Influence of Differences between Inverse Filtering Techniques on the Residual Signal of Speech

Hartmut R. Pfitzinger

Inst. of Phonetics and Speech Communication, Ludwig-Maximilians-Univ., 80799 Munich, Germany, Email: hpt@phonetik.uni-muenchen.de

Introduction

Human speech production is characterized by two major processes: *i)* the source signal generation, which is either the quasi-periodic vibration of the vocal folds in voiced sounds, a turbulent airstream in voiceless sounds, or a combination of both in voiced fricatives, and *ii)* the slowly varying shape of the vocal tract causing a time-varying modulation of the spectral envelope of the speech signal. This dichotomy is described by the *source-filter model of speech production* [3]: the speech signal is the convolution of the excitation signal with the impulse response of the vocal tract.

Since the late sixties efficient algorithms based on the principle of LPC (*linear predictive coding*) have been applied to speech in order to decompose it into filter coefficients and the residual signal. This decomposition is also known as the *inverse filtering* of speech. Usually, every 10 ms an all-pole filter is estimated which minimizes the sum of the squares of the residual signal samples. This minimization criterion guarantees that the residual signal typically has a flat (or a so-called *white*) spectrum while the all-pole filter has ideally adapted to the formant structure of the vocal tract.

However, two inadequacies of LPC are widely tolerated: first, the least-squares minimization criterion equally distributes the residual error over all samples although it is well-known that the rapid closing of the vocal folds is the main excitation of the vocal tract and thus periodically produces a short-term large error. More sophisticated extensions to standard LPC overcome this problem using two different approaches: some detect these samples and exclude them from the minimization procedure (e.g. *robust linear prediction* [5], *weighted LPC* [7]), others minimize the squared error between the residual signal and the derivative of a glottal excitation model [4, 6, 9].

The second inadequacy is related to the varying overall slope of the short-term spectrum of speech which occurs not only between voiced and voiceless sounds but also between several instances of the same sound. The average slope of the spectrum envelope of the voice source is approx. -12 dB per octave. However, it is well-known that this slope varies between at least -9 dB to -15 dB per octave [2] depending on e.g. voice effort. Nevertheless, the widely applied fixed pre-emphasis of 6 dB per octave attributes all spectral voice source variations to the filter function of the vocal tract.

Method

This study aims at comparing four well-known LPC methods with two new LPC techniques which we recently developed for inverse filtering and reconstruction of the underlying glottal waveform. The problem of how to assess the adequacy of the estimated filter coefficients is usually solved by referring to the mean squared residual error [4] or to the deviation between synthetic and extracted formant frequencies [7]. But since we intend to use real speech signals for evaluation, the true formant frequencies are unknown and therefore not available as a reference. Besides, a smaller residual error is not necessarily better: the residual error should be small *between* and large *at* the excitation impulses. Therefore we developed a new evaluation measure: the *normalized standard deviation of absolute error* (*NSDAE*) which is the standard deviation of the absolute residual samples divided by their mean.

Four standard LPC techniques have been implemented: *autocorrelation LPC* [8], *covariance LPC* [8], Burg's LPC which is also known as the *maximum entropy method LPC* (MEM-LPC) [1], and *weighted LPC* [7]. Additionally, our two new LPC methods have been implemented which we would like to name *minimized linear error LPC* (MLE-LPC) and *smoothed glottal derivative LPC* (SGD-LPC). Both are two-pass algorithms: the MLE-LPC uses the filter coefficients of the covariance LPC as a first guess and then minimizes the linear absolute residual error instead of the quadratic residual error. The SGD-LPC is more complex: first, the residual signal is estimated via covariance LPC and integrated, then the resulting derivative glottal flow is smoothed with a 0.3-ms-Hanning window and differentiated to achieve the new residual signal. Finally, the filter coefficients are re-estimated to fit this signal.

Adaptive Pre-Emphasis

To overcome the problem that the slope of the spectrum envelope of the voice source varies as described in the Introduction we applied adaptive pre-emphasis prior to all LPC techniques. Fig. 1 shows the basic scheme in which the 18th order LPC analysis is one of the five different methods to be compared.

