately, we did not find a quantitative way to describe this type of problem, thus only a qualitative description will be given.

Informally, it was observed that ALM methods had far fewer discontinuities of filters switching formants than LPC. In fact, the method using cost function (4) and filter constraint (A) never had this type of discontinuity during voicing except occasionally at the articulator closure point between a vowel and a voiced stop, voiced fricative, or nasal. A discontinuity might be expected at these points, and further, the discontinuities did not adversely affect the resynthesis. Overall, it could be stated that the ALM methods had a much more continuous and smooth formant tracking.

To evaluate accuracy, two spectral distance measures were implemented, and a comparison test was run on synthetic speech. The first measure is based on the distance, in the z-plane, between the target pole and the pole that was estimated by the analysis method. The distance was calculated separately for formants one through four, and also for the sum of all four, and was accumulated over the whole test utterance.

The second measure is the (spectral peak sensitive) Root-Power Sums (RPS) distortion measure [9], defined by

$$dist = \sum_{k=1}^{N} (k \cdot (c1_k - c2_k))^2$$

where $c1_k$ and $c2_k$ are the kth ceptsral coefficient of the target spectrum and analyzed spectrum respectively, and N was chosen large enough to adequately represent the log spectrum.

The analysis was performed on a completely voiced sentence, “Where were you a year ago?” which was produced by a rule based formant synthesizer. Several words were emphasized to cause a fairly extreme intonation pattern. The formant synthesizer produced six formants, and each analysis method tracked six, however, only the first four formants were considered in the distance measures. The known formant parameters from the synthesizer served as the target values.

For reference, the sentence was analyzed by standard LPC of order 16, using the autocorrelation estimation method. The LPC was done pitch synchronously, similar to the other methods, and the window was a Hanning window centered on two pitch periods. Formant modeling poles were separated from source modeling poles by selecting the stronger resonances (i.e. narrower bandwidths). The LPC analysis made several discontinuity errors, but for the accuracy measurements, these errors were corrected by hand by reassigning formants.

Any combination of cost function and filter constraint can be used for analysis, however some of these combinations give very poor results. The non-productive combinations were eliminated from consideration. Combinations that performed fairly well are listed in Table 1, to be compared with themselves and LPC. The scale or units associated with these numbers is arbitrary, but the relative values within a column are comparable.

	Z-plane pole distance for formants					
	1	2	3	4	sum	RPS
LPC	3.57	3.24	2.93	3.63	13.4	17.6
1C	9.32	5.45	4.73	5.07	24.6	81.1
1A	4.51	5.86	5.63	7.03	23.0	38.7
2A	11.80	11.08	6.56	9.54	39.0	115.0
3A	2.12	2.43	1.81	2.07	8.4	12.2
4A	1.26	2.37	2.32	2.83	8.8	11.1
4B	3.22	7.82	4.98	4.13	20.2	46.7
5A	1.57	4.13	4.27	8.30	18.3	24.8
6A	1.23	2.88	2.51	2.84	9.5	7.6

Table 1. Error measurement of analysis methods. Methods are named by cost-function number and constraint letter.

Assuming that these distance measures are valid, we conclude generally that the cost functions based in the frequency domain and using the DC unity gain constraint outperform LPC in accuracy. Especially noticeable is their improvement to accuracy in the first formant.

One might conclude that methods (3A), (4A), and (6A) are equally likely candidates for an analysis application, however there is further factors to be considered. This concerns local minima and convergence. Methods (3A) and (6A), which involve the logarithm, are much more likely to encounter local minima and converge more slowly. This is unfortunate since these are the most likely to also track zeros.

Methods (4A) and (5A) rarely encounter local minima, in fact, no local minima has yet been observed for method (5A). On the other hand, these methods tend to estimate overly narrow bandwidths. Hence, for these, a small penalty was added to the cost function to discourage overly narrow bandwidths. Athough method (5A) is inferior overall, it may be very useful since it accurately tracks formant one with faster convergence and no local minima.

5. ANALYTIC SOLUTION

It was found that cost function (5) could be solved analytically for both constraint A and B. Likewise an approximate analytic solution was found for (4A) and (4B). This may be important for gaining speed and reliability for these methods. The solution is analogous to that for LPC, and ends up with inversion of an autocorrelation based matrix.

