QUASI CLOSED PHASE ANALYSIS OF SPEECH SIGNALS USING TIME VARYING WEIGHTED LINEAR PREDICTION FOR ACCURATE FORMANT TRACKING

Dhananjaya Gowda, Manu Airaksinen, Paavo Alku

Dept. of Signal Processing and Acoustics, Aalto University, Finland

{dhananjaya.gowda, manu.airaksinen, paavo.alku}@aalto.fi

ABSTRACT

Recent research on temporally weighted linear prediction shows that quasi closed phase (QCP) analysis of speech signals provides better modeling of the vocal tract and the glottal source. Quasi closed phase analysis gives more weightage on the closed phase of the glottal cycle, at the same time deemphasizing the region around the instant of significant excitation which is often poorly predicted. However, all the traditional analysis techniques including the QCP analysis is performed over short intervals of time. They do not impose any continuity constraints either on the vocal tract system or the glottal source. Such constraints are often imposed at a later stage to either smooth or track the estimated features over time. Time varying linear prediction (TVLP) provides a framework for modeling speech with a long-term continuity constraint imposed on the vocal tract shape. In this paper, we propose a new method for accurate modeling and tracking of the vocal tract resonances by integrating the advantages of a QCP analysis with that of TVLP. Formant tracking experiments show consistent improvement in performance over traditional LP or TVLP methods under a variety of conditions including different voice types and over a wide range of fundamental frequency.

Index Terms— Quasi closed phase analysis, time varying linear prediction, weighted linear prediction, formant tracking

1. INTRODUCTION

Linear prediction (LP) analysis of speech signals is widely used to approximate the speech production apparatus as a source-system model [1, 2]. The conventional LP modeling gives equal emphasis towards predicting all the samples within an analysis frame. However, temporally weighted LP algorithms give differential emphasis on the samples by defining a weight function on the error signal being minimized [3–8]. A weight function based on the short-time energy variations within the glottal cycle provides robustness against degradations [3–5]. However, such an energy based weight function does not take into account the presence of impulse-like excitations within voiced speech, which often leads to high prediction errors. Quasi closed phase (QCP) analysis of speech tries to address this problem by designing a weight function that emphasizes the closed phase of the glottal cycle while at the same time deemphasizing the region around the instants of significant excitation [7, 8].

Speech signal is conventionally analyzed over short segments (5-50 ms) with an inherent assumption of quasi stationarity [1]. This assumption is justified to an extent that the vocal tract apparatus being a physical inertial system tends to change slowly within such short time intervals and over adjacent glottal cycles. This natural redundancy built into the speech production apparatus is utilized only to the extent of estimating an average vocal tract model over a few adjacent glottal cycles. The continuity of the vocal tract system beyond the window sizes of the short-time analysis is seldom utilized. Some amount of continuity is enforced in this piecewise approximation of the vocal tract system by overlapping the adjacent segments of analysis. Time varying linear prediction (TVLP) tries to bridge this gap by modeling the speech signal over longer intervals of time by defining the vocal tract model parameters as a function of time [9–11].

The conventional least squares solution to the LP problem involves minimizing the L_2 norm of the excitation source signal approximated by the prediction error signal [12, 13]. The inherent assumption here is that the excitation signal is a Gaussian process. However, neither the speech signal nor the excitation signal is Gaussian in nature. Sparsity constraints based on the theory of compressed sensing may be used to utilize the super Gaussian nature of the excitation signal [14, 15]. This involves minimizing the L_0 norm (the number of non-zero elements) of the error signal. However, due to the non-convex nature of the cost function the L_0 norm is often approximated by a L_1 norm optimization which provides a more tractable convex problem [14]. Also, it has been shown that an iterative reweighted minimization of the norm can achieve increased sparsity of error signal and thereby yielding a solution more closer to L_0 norm optimization [15].

In this paper, we propose a new method for accurate modeling and tracking of the vocal tract resonances which integrates the advantages of temporally weighted LP, time varying LP and sparse LP.

2. QUASI CLOSED PHASE ANALYSIS

Quasi closed phase analysis of speech signals belongs to the family of weighted linear prediction (WLP) methods. WLP analysis involves minimizing the prediction error by giving a differential emphasis to different regions of the signal within a glottal cycle. It is important to note that the weighting is on the error signal as against the traditional windowing of the signal for short-time analysis. The cost function optimized in WLP is given by

$$E = \sum_n w[n] e^2[n] \tag{1}$$

where $w[n]$ is the weight function on the error ($e[n] = x[n] - \hat{x}[n]$) in predicting the current speech sample ($x[n]$) based on the past p samples (given by $\hat{x}[n] = -\sum_{k=1}^{p} a_k x[n-k]$). The prediction coefficients a_ks represent the vocal tract system modeled as an all-pole system of order p. Different weight functions have been proposed in the literature based on different criterion and for different purposes. Weight functions that follow the short-time energy of the speech signal within a glottal cycle have been used to increase the robustness of the analysis against degradations [3, 4, 6]. A weight

function inversely proportional to the prediction error signal is often used in sparse linear prediction analysis to increase the sparsity of the error signal [14, 15]. Another weight function with an attenuated main excitation (AME) reduces the effect of glottal source on the vocal tract estimation [8]. The QCP analysis uses a weight function that combines the advantages of WLP, sparse LP, and AME weight function. The weight function is designed so as to emphasize the closed phase region of the glottal cycle while at the same time attenuate the region around the main excitation [7]. Attenuation of the main excitation automatically imposes sparsity constraints on the error signal. By defining a continuous weight function on the error signal the QCP analysis provides a flexible framework to approximate a closed phase analysis over multiple glottal cycles using either an autocorrelation-based or a covariance-based formulation.

3. TIME VARYING WEIGHTED LINEAR PREDICTION

The time varying linear prediction (TVLP) formulation is very similar to the conventional LP formulation, except that the predictor coefficients are defined as functions of time instead of being constants. This allows for modeling of the speech signal over intervals of time longer ($> 50ms$) than the typical short-time analysis window lengths. In TVLP, an estimate $\hat{x}[n]$ of the current sample $x[n]$ is predicted based on the past p samples given by

$$\hat{x}[n] = \sum_{k=1}^{p} a_k[n]x[n-k] \tag{2}$$

where the time varying predictor coefficients $\{a_k[n]\}_{k=1}^{p}$ can be modeled using different approximations either using a power series or a trigonometric series [9–11]. In this paper we use the power series or polynomial approximation of the coefficients given by

$$a_k[n] = \sum_{i=0}^{q} b_{k_i} n^i. \tag{3}$$

A piecewise constant approximation using a zeroth order polynomial ($q = 0$) yields the traditional short-time analysis.

The TVLP coefficients are estimated by minimizing the L_m-norm of the error signal given by

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} ||\boldsymbol{x} - \boldsymbol{X}\boldsymbol{b}||_m \tag{4}$$

where

$$\boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]^T_{N\times 1} \tag{5}$$

$$\boldsymbol{b} = [b_{1_0}, b_{1_1}, \ldots, b_{1_q}, \ldots, b_{p_0}, b_{p_1}, \ldots, b_{p_q}]^T_{p(q+1)\times 1} \tag{6}$$

$$\boldsymbol{X} = [X_0, X_1, \ldots, X_{N-1}]^T_{N\times p(q+1)} \quad \text{and} \tag{7}$$

$$X_n = [x[n-1], nx[n-1], \ldots, n^q x[n-1], \ldots, x[n-p], nx[n-p], \ldots, n^q x[n-p]]^T_{p(q+1)\times 1}. \tag{8}$$

The traditional autocorrelation and the least square formulations typically minimize the L_2-norm. L_1-norm minimization can also be used as an alternative, and as a convex approximation of L_0-norm optimization problem, which adds an additional sparsity constraint on the error signal for better modeling of the excitation source and vocal tract system [11, 14, 15].

Time varying weighted linear prediction (TVWLP) is analogous to WLP where the predictor coefficients are estimated by minimizing a weighted error signal given by

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \boldsymbol{W}||\boldsymbol{x} - \boldsymbol{X}\boldsymbol{b}||_m \tag{9}$$

where $\boldsymbol{W}$ is a diagonal matrix with its diagonal elements corresponding to the weight function defined on the error signal. In sparse LP literature an iterative reweighted norm minimization approach is often employed to increase the sparsity of the error signal [15]. A weight function inversely proportional to the error signal in used to reestimate the predictor coefficients for the next iteration.

In this paper, we propose to use the QCP weight function within the TVWLP framework. The resultant QCP analysis using TVWLP can provide more accurate closed phase estimates of the vocal tract over multiple glottal cycles with reduced effect of the glottal source. It also imposes sparsity constraints on the excitation signal, and at the same time ensures continuity of the vocal tract system.

4. FORMANT TRACKING EXPERIMENTS

The performance of TVLP and TVWLP based QCP analysis methods in tracking vocal tract resonances is studied in this section. The initial set of experiments address issues in the conventional TVLP analysis on the choices of norm function, analysis window size, and polynomial order for the time varying parameters. Based on these observations, subsequent experiments compare the performance of the TVLP and TVWLP methods for different phonation types and a wide range of fundamental frequency. The performance is evaluated using a variety of data including artificial resonance contours, synthetic as well as real speech signals using both L_2 and L_1 norm minimization. In this paper, we use a linear programming solution to L_1-norm minimization [14] and the least squares solution to the L_2-norm optimization [12].

4.1. Choice of L_2 or L_1 norm

A comparison between L_2 and L_1 norm minimization is studied using an artificial signal as well as real and synthetic speech signals.

4.1.1. Analysis of synthetic signals

Performance of the TVLP methods using either L_2 or L_1 norm in tracking artificial resonance tracks is shown in Fig. 1. The resonance contours are simulated as a sum of three damped chirp signals with a quadratic sweep in the center frequency with time. The ground truth resonance tracks (red dashed lines) and the estimated tracks are shown for different polynomial orders. It can be seen that L_2 norm minimization is optimal at an order ($q = 3$) close but higher than the actual quadratic nature of the contours which may be necessary due to the complexity of a resonance cross over in the signal. The performance deteriorates for both lower as well as higher polynomial orders. However, L_1 norm optimization leads to consistent improvement with increasing polynomial order.

4.1.2. Analysis of real speech signals

A qualitative analysis of the TVLP methods on real speech signals using different norms and polynomial order is presented here. Fig. 2 shows the first three formant contours estimated (red lines) using an 8th order TVLP analysis for different polynomial orders using L_2 as well as L_1 norm optimization. The analysis window size used is 200 ms. It can be seen that L_1 norm minimization seem to track the reference formant values (green dots) more consistently and accurately with increasing polynomial order, as was the case with simulated resonance contours. Increasing the analysis window size and the order of polynomial seem to work well for simulated signals, but with real speech signals convergence and stability issues are often encountered resulting in poor estimates of the formant tracks. The

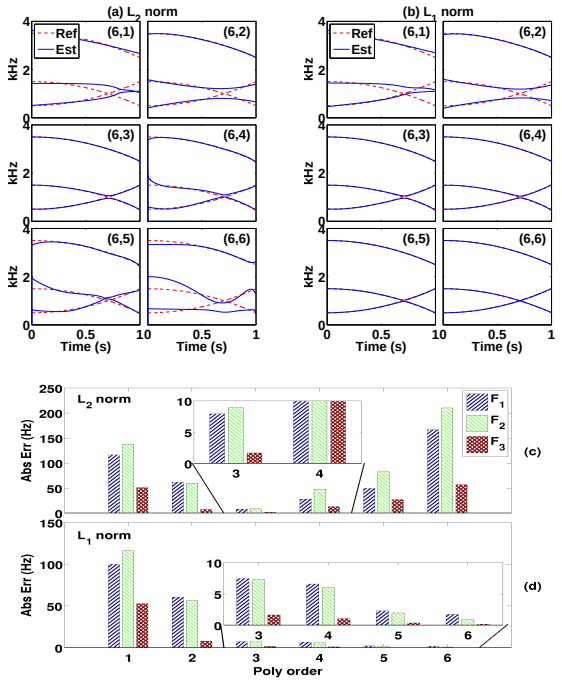

Fig. 1. TVLP analysis on sum of chirp signals with quadratic sweep. (a) and (b) Estimated resonance contours (solid blue lines) for varying polynomial order q using L_2 and L_1 norms, respectively. The reference contours are shown as dashed red lines. $(p, q) = \{(6, 1), \ldots, (6, 6)\}$ denotes (LP order, polynomial order). (c) and (d) Average errors in contour estimation using L_2 and L_1 norms, respectively.

L_1 norm minimization fails to converge for arbitrary segments of speech at higher orders and longer window size (300 ms or more). Similarly L_2 norm optimization suffers with the problem of rank deficient covariance matrices. Advanced numerical methods with additional constraints on convergence, stability and continuity of the system parameters would be essential. However using either of the norms works reasonably well for orders less than 5 and window sizes up to 300 ms.

It can be seen from Fig. 2 that the estimates (red lines) of lower formants seem to match and track the reference formant locations (green dots) better than the weaker higher formants. However, availability of the real ground truth is a major issue in evaluating the accuracy of the estimates. For example the estimates of the third formant are biased towards what appears to be a peak at around 2.5 kHz on the spectrogram at 0 ms. The reference ground truth (green dots) plotted here are part of the VTR database [16] estimated in a semi-automated way. Estimates from a short-time autocorrelation based LP analysis are smoothed and manually corrected against spectrographic evidences. In such a scenario it becomes difficult to determine which of the two estimates is accurate. In view of the lack of an authentic ground truth, the accuracy of the proposed methods will have to be evaluated on synthetic speech utterances.

4.2. Choice of window size and polynomial order

One important issue in time-variant LP modeling is the choice of the window size and polynomial order used to approximate the predictor coefficient contours. Longer analysis window sizes are attractive for efficient parameterization and coding of speech signals but would introduce longer delays. Longer segments require higher polynomial orders to approximate the contours of the vocal tract model parameters better, and can lead to computational problems as mentioned earlier. However, it is not absolutely essential to analyze speech over very long segments. Moderate window sizes that can capture the slow time varying nature of the vocal tract with lower polynomial orders may be a good compromise overall.

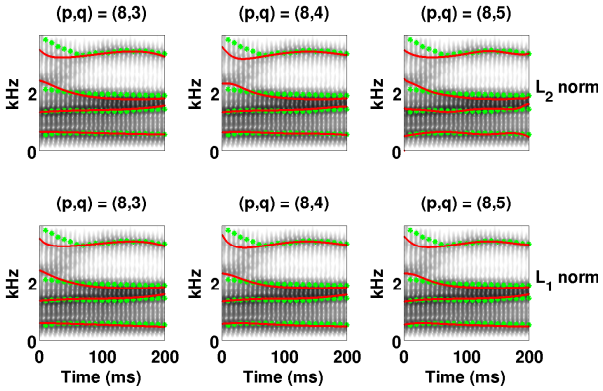

Fig. 2. TVLP analysis of real speech for different polynomial orders using L_2 (top row) and L_1 (bottom row) norms. Estimated and reference formants are shown by red lines and green dots, respectively.

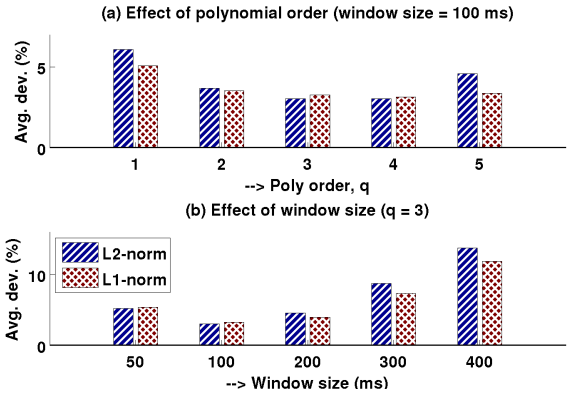

Fig. 3. Average error in formant estimation (F_1, F_2 and F_3) as a function of (a) polynomial order and (b) window size.

Fig. 3 shows the absolute error in formant estimation averaged over the first three formants of a synthetic speech utterance for different values of polynomial order and window size. Fig. 3(a) shows the performance of the TVLP methods using L_2 and L_1 norms for a fixed window size of 100 ms but with varying polynomial order q. It can be seen that optimal performance is reached at an order of $q = 3$ or $q = 4$ and the performance starts to deteriorate at an order of $q = 5$. Fig. 3(b) shows the performance for varying window sizes at a fixed polynomial order $q = 3$. It can be seen that the performance is good at 100 ms or 200 ms, and it deteriorates for longer window sizes. In the experiments to follow in the remainder of the paper we use a fixed window size of 100 ms and a polynomial order of $q = 3$.

4.3. Quasi closed phase TVWLP analysis

Performance of the proposed QCP analysis of speech signals using TVWLP in formant