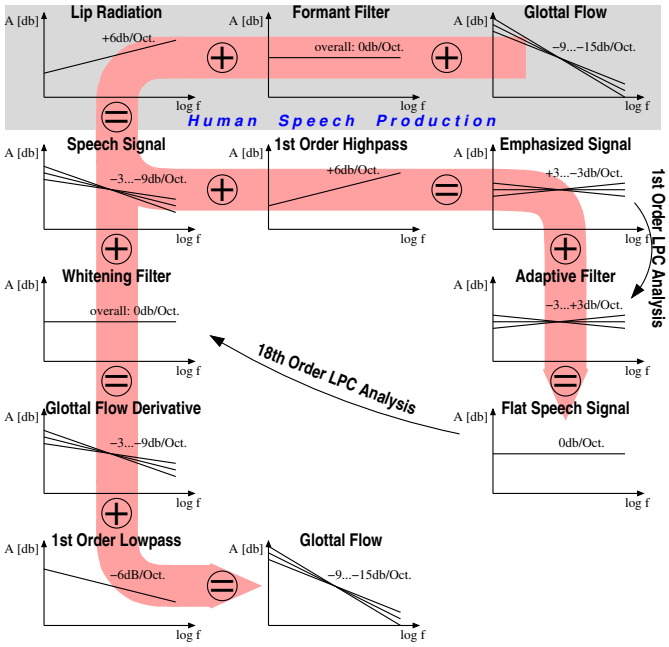

Figure 1: Estimation of the excitation signal via adaptive pre-emphasis, LPC-based inverse filtering, and integration.

