2020-11-24

Power-Weighted LPC Formant Estimation

Ruairí de Fréin
Dublin Institute of Technology, ruairi.defrein@tudublin.ie

Recommended Citation

de Fréin, R. (2020). Power-Weighted LPC Formant Estimation. *IEEE Transactions on Circuits and Systems II: Express Briefs.* doi: 10.1109/TCSII.2020.3040194.

Funder: Science Foundation Ireland

Power-Weighted LPC Formant Estimation

Ruairí de Fréin
Technological University Dublin,
Ollscoil Teicneolaíochta Bhaile Átha Cliath,
Ireland
web: https://robustandscalable.wordpress.com

in: IEEE Transactions on Circuits and Systems II: Express Briefs. See also BibTeX entry below.

BibTeX:

```
@article{deFrein21Power,
  author={Ruair\'{i} de Fr\'{e}in},
  journal={IEEE Transactions on Circuits and Systems II: Express Briefs},
  title={Power-Weighted LPC Formant Estimation},
  year={2020},
  volume={},
  number={},
  pages={1-1},
  doi={10.1109/TCSII.2020.3040194},
  url={https://ieeexplore.ieee.org/document/9268171},}
```


document created on: January 11, 2021
created from file: rdefreinTCASIIcoversheet.tex
cover page automatically created with CoverPage.sty
(available at your favourite CTAN mirror)

Power-Weighted LPC Formant Estimation

Ruairí de Fréin

***Abstract*—A power-weighted formant frequency estimation procedure based on Linear Predictive Coding (LPC) is presented. It works by pre-emphasizing the dominant spectral components of an input signal, which allows a subsequent estimation step to extract formant frequencies with greater accuracy. The accuracy of traditional LPC formant estimation is improved by this new power-weighted formant estimator for different classes of synthetic signals and for speech. Power-weighted LPC significantly and reliably outperforms LPC and variants of LPC at the task of formant estimation using the VTR formants dataset, a database consisting of the Vocal Tract Resonance (VTR) frequency trajectories obtained by human experts for the first three formant frequencies. This performance gain is evident over a range of filter orders.**

***Index Terms*—Least Squares Methods, All-Pole Filter, Spectral Estimation, Power Weighted Estimators.**

I. Introduction

LINEAR Predictive Coding (LPC) is a widely used approach for modelling the vocal tract. It learns a time-varying linear digital filter [1]. LPC extracts a set of parameters from the signal which specify the filter transfer function that best models the signal [2]. An all-pole filter of order p, typically in the range of $10 \leq p \leq 20$ is used for speech [3]. The spectral envelope of short-term speech contains peaks at frequencies related to the formant frequencies. Formants are the resonant frequencies of the vocal tract. The problem is that LPC does not always extract the correct formant frequencies. We introduce LPC analysis and outline why. Our objective is to improve the accuracy of formant frequency estimation of speech by LPC when the formants are an arbitrary set of frequencies which are less than the Nyquist frequency and when the signal is corrupted by noise. We contribute Power-Weighted (PW) LPC estimators that achieve this objective. To demonstrate the improvement achieved we perform estimation when the true frequencies are known, to exactly quantify the improvement, and then formant frequency estimation on real speech where formants are estimated by experts.

We chart the progress of applications that address LPC. LPC has been influential in the field of speech coding for the past 40 years. Since the adoption of the LPC10 standard [3], LPC has had a central role in the development of present day audio codecs such as MPEG-4 ALS [4], FLAC, SILK audio codec (developed by Skype), and other lossless codecs. During the initial wave of interest in LPC, notable advances included multi-pulse LPC [5] and Code Excited LPC (CELP) [6]. LPC is the first step of many main-stream codecs which are based on CELP, therefore, we use it as a baseline in our evaluation. Its role in CELP is now summarized: (1) LPC is applied to speech and the effect of the LPCs are filtered out producing a residual signal; (2) fundamental frequency estimation is performed on the residual and the fundamental is then also removed; finally, (3) an analysis-by-synthesis step picks a good quantisation scheme for the residual. By comparing PWLPC with the first step of CELP, we compare its operation with the analysis step of many modern codecs.

