Doppler-variant modeling of the vocal tract

Heim, Axel; Sorger, Uli; Hug, Florian

Published in:
ITG Conference on Voice Communication [8. ITG-Fachtagung]

DOI:
10.1109/ICASSP.2008.4518580

2008

DOPPLER-VARIANT MODELING OF THE VOCAL TRACT

Axel Heim[†]*, Uli Sorger*[‡] *and Florian Hug*[†]

[†] Institute of Telecommunications and Applied Information Theory, Ulm University, Germany
[‡] Computer Science and Communications, University of Luxembourg, Luxembourg

ABSTRACT

A common technique to deploy linear prediction to non-stationary signals is time segmentation and local analysis. Variations of a process within such a segment cause inaccuracies. In this paper, we model the temporal changes of linear prediction coefficients (LPCs) as a Fourier series. We obtain a compact description of the vocal tract model limited by the predictor order and the maximum Doppler frequency. Filter stability is guaranteed by all-pass filtering, deploying the human ear's insensitivity to absolute phase. The periodicity constraint induced by the Fourier series is counteracted by oversampling in the Doppler domain. With this approach, the number of coefficients required for the vocal tract modeling is significantly reduced compared to a LPC system with block-wise adaptation while exceeding its prediction gain.

As a by-product it is found that the Doppler frequency of the vocal tract is in the order of $10\,\mathrm{Hz}$. A generalization of the algorithm to an auto-regressive moving average model with time-correlated filter coefficients is straight forward.

***Index Terms*—** Speech coding, Linear predictive coding, Fourier series

1. INTRODUCTION

There has been a vast amount of research on how to remove redundancy from speech signals allowing for efficient transmission over digital communication channels. One approach in speech coding is to model the speech generation process by a (modulated) excitation source and a subsequent filter, according to the airflow from the lungs passing the vocal chords and the vocal tract, respectively. By this, a speech signal with seemingly high entropy can be broken down into two low entropy processes. The vocal tract filter is commonly modeled as an inverse Linear Predictor (LP), wherefore this technique is referred to as linear predictive coding.

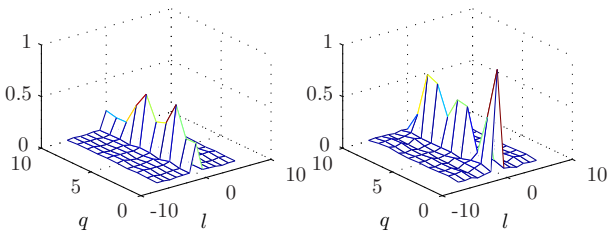
Speech is a non-stationary process, meaning its properties are changing over time, e.g., as the shape of the vocal tract is changing. In classical linear predictive coding, the difficulty of non-stationarity is addressed either by fixed-length windowing of a speech signal, as it is assumed to be stationary for relatively short periods of 20 to 400 ms, or the predictor coefficients are sequentially adapted (but still locally optimized) at each discrete time instance – to the expense of increase in computational complexity [7]. In [6] the importance of window size and placement when applying block-oriented adaptation is pointed out, and an adaptive algorithm to minimize the final cost (in bits) of the linear predictor description is introduced. This global optimization implicates increased processing delay, making it inappropriate for time-critical applications.

In this paper we address the problem of global optimization by introducing a time-variant description of the linear prediction coefficients (LPCs), allowing significantly larger window sizes and thus eliminating the placement problem. The time-variance of each LPC is modeled as a Fourier series, yielding the compact representation of a so-called spreading matrix[1]: Increasing the prediction filter order beyond an empirical maximum value the additional prediction gain is small, hence higher order predictor coefficients $a_q, q > Q$, are disregarded. The Doppler frequencies, i.e., the number of Fourier coefficients, are limited due to the finite-speed shape changes of the vocal tract, e.g. by movement of tongue or jaws.