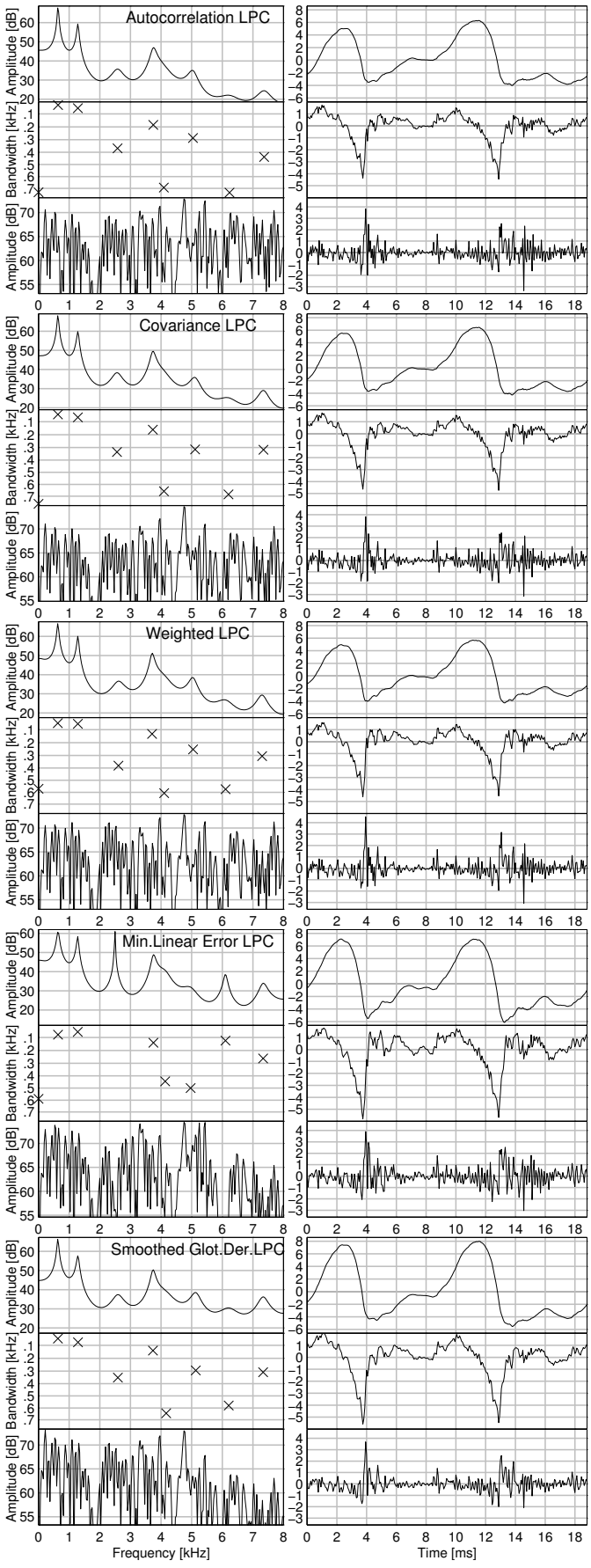

Figure 2: Each of the five different LP methods is displayed in one 6-section-panel: *left 3 sections of each panel:* transfer function of the filter coefficients, their poles, and spectrum of the residual signal; *right 3 sections:* glottal flow, its derivative, and the residual signal.

Results

As an example, Fig. 2 shows spectral and temporal properties of five LPC methods analysing 20 ms of an /a/. Burg LPC was omitted because its panel was almost identical to that of the covariance method. While autocorrelation, covariance, Burg, and weighted LPC yield an almost white residual signal, the MLE and SGD methods show less energy for higher frequencies (compare lower left sections of each 6-section panel in Fig. 2). The SGD-LPC reduces the amplitude of a negative outlier at 14.5 ms more than the other methods do. It also emphasizes the glottal excitations at 4 ms and 13 ms as does the weighted LPC. None of the methods manage to remove the ripple which directly follows the glottal closure instants at 4 ms and 13 ms (see the middle right section of each panel in Fig. 2). But a strong effect of the particular inverse filtering technique on the shape of this ripple is obvious.

LPC technique	*NSDAE*
Autocorrelation	1.085
Covariance	1.092
Maximum entropy method (Burg)	1.089
Weighted LPC	1.129
MLE-LPC (Minimized linear error LPC)	1.020
SGD-LPC (Smoothed glottal derivative LPC)	1.111

For objective evaluation we estimated the *NSDAE* (*normalized standard deviation of absolute error*) of the residual signals of all voiced sounds taken from a 45-minutes 16-speaker read-speech database. A bigger *NSDAE* indicates a more adequate residual signal. These results show that while MLE-LPC performs worst, weighted LPC yields an excitation signal which is slightly better than SGD-LPC. This is mainly due to the increased amplitudes of the excitation impulses at 4 and 13 ms.

Conclusions

Our new evaluation measure provides reasonable results: residual signals which better represent the underlying speech production process, yield higher *NSDAE* values. Although weighted LPC performs slightly better than SGD-LPC, the latter produces a glottal flow signal which is closer to widely used glottal flow models (e.g. the *Liljencrants-Fant-Model*).

References

[1] Burg, J. P. (1967). Maximum entropy spectral analysis. Report, Proc. of the 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City; Oklahoma.

[2] Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics*, 15(1): 3–108.

[3] Fant, G. (1970). *Acoustic theory of speech production.* Mouton & Co., The Hague; Niederlande, 2. Ed.

[4] Fujisaki, H.; Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP86)*, pp. 1605–1608, Tokyo.

[5] Lee, C.-H. (1988). On robust linear prediction of speech. *IEEE Trans. on Acoustics, Speech and Signal Processing (ASSP)*, 36(5): 642–650.

[6] Lobo, A. P.; Ainsworth, W. A. (1992). Evaluation of a glottal ARMA model of speech production. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP92)*, vol. 2, pp. 13–16, San Francisco.

[7] Ma, C.; Kamp, Y.; Willems, L. F. (1993). Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(1): 69–81.

[8] Markel, J. D.; Gray Jr., A. H. (1976). *Linear prediction of speech.* Communication and Cybernetics, 12. Springer-Verlag, Berlin, Heidelberg, New York.

[9] Strik, H.; Cranen, B.; Boves, L. (1993). Fitting a LF-model to inverse filter signals. In *Proc. of EUROSPEECH '93*, vol. 1, pp. 103–106, Technische Universität Berlin.