The second wave of interest in LPC arose in response to the confluence of the demand for real-time services and the availability of IP networks, and has given rise to approaches that consider robustness to packet loss [7], [8] and processing speed [9]. The task of reducing the bit rate of high quality speech is related to, but not the same as, detecting the formants of speech. Codecs look to explain *all* of the signal using a filter whereas formant frequency estimators search for resonant frequencies of the vocal tract. The dichotomy lies in placing poles at locations to capture the formants and placing poles at frequencies in order to perform good prediction. For example the LPC frequency tracking approach in [10] incorporates state dynamics into LPCs based on the Kalman filter. Their approach is optimal in the minimum mean square error sense, when the signal and noise are jointly Gaussian; in many cases accurate estimates of the formant frequencies are required and the goodness of the least-squares fit might not be as important. The current wave of interest in LPC is due to the requirement for accurate LPC-based formant estimates for machine learning. LPC-based features are used in [11] in a recurrent neural network (Long Short-Term Memory architecture) for formant tracking. Recent weighting schemes for LPC for feature exctraction include Stablized Weighted LPC (SWLP) [12] and Extended Weighted LPC (XWLP) [13]. SWLP weights each value of the squared prediction error by the short-time energy of the previous samples to obtain smoother spectral shapes when learning formants in the presence of zero-mean Gaussian noise. The weighting function used in XWLP is reported to yield improved robustness in feature extraction in speaker verification and automatic speech recognition tasks, however, the authors report that the resulting synthesis filter is not guaranteed to be stable.

The idea of Power Weighting estimators was recently introduced to improve a number of Time-Frequency (TF) domain estimators, including relative attenuation and delay estimation [2] and multi-channel TF methods [14]. These estimators assume that speech is compactly supported in TF. Consequently the energy of a speech signal dominates a few TF components. It follows that these components should have an important role in speech parameter estimation, as these components are where the majority of the speech energy lies. Source extraction methods which are based on second-order statistics are based on linear predictors [15], [16] and could be adapted so that the estimates are power-weighted similar

R. de Fréin is with the School of Electrical and Electronic Engineering, Technological University Dublin, Ireland e-mail: ruairi@kth.se

Manuscript received June 15, 2020; revised July 15, 2020.

to our proposed approach. In contrast, the approach taken in [16] was to introduce a new cost function to account for the presence of noise. Source separation was the motivation for Power-Weighted estimation in [2]; the underpinning idea was to emphasize the desired signal given that its support was compact in TF. Up to now, power weighting has been achieved in the TF domain where the goal is to improve signal separation for relative attenuation and delay estimation; our goal is to improve formant estimation. We introduce a time-domain power-weighting scheme suited to focusing in on the formant frequencies, as opposed to smoothing the spectral shapes, in LPC spectral envelope estimation.

This paper is organized as follows: we introduce notation and examine the effects of sampling a signal in the frequency domain by considering signals that have and do not have frequency components that are in the set of the Discrete Fourier Transform (DFT) frequencies in Section II. In Section III we investigate what happens when LPC is performed on both types of signal. We demonstrate when LPC produces frequency estimates which are inaccurate. Finally we introduce a set of power-weighted estimators in Section IV which decrease this inaccuracy and evaluate them on an annotated database in Section V.

II. Motivation

Sampling in Frequency: We consider discrete time signals and how they are sampled in the frequency domain. The Discrete Time Fourier Transform of x_n is defined as

$$X(e^{j\omega}) = \sum_n x_n e^{-j\omega n}. \tag{1}$$

The variable ω is continuous. It is common to compute $X(e^{j\omega})$ at a finite set of frequencies. Typically N equally spaced samples around the unit circle are chosen (where $\omega_k = \frac{2\pi k}{N}$),

$$X(e^{j\omega_k}) = \sum_n x_n e^{-j\omega_k n}, \tag{2}$$

and $k = 0, 1, \ldots, N-1$. In doing so, we consider N samples of $X(e^{j\omega})$,