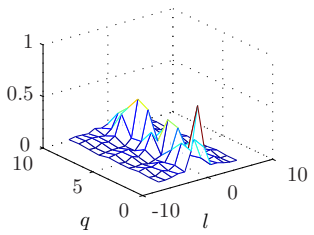
Example: Transition between two vowels
Figure 1 shows the magnitude of the spreading coefficients for the German sound of the vowels 'i' (a) and 'o' (b) and the transition 'io' between them (c), where q indexes the predictor coefficient and l the discrete Doppler frequency. For constant sounds, in (a) and (b) the energy at non-zero frequencies ($l \neq 0$) is nearly zero as the vocal tract does not change. The constant components ($l = 0$) of the spreading matrix reflect the values of 'normal' LPCs. For (c) the vocal tract is changing, causing non-zero values for $l \neq 0$.

The remainder of this paper is structured as follows. After a brief review of linear predictive coding in Section 2 we will formally introduce the representation of LPCs as spreading (matrix) coefficients and their calculation in Section 3, discussing problems coming up with the Fourier series representation and providing solutions. Suitable model parameters are experimentally found in Section 4 before we conclude in Section 5.

[1] This terminology originates from the field of wireless communications where spreading matrices are used to represent time-variant channels [2].



(a) vowel 'i' (b) vowel 'o'



(c) transition between 'i' and 'o'

Fig. 1. Magnitude $|\mathcal{S}_{q,l}|$ of spreading coefficients for transition between vowels 'i' and 'o', with Doppler frequency index l and predictor coefficient index q

2. LINEAR PREDICTIVE CODING

For linear predictive coding, the speech generation process is approximated by a source-filter model, with the excitation source being the airflow from the lungs, potentially modulated by the vocal chords, and the filter being the vocal tract, cf. Figure 2. For the latter, an all-pole filter

$$H(z) = \frac{X(z)}{D(z)} = \frac{1}{A(z)}\,, \tag{1}$$

where

$$A(z) = 1 - \sum_{q=1}^{Q} a_q \cdot z^{-q} \tag{2}$$

is the inverse filter of $H(z)$, and a sufficient number of poles is a good approximation [3]. By minimizing the expectation of the quadratic prediction error we can correctly estimate the coefficients a_q of $A(z)$ [7], where Q is the prediction order. Equation (1) is equivalently given in the time domain by

$$d(k) = x(k) - \sum_{q=1}^{Q} a_q \cdot x(k-q)\,. \tag{3}$$

Hence for speech segments of finite length K the optimization task becomes the minimization of the quadratic prediction error

$$\sum_{k=0}^{K-1} |d(k)|^2 = \sum_{k=0}^{K-1} \left| x(k) - \sum_{q=1}^{Q} a_q \cdot x(k-q) \right|^2 \rightarrow \min\,. \tag{4}$$

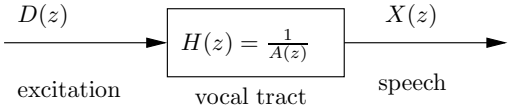


Fig. 2. Source-filter model

This results in a set of equations called the *Yule-Walker* equations. These can be efficiently solved with the Levinson-Durbin algorithm [5, 1], which moreover guarantees the resulting filter to be stable.

As a quality measure for linear predictors we consider the *prediction gain* . It is defined by the ratio

$$G_p = \frac{\sigma_x^2}{\sigma_d^2} \tag{5}$$

of speech signal power and excitation power

$$\sigma_x^2 = \frac{1}{K} \cdot \sum_{k=0}^{K-1} |x(k)|^2 \quad \text{and} \quad \sigma_d^2 = \frac{1}{K} \cdot \sum_{k=0}^{K-1} |d(k)|^2\,,$$

respectively. It is a measure for the bit rate reduction achievable by predictive coding and increases monotonically with the prediction order.

3. DOPPLER-VARIANT REPRESENTATION OF LINEAR PREDICTION COEFFICIENTS

In this section we introduce a new, time-variant representation of LPCs. It allows to significantly increase the segment length $N \gg K$, thus reducing the window placement problem [6] without surrender of temporal adaptivity of the LPCs.