For the case of cost function (5), define

$$P_{i,j} = \sum_{k=0}^{N-1} x_{k-i} \cdot x_{k-j} \cdot (1 - \cos(\frac{2\pi(k - cntr)}{N}))$$

where x_n is the residual waveform, M is the order of analysis, N is the size in points of the analysis window, and cntr is the estimated pitch epoch sample point index.

Then if A_i is the sequence of inverse filter coefficients, and B_i is a sequence of constants defining a linear constraint on the coefficients A_i, such that B_0*A_0 + ... + B_M*A_M = 1, then A_i can be solved in the following matrix equation:

$$\left[\begin{array}{c} B_0 \;\; B_1 \;\; B_2 \quad \dots \quad B_M \\ \\ \left[\begin{array}{c} P_{j,n} - B_j * P_{o,n} \\ \text{for } j=1,\dots\ M \end{array}\right] \end{array}\right] \left[\begin{array}{c} A_0 \\ A_1 \\ \\ A_M \end{array}\right] = \left[\begin{array}{c} 1 \\ 0 \\ \\ 0 \end{array}\right]$$

Setting B_i = 1 for i = 0, ...M gives constraint (A). Setting B_I = 1, and B_i = 0 for i = 1, ... M gives constraint (B).

To find an approximate solution for cost function (4), in the above matrix equation, replace $P_{i,j}$ by:

$$P_{i,j} = \sum_{k,l=0}^{N-1} \left\{ x_{k-i} \cdot x_{l-j} \cdot (\cos(\pi \frac{k-l}{N}) - \cos(\pi \frac{k+l-2 \cdot cntr}{N})) \cdot S_{k-l} \right\}$$

$$where:$$

$$S_m = \sum_{n=0}^{N/2-1} \left\{ (n+1)^a \cdot \cos(2\pi \frac{(n+0.5)m}{N}) \right\}$$

In this equation, the term, $(n+1)^{\alpha}$, represents an idealized source. When α = 0, the equation reduces to that of cost function (5). Setting α = 2 gives approximately equivalent results to cost function (4).

6. DISCUSSION

A possible explanation of why the frequency domain cost functions work well is as follows. This method focuses on the effect of a resonance filter on an ideal source. The ideal source has linear phase and smoothly falling spectral envelope. A resonant filter would cause a circular detour in the otherwise short path of the complex spectrum. The arc-length minimization aims at eliminating the detour by using both magnitude and phase information.

In comparison, LPC assumes a white source, and equivalently tries to flatten the magnitude spectrum, and does not take phase into account. Thus, it predicts resonances to model the source characteristics.

7. SUMMARY

We have proposed an iterative formant analysis method, based on minimizing the arc-length of various curves, and under various filter constraints. Careful trials have shown the method to overcome some of the pitfalls of LPC. In particular, an experiment has shown the methods to be more accurate than LPC in estimating formant frequencies. An analytic solution was shown for two of the particular methods.

For future directions, there are a number of things that should be done. First it would be interesting to find analytic solutions to more of the methods. But it is suspected that this will require series approximation. Secondly, it would be useful to evaluate these methods with additional voices and with noisy data.

REFERENCES

1. Pearson, Javkin, Matsui, Kamai, “Text-to-Speech Synthesis Using a Natural Voice Source”, Proc. ICSLP, Kobe, Japan, 1990.

2. Matsui, Pearson, Hata, “Improving Naturalness in Text-to-Speech Synthesis using Natural Glottal Source”, Kamai, Proc. ICASSP 2.769-772, 1991.

3. Program “formant”, ESPS Software Manual, Entropic Research Labs., Washington D.C.

4. Kopec, G. “Formant tracking using hidden markov models and vector quantization”. IEEE ASSP-34, 1986.

5. J. de Veth, W. van Golstein Brouwers, L. Boves, “Robust ARMA Analysis for Estimation of Vocal Tract Parameters”, Proc. FASE-88, pp. 305-312, 1988.

6. Fujisaki H., Ljungqvist M., “Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform”, Proc. ICASSP 1987, pp. 637-640.

7. Ding, W., Kasuya, H. “A novel approach to the estimation of voice source and vocal tract parameters from speech signals”, Proc. ICSLP 1996.

8. Olive, J. “Automatic formant tracking by a Newton-Raphson technique”, JASA vol. 50, no. 2, pp661-670, 1971.

9. Hanson B., Wakita H., “Spectral Slope Based Distortion Measures for All-Pole Models of Speech”, Proc. ICASSP-86, pp 757-760.