$$X_k = X(e^{j\omega})|_{\omega=\frac{2\pi k}{N}} = \sum_{n=0}^{N-1} x_n W_N^{kn} \tag{3}$$

where the term, $W_N = e^{-j\frac{2\pi}{N}}$, simplifies notation. Many speech signals contain approximately $R = 3$ formant frequencies $\{\omega_r\}_{1\le r\le R}$. Generally, they are not members of the set of frequencies $\{\omega_k\}_{0\le k\le N-1}$ used by the DFT. We consider LPC's performance on three synthetic signals where the frequencies are known to evaluate its performance.

Exemplars: Cosines with known frequencies are used to motivate our method. The first signal has a frequency component which corresponds to one of the DFT basis functions; the second signal has a frequency component which lies between two of the DFT basis functions. A third cosine signal is corrupted by Additive White Gaussian Noise. This noise has the effect of adding in multiple frequencies which are not members of the set $\{\omega_k\}_{k=0,\ldots N-1}$. In the first case

$$x_n = .5\left(e^{j\omega_r n} + e^{-j\omega_r n}\right), \quad 0 \le r \le N-1 \text{ and } r \in \mathbb{Z}. \tag{4}$$

The following identity is useful. For any complex number $z \in \mathbb{Z}$, if $z \neq 1$ the finite geometric series $\sum_{n=0}^{N-1} z^n$ may be expressed in closed form as

$$\sum_{n=0}^{N-1} z^n = \frac{1-z^N}{1-z}. \tag{5}$$

The DFT of x_n, $X_k = \text{DFT}\{x_n\}$, at ω_k yields

$$X_k = .5\sum_{n=0}^{N-1}\left(e^{j(\omega_r-\omega_k)n} + e^{-j(\omega_r+\omega_k)n}\right). \tag{6}$$

In the case that $\omega_k = \omega_r$, for one of the frequency indices, k,

$$X_k = \sum_{n=0}^{N-1}\frac{e^{j0n}}{2} + \sum_{n=0}^{N-1}\frac{e^{-j2\omega_k n}}{2} = \frac{N}{2} + \frac{1-e^{-j2\omega_k N}}{1-e^{-j2\omega_k}} = \frac{N}{2}. \tag{7}$$

Given $e^{j4\pi k} = \cos(4\pi k) + j\sin(4\pi k) = 1$, the second term is $\frac{1-1}{1-e^{j2\omega_k}} = 0$. For $\omega_k = -\omega_r$, we get a similar real value, $\frac{N}{2}$. More generally, for $\omega_k \neq \omega_r$ and $r \notin \mathbb{Z}$, it holds that

$$X_k = \frac{1-e^{j\frac{2\pi(r-k)N}{N}}}{1-e^{j\frac{2\pi(r-k)}{N}}} + \frac{1-e^{-j\frac{2\pi(r+k)N}{N}}}{1-e^{-j\frac{2\pi(r+k)}{N}}} = 0. \tag{8}$$

For all other integer values of k, $X_k = 0$ because the basis vectors of DFT form an orthogonal basis. In the second case

$$x_n = .5\left(e^{j\omega_r n} + e^{-j\omega_r n}\right),\ 0 < r < N-1 \text{ and } r \notin \mathbb{Z}. \tag{9}$$

There is no k that admits to $\omega_k = \pm\omega_r$; r is non-integer.

$$X_k = \frac{1-e^{j\frac{2\pi(r-k)N}{N}}}{1-e^{j\frac{2\pi(r-k)}{N}}} + \frac{1-e^{-j\frac{2\pi(r+k)N}{N}}}{1-e^{-j\frac{2\pi(r+k)}{N}}}. \tag{10}$$

In words, the analyzed signal x_n is not a member of the set of frequencies ω_k which are used to sample the frequency domain by the DFT. Many of the DFT basis functions with frequencies neighbouring ω_r are activated as they are not orthogonal to x_n. The third signal consists of R cosines which take any frequency up to the Nyquist frequency and noise, z_n which is iid, zero mean with variance σ^2,

$$x_n = \sum_{r=1}^{R}\cos(\omega_r n) + z_n. \tag{11}$$

This final test signal is consistent with the noise model used in [12] and the assumptions that underpin CELP in [6].