3.1. LPCs in the Doppler Domain

We model the temporal changes of the LPCs a_q as a discrete time Fourier series

$$a_q(k) = \sum_{l=-L}^{L} \mathcal{S}_{q,l} \cdot e^{j\frac{2\pi lk}{N}}, \quad q = 1,\ldots,Q;\; k = 0,\ldots,N-1. \tag{6}$$

The coefficients $\mathcal{S}_{q,l}$ will be referred to as *Spreading Coefficients* (SCs) with the filter tap index q and the discrete Doppler frequency index l. Note that instead of summing over a complete period $l \in \{0,\ldots,N-1\}$ according to the inverse DFT, we restrict l in the range $-L \le l \le L$, thus truncating higher (positive and negative) Doppler frequencies. The physical frequency corresponding to the l-th discrete Doppler coefficient is thus given by

$$f_D^{(l)} = l \cdot \frac{f_s}{N}$$

where f_s is the sampling frequency. Hence, the maximum Doppler frequency contained in the model is $f_D^{(L)} = L \cdot \frac{f_s}{N}$.

Replacing a_q in Equation (3) by its time-variant representation in (6), the optimization task is now to find the values of $\mathcal{S}_{q,l}$ which minimize the mean square prediction error

$$\sum_{k=0}^{N-1} \left| x(k) - \sum_{q=1}^{Q} x(k-q) \sum_{l=-L}^{L} \mathcal{S}_{q,l} \cdot e^{j\frac{2\pi lk}{N}} \right|^2 \rightarrow \min . \quad (7)$$

This can be done, e.g., by using the Least-Mean-Square (LMS) or the Recursive Least-Squares algorithm. Of course, as speech samples and therefore LPCs are real-valued, $\mathcal{S}_{q,l} = \mathcal{S}^*_{q,-l}$, where * denotes the complex conjugate. Hence, the number of real valued coefficients is cut in half when using the real-valued notation of the Fourier series. However, the complex valued notation will be preferred in the sequel as it is more compact. To distinguish from classical linear prediction (LP) analysis, in the following we will denote the proposed approach as Spreading Coefficient (SC) analysis.

3.2. Stability Considerations

As opposed to frame-wise calculation of LPCs with the Levinson-Durbin algorithm which guarantees stability of the resulting filter [4], this does not hold for the LMS solution of the proposed approach. These instabilities occur relatively seldom and primarily during speech pauses or plosives. When at a time instance k a filter realization $A(z)$ with coefficients from Equation (6) and with roots $|z_0| > 1$ is detected, all-pass filtering is applied. The inferred phase shift is disregarded as human speech perception is generally insensitive to absolute phase. The resulting minimum-phase filter is

$$A_{\min}(z) = z^{-Q} \cdot \prod_{|z_0|\leq 1} (z - z_0) \cdot \prod_{|z_0|>1} (z - 1/z_0^*) ,$$

where z_0 are the roots of $A(z)$.

3.3. Oversampling in the Frequency Domain

As discrete Doppler frequencies imply a periodic time signal respectively a periodic sequence of LPCs, errors will raise at the segment boundaries. As in [2], we address this problem by virtually lengthening the period N to $N \cdot \nu$ with the oversampling factor $\nu \geq 1$, while retaining the optimization interval to the first N samples. By this Equation (7) becomes

$$\sum_{k=0}^{N-1} \left| x(k) - \sum_{q=1}^{Q} x(k-q) \sum_{l=-L}^{L} \mathcal{S}_{q,l} \cdot e^{j\frac{2\pi lk}{N\nu}} \right|^2 \rightarrow \min$$

(note the 'ν' in the exponent). Hereby, the values of $a_q(N-1)$ and $a_q(0)$, $q = 0, \ldots, Q$ are no longer 'glued' together but can follow their 'true' value while their difference is balanced over the virtual interval from N to $N\nu - 1$. Note that $N\nu$ does not even have to be an integer. By this virtual lengthening also the spectral distance between the discrete frequencies in the Fourier domain as well as the maximum Doppler frequency $f_D^{(L)}$ are reduced to

$$\Delta(\nu) = \frac{f_s}{N\nu} \quad \text{and} \quad f_{D,\max} = f_D^{(L)} = L \cdot \frac{f_s}{N\nu} , \quad (8)$$

respectively, wherefore we refer to this procedure as *oversampling in the frequency domain*. Note the difference to the more common technique of *zero-padding* where the minimization would take place on $k = 0, \ldots, N\nu - 1$ with $x(k) = 0$ for $N \leq k < N\nu$.