III. Linear Predictive Coding

If we compute the LPC of the signals in Equations 4, 9 or 11, we should be able to correctly determine its component frequencies. Equation 10 is an example of a signal where a TF method that uses the DFT might fail to estimate these component frequencies. Fig. 1 overlays the magnitude frequency response of the second order LPC filters estimated for two signals. The first signal is a cosine with a frequency of 16Hz, which corresponds to one of the DFT basis functions (Equation 4). The second cosine's frequency, 6.4Hz, is not one of the DFT basis functions (Equation 9). LPC gives the

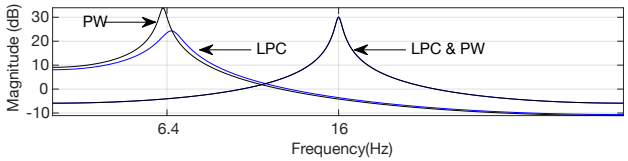

Fig. 1. LPC correctly estimates the 16Hz cosine as it has an integer number of cycles in the observation period. The error exhibited by LPC for the 6.4Hz cosine is 3.5%. The error in the PWLPC estimate is 3.2%.

correct frequency estimate for the 16Hz signal but not the 6.4Hz signal. The estimate given for a pure cosine signal is 6.625Hz. The error is 3.5% of 6.4Hz. It is tempting to view this result as an artefact of the discontinuities at the edges of the signal, and to appeal to a window function to improve the estimate; however, we have not yet considered the effects of noise (Equation 11) which are unlikely to be adequately resolved by a window.

The LPC all-pole model of radiation, vocal tract and glotal excitation is represented as

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{A(z)}. \tag{12}$$

The gain parameter is denoted as G, the transfer function is $H(z)$, the filter coefficients are a_k, the order of the filter is p. The DFT considers N samples of $X(e^{j\omega})$ whereas in the unilateral z-transform, $z = Ae^{j\omega}$, and thus ω may assume any value in the range $0 \leq \omega \leq 2\pi$. As there are no restrictions on the locations of the poles determined by LPC, for the correct value of p, LPC should identify the frequencies exactly. The poor performance of LPC for signals such as Equations 9 and 11 is due to the form of the signal used by the estimator. Can we improve the presented signal using some form of pre-emphasis filter? LPC produces a linear predictive estimate, $\hat{x}_n$, for x_n using a p-th order prediction filter

$$\hat{x}_n = -\sum_{k=1}^{p} a_k x_{n-k}. \tag{13}$$

The total error is $E = \sum_n (x_n - \hat{x}_n)^2$. Computing the derivative of E, setting it to zero and solving for the coefficients, $\frac{dE}{da_k} = 0$, where $\theta_{i,k} = x_{n-i}x_{n-k}$, yields

$$-\sum_n \theta_{0,k} = \sum_{i=1}^{p} a_i \sum_n \theta_{i,k}. \tag{14}$$

The function $\theta_{i,k}$ consists of up to two components for the signals considered here, $\theta_{i,k} = T_{i,k} + C_{i,k}$. For the signal in Equation 4 $\theta_{i,k} = T_{i,k} = \frac{1}{4}\sum_n e^{i\omega rA} + e^{-i\omega rA} + e^{i\omega rB} + e^{-i\omega rB}$, where $A = 2n + (i + k)$ and $B = k - i$. For the signal in Equation 11, when $R = 1$, $T_{i,k}$ is the same as above, but a new term $C_{i,k} = z_{n-i}z_{n-k} + z_{n-k}\cos(2\pi\omega_r(n-i)) + \cos(2\pi\omega_r(n-k))z_{n-i}$ perturbs the matrix inverse used to solve Equation 14 from the desired solution when $C_{i,k} \neq 0, \forall, i, k$. Our solution to this problem is to emphasize the desired component in the matrix inverse. For clarity, both components are expressed below. We solve for a_i in

$$-\left(\sum_n T_{0,k} + C_{o,k}\right) = \sum_{i=1}^{p} a_i \sum_n T_{i,k} + C_{i,k}. \tag{15}$$