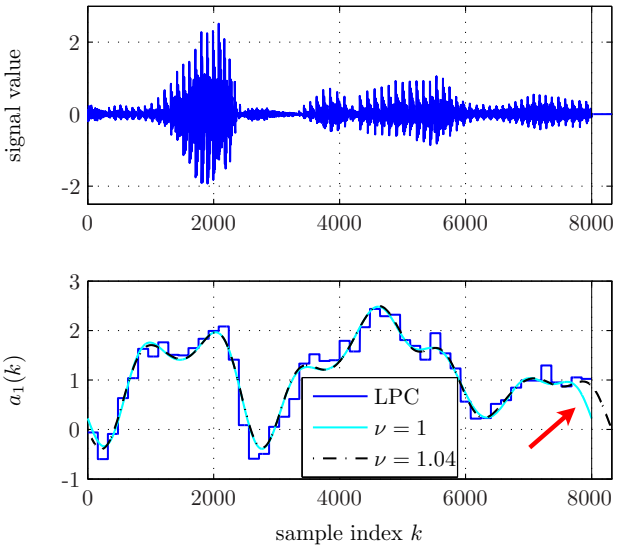


Fig. 3. Speech segment of length $N = 8000$ (top) and values of 1^{st} predictor coefficient $a_1(k)$ obtained by LP analysis and SC analysis without ($\nu = 1$) and with oversampling ($\nu = 1.04$) (bottom)

Figure 3 shows a speech segment of 1 second length sampled at $f_s = 8\,\text{kHz}$ (top) and the 1^{st} predictor coefficient $a_1(k)$ obtained by classical LP analysis with frames of 20ms as well as by SC analysis for oversampling factors $\nu = 1.0$ and $\nu = 1.04$ (bottom). We observe that the LPCs obtained by SC analysis are smoothing the staircase-like LPCs from LP analysis. For $\nu = 1$, i.e., without oversampling, we note an increased deviation from the classical LPCs at the segment boundary due to the periodicity constraint. For $\nu = 1.04$, the difference between $a_1(N-1)$ and $a_1(0)$ is balanced over the virtually extended interval.

4. EXPERIMENTAL RESULTS

We will now experimentally investigate the performance of the SC representation of LPCs in dependence of the parameters provided by the Fourier transform. For this evaluation we use 100 seconds of speech from a standard test corpus sampled at $f_s = 8\,\text{kHz}$ and a predictor order $Q = 10$. For classical linear prediction we choose 20 ms frames and apply a Hamming window before calculating the LPCs. The

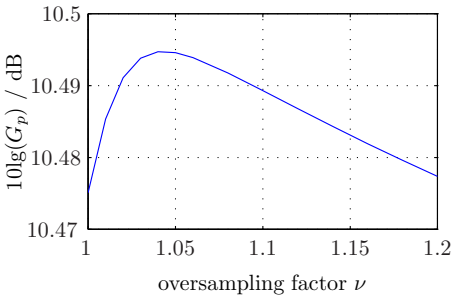


Fig. 4. Logarithmic prediction gain over ν for $L = 10$

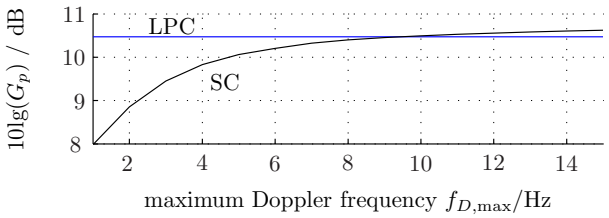


Fig. 5. Prediction gain G_p over maximum Doppler frequency

predictor coefficients calculated from SCs are checked for the minimal phase property and - if not fulfilled - all-pass filtering according to Section 3.2 is applied.