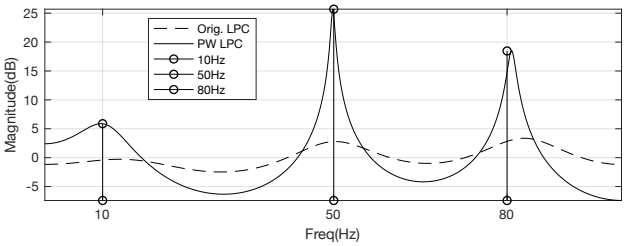

Fig. 2. PWLPC vs Traditional LPC for a mixture of three sinusoids in AWGN. Stems at 10, 50 and 80Hz indicate the correct frequencies. PWLPC gives excellent frequency estimates.

TABLE I
COMPARISON OF PW AND TRADITIONAL LPC ON NOISE CORRUPTED SINUSOIDS. THE SUM OF THE ABSOLUTE ERROR FOR EACH SINUSOID IS 1.05HZ FOR PWLPC AND 7.07HZ FOR LPC

True Frequency	10	50	80
PW Estimates	10.1562	49.2187	79.8828
LPC Estimates	13.0859	47.8516	81.8359

IV. POWER-WEIGHTING

Consider the effect of power-weighting the signal in the discrete frequency domain as a way of emphasizing the desired term $T_{i,k}$ via the operation $\text{DFT}\{x_n\}\|\text{DFT}\{x_n\}\|^2$. An inverse DFT produces a new time domain signal,

$$\hat{x}_n = \text{P}\{x_n\} = \text{iDFT}\{\text{DFT}\{x_n\}\|\text{DFT}\{x_n\}\|^2\}. \tag{16}$$

Applying this power weighting to Equation 4, e.g. $\hat{x}_n = \text{P}\{.5\left(e^{j\omega_r n} + e^{-j\omega_r n}\right)\}$ has the appealing property of scaling the signal at one frequency, the correct frequency, producing

$$x_n = \left(\frac{N}{2}\right)^2 .5\left(e^{j\omega_r n} + e^{-j\omega_r n}\right), \tag{17}$$

which is submitted to LPC analysis. Recall the case when the signal frequency corresponds to one of the DFT basis function frequencies (cf. the 16Hz cosine in Fig. 1). LPC yielded the correct frequency estimate for the 16Hz frequency component in Fig. 1. PWLPC also gives the correct estimate of 16Hz. PWLPC does not adversely affect LPC when LPC is accurate. When the cosine's frequency (6.4Hz in Fig. 1) does not correspond to one of the DFT basis function frequencies, power weighting focuses the magnitude frequency response on the frequencies neighbouring the true frequency, and improves the estimate. We apply LPC to the power-weighted signal,

$$\text{P}\{x_n\} = \text{iDFT}\left\{\left(\frac{1 - e^{j2\pi(r-k)}}{1 - e^{j\frac{2\pi(r-k)}{N}}} + \frac{1 - e^{-j2\pi(r+k)}}{1 - e^{-j\frac{2\pi(r+k)}{N}}}\right) \left|\frac{1 - e^{j2\pi(r-k)}}{1 - e^{j\frac{2\pi(r-k)}{N}}} + \frac{1 - e^{-j2\pi(r+k)}}{1 - e^{-j\frac{2\pi(r+k)}{N}}}\right|^2\right\} \tag{18}$$

and observe a frequency estimate of 6.18Hz, which has a smaller error than for LPC. The error is 3.2% of 6.4Hz. Power weighting scales the activated frequency components of the signal by the instantaneous power of each of the components.

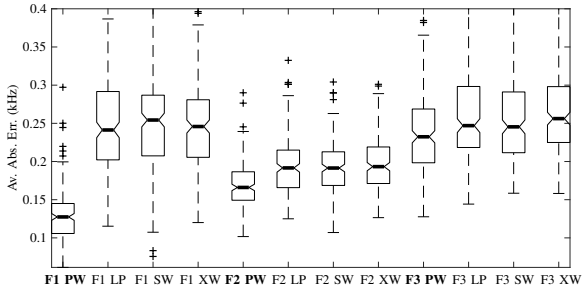

Fig. 3. PWLPC, LPC, SWLP and XWLP estimation errors for F1, F2 and F3 for the VTR Formants database. PWLPC gives better formant estimates.

V. Numerical Evaluation

We demonstrate the benefits of PWLPC by showing that: (1) it outperforms benchmark LPC methods on synthetic signals; (2) the median of the average absolute error achieved by PWLPC over all 192 utterances in the VTR Formants database is $\approx 47\%$ smaller than the median of the average absolute error achieved by LPC; (3) the improvement achieved by PWLPC is not significantly affected by window discontinuities; and finally, (4) PWLPC's computational cost is reasonable.

We start by establishing that PWLPC performs better than LPC in the presence of noise. We generate the noise-corrupted signal in Equation 11. The cosine frequencies are at 10, 50, and 80 Hz. Sixth-order LPC and PWLPC filters are estimated. Additive White Gaussian Noise of -4.5dB is added to the sum of three sinusoids. The aim is to detect the correct cosine frequencies and to see if frequencies given by the sinusoids are emphasized sufficiently by power weighting to improve LPC frequency estimation. Table I summarizes the locations of the frequencies estimated by both PWLPC and LPC. Each of the PW estimates are less than 1Hz from the correct value. The sum of the errors experienced by PWLPC is approximately 1Hz for three frequencies. To better understand why PWLPC works so well, we plot the magnitude response of the PWLPC and LPC filters in Fig. 2. Pre-emphasis via power weighting emphasizes the magnitude in the locations of the cosine components and de-emphasizes the other components.

PWLPC is compared with the baseline techniques, LPC, SWLP and XWLP in Fig. 3 using the VTR Formants database [17] for different values of p. The VTR Formants database provides estimates of the first three Vocal Tract Resonance (VTR) frequency trajectories obtained by human experts for the VTR frequencies F1, F2 and F3. We estimate F1, F2 and F3 for 192 test utterances (8 utterances from 24 speakers) using PWLPC, LPC, SWLP and XWLP and compute the error, which is the mean absolute difference between these estimates and the values determined by the VTR database experts. The experimental parameters outlined in [17] are used. They consist of a pre-emphasis filter with taps 1 and .97 and a sampling rate of 16kHz. The analysis window is a Hamming window of length 1024 samples. It is advanced by 10ms. Unlike the cosine signals analyzed above, we do not have the ground truth solution, but instead compare our estimates with the values in the VTR database. Our hypothesis is that PWLPC outperforms traditional LPC, SWLP and XWLP as the ridge-like features selected by the experts are likely to be detected by the PWLPC that emphasizes these ridges. The length and delay of the short-time energy window in SWLP is set to the values recommended in [12] (p and 1 respectively). The comparable fixed window used by XWLP is also set to the value, p, which is recommended by the authors in [13].

TABLE II
Percentage improvement in the median error achieved by PWLPC over LPC for the VTR Formants database.

Median error reduction	F1	F2	F3
6 coefficients	47%	13%	6%
8 coefficients	41%	19%	10%
10 coefficients	34%	23%	21%

TABLE III
Improvement in the median error achieved by PWLPC over LPC for $p = 8$ using a Hann window and two Kaiser windows.

window	F1	F2	F3
Hann	40%	18%	8%
Kaiser ($\beta = 9$)	41%	18%	9%
Kaiser ($\beta = \frac{1}{2}$)	43%	17%	11%

Guidance on detecting formant locations and their bandwidths is given in [18]. We adopt the criteria that formant frequencies should be greater than 90Hz and have bandwidths which are less than 400Hz. Regarding the model order, the rule of thumb is that the order of the filter should be twice the expected number of formants plus 2. We fit models of the order $p = \{6, 8, 10\}$. For $p = 8$, the bandwidth criteria resulted in 2-4 formant frequency estimates for PWLPC and the LPC variants for each time-window of speech. We assigned the estimated coefficients to the closest VTR database formant frequency, and calculated the error between the expert estimate and the estimate obtained by the algorithms for each formant for each utterance in the VTR database.

Performance Gain: Boxplots of the error for F1, F2 and F3 for the PWLPC, LPC, SWLP and XWLP estimators are illustrated in Fig. 3. PWLPC always gives a better estimate of F1 than LPC and its variants. PWLPC generally gives a better estimate than LPC and its variants for F2 and F3. The median error of PWLPC estimation of F2 and F3 is less than the median error for the same formants using traditional LPC, SWLP and XWLP. The median of the average absolute error achieved by PWLPC over all 192 utterances is approximately 47% smaller than the median of the average absolute error achieved by LPC estimation. Table II tabulates the percentage reduction in the median of the error achieved by using PWLPC over LPC for formant estimation when $p = \{6, 8, 10\}$. It establishes that PWLPC improves the LPC estimates given a range of filter orders. The improvement for F1 is explained by the fact that the power of the F1 component dominates F2 and F3, which allows PWLPC to better emphasize this component. Improvements for F2 and F3 are also generally achieved given that these components have smaller power. Table II supports the claim that PWLPC significantly and reliably outperforms LPC at the task of formant estimation using the VTR dataset.

Discussion: The goal of LPC is to produce a linear predictive

estimate of speech. The problem of detecting formant frequencies in speech is related, but different. Input speech may include noise, and other effects such as those introduced by taking a short-time portion of the speech signal to compute local-in-time formant frequencies estimates. The result of the traditional LPC approach is that the filter coefficients learned also consider these artefacts. Therefore, the formant frequencies estimated explain *all* of the data and not just the formant frequencies. As a consequence, the formant frequencies learned by LPC in the VTR formants database are frequently in the valleys of the magnitude TF representation of the speech signal. Power weighting of speech causes estimated PWLPC formants to be re-aligned with the tops of the ridges of the magnitude TF representation. Experimental evidence supports the claim that in general PWLPC pre-emphasizes speech so that the formants estimated are aligned with the frequencies of the vocal tract resonances. The authors of the VTR database in [17] note that the panel of experts determined that formants lay where spectral valleys were instead of spectral peaks in some cases. The justification for this was that the experts deemed values should be consistent with their prior knowledge. PWLPC exploits spectral peaks to improve formant estimation. When formants are not aligned with spectral peaks, power weighting may not be appropriate. This is also true for the LPC estimator. Visual inspection indicates that when the formants correspond to spectral peaks, the PWLPC produces an estimate which is similar to the judgement of the VTR database experts.

Window Comparison: Regarding the window discontinuities introduced into the theoretical signals as a result of short-time processing, PWLPC takes some steps towards reducing the errors in formant frequency estimates using a Hamming window. We consider the effect of discontinuities by using a Kaiser window with different values of β, the relative side-lobe attenuation parameter. A Hann window, which is commonly used for speech, is also considered. Table III demonstrates that PWLPC gives a larger improvement in the median error achieved by PWLPC over LPC for a flatter time-domain window ($\beta = \frac{1}{2}$) than when the Kaiser window has greater attenuation at either end ($\beta = 9$). PWLPC outperforms LPC for the higher frequeny formant, e.g. F3 using the flatter kaiser window ($\beta = \frac{1}{2}$). Improvement achieved by PWLPC in Table III is not significantly affected by the analysis window.

Computational Complexity: PWLPC analysis can be efficiently implemented due to the central role of the FFT. A first N-point FFT, which costs $O(N \log_2(N))$ FLOPS, is multiplied by its magnitude squared, which costs $4N$ FLOPS. An inverse FFT, costing $O(N \log_2(N))$ FLOPS, produces the power weighted signal which is passed to LPC. The Levinson-Durbin algorithm solves the Yule-Walker equations in $O(p^2)$ FLOPS. The sampling rate, window size N and filter order p, determine the burden incurred by choosing PWLPC analysis. This amounts to an additional cost of two $O(N \log_2(N))$ FLOPS operations and $4N$ FLOPS over LPC's $O(p^2)$ cost. The performance gain demonstrated above outweights this additional computational cost, which amounts to two additional FFT-like operations in practice.

VI. Conclusion

A central concern in speech processing is the estimation of formant frequencies. We demonstrated that by accounting for the entire signal, and not the component due to the formants, the traditional approach for estimating formants, LPC, was inaccurate. We introduced a time-domain pre-emphasis method, which weighted the input signal so that the components with the most energy had the largest input into determining where LPC placed its formants. PWLPC outperformed traditional LPC, which is commonly used as a feature selection method in machine learning formant tracking systems.

Acknowledgment

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/SIRG/3459

References

[1] L.R. Rabiner and R.W Schafer, "Digital processing of speech signals," *Prentice Hall, Englewood Cliffs, New Jersey*, 1973.
[2] R. de Fréin and S.T. Rickard, "Power-weighted divergences for relative attenuation and delay estimation," *IEEE Sig. Proc. Let.*, vol. 23, no. 11, pp. 1612–1616, 2016.
[3] T.E Tremain, "The government standard linear predictive coding algorithm: LPC10," *Speech Technol.*, vol. 1, pp. 40–49, 1982.
[4] T. Tsai and C. Liu, "Low-Power System Design for MPEG-2/4 AAC Audio Decoder Using Pure ASIC Approach," *IEEE TCAS I*, vol. 56, no. 1, pp. 144–155, 2009.
[5] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *ICASSP*, 1982, vol. 7, pp. 614–617.
[6] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *ICASSP*, 1985, vol. 10, pp. 937–940.
[7] S.V. Andersen, W.B. Kleijn, R. Hagen, J. Linden, M.N. Murthi, and J. Skoglund, "iLBC - a linear predictive coder with robustness to packet losses," in *IEEE Wksh Proc. Sp. Coding*, 2002, pp. 23–25.
[8] T. Gueham and F. Merazka, "An enhanced interleaving frame loss concealment method for voice over IP network services," in *EUSIPCO*, 2018, pp. 1302–1306.
[9] G. Fuchs, C.R. Helmrich, G. Markovi, M. Neusinger, E. Ravelli, and T. Moriya, "Low delay LPC and MDCT-based audio coding in the EVS codec," in *ICASSP*, 2015, pp. 5723–5727.
[10] Z. G. Zhang, S. C. Chan, and K. M. Tsui, "A recursive frequency estimator using linear prediction and a Kalman-filter-based iterative algorithm," *IEEE TCAS II*, vol. 55, no. 6, pp. 576–580, 2008.
[11] Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," *J. Acous. Soc. America*, vol. 145, no. 2, pp. 642–653, 2019.
[12] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Comm.*, vol. 51, no. 5, pp. 401–411, 2009.
[13] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo, "Noise robust feature extraction based on extended weighted linear prediction in LVCSR," in *INTERSPEECH*, 2011.
[14] R. de Fréin and S.T. Rickard, "The synchronized short-time-Fourier-transform: Properties and definitions for multichannel source separation," *IEEE Trans. Sig. Proc.*, vol. 59, no. 1, pp. 91–103, 2011.
[15] W. Liu, D. P. Mandic, and A. Cichocki, "A class of novel blind source extraction algorithms based on a linear predictor," in *IEEE Int. Symp. Circ. Sys.*, 2005, pp. 3599–3602 Vol. 4.
[16] W. Liu, D. P. Mandic, and A. Cichocki, "Blind second-order source extraction of instantaneous noisy mixtures," *IEEE TCAS II*, vol. 53, no. 9, pp. 931–935, 2006.
[17] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *ICASSP*, 2006, vol. 1, pp. I–I.
[18] R.C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. Sp. Aud. Proc.*, vol. 1, no. 2, pp. 129–134, 1993.