First we evaluate the oversampling factor ν. For small values, it is assumed to mostly affect the boundary regions of the analyzed speech segments due to the balancing. However, with increasing ν the frequency band of the model is shrinking, cf. Equation (8). Hence we can expect G_p to have a maximum, which is shown in Figure 4 for $N = 8000$. The maximum gain of $0.02\,\mathrm{dB}$ for $\nu = 1.04$ seems small, but this is the average over the complete speech segment, while the actual gain primarily origins from the boundary regions.

Next we consider the prediction gain G_p as a function of $f_{D,\max}$ and the segment length N. For speech segments of length $N = 8000$ (one second), Figure 5 shows the prediction gain over $f_{D,\max} \approx L$. The horizontal line gives the reference prediction gain with LP analysis. In fact, three SP analysis curves are plotted in Figure 5 for $N = 8000, 16000$ and 32000, respectively, which, however, completely overlap. We conclude that for a given target G_p the segment length N can be freely chosen as long as L is set such that a certain $f_{D,\max}$ is reached.

The monotonically increasing SP analysis curve intersects the LP reference line at $f_{D,\max} = 10\,\mathrm{Hz}$ and is relatively flat thereafter. Hence we conclude that our model represents the temporal changing of the vocal tract well for this value. The other way round, this means that the true maximum Doppler frequency of the vocal tract is in the order of $f_{D,\max} \approx 10\,\mathrm{Hz}$.

For our system setup, the number of real valued coefficients necessary to define the filter is 500 per second for LP analysis compared to only

$$(2L+1) \cdot Q \cdot \frac{f_s}{N} = 210/\mathrm{s}$$

for SC analysis with $f_{D,\max} = 10\,\mathrm{Hz}$ and $N = 8000$.

5. CONCLUSIONS

We introduce a time-variant representation for temporally correlated linear predictor coefficients which allows the modeling of non-stationary processes such as speech. We apply oversampling to counteract the boundary constraint induced by the Fourier series and use all-pass filtering to stabilize estimated vocal tract filter coefficients. The prediction gain of a comparable frame-wise LP system is exceeded when the maximum Doppler frequency contained in the model is $10\,\mathrm{Hz}$ or above, while dramatically reducing the number of coefficients required for the filter model. We hope that this reduction transfers to quantized coefficients allowing better data compression. Due to the inferred delay, however, the approach is not adequate for real-time application. We present our idea as a starting point for this line of research. Further investigations will show whether the performance transfers to quantized spreading coefficients and will concentrate on efficient optimization algorithms.

6. REFERENCES

[1] J. Durbin. The Fitting of Time-series Models. *Revue de l'Institut International de Statistique*, Vol. 28, 1960.

[2] Snjezana Gligorevic. Joint Channel Estimation and Equalisation of fast Time-varying Frequency-selective Channels. *European Transactions on Telecommunications*, 2006.

[3] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing*. Prentice Hall, 2001.

[4] K. D. Kammeyer and K. Kroschel. *Digitale Signalverarbeitung*. Teubner Studienbücher, 1998.

[5] N. Levinson. The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction. *Journal of Mathematical Physics*, Vol. 25:pp. 261–278, 1947.

[6] Paolo Prandoni and Martin Vetterli. R/D Optimal Linear Prediction. *Speech and Audio Processing, IEEE Transactions on*, vol.8(no.6):pp.646–655, Nov 2000.

[7] Peter Vary and Rainer Martin. *Digital Speech Transmission*. John Wiley & Sons Ltd, 2006.

Acknowledgments

The authors would like to thank Peter Vary and Hauke Krüger for sharing their experience in speech coding and fruitful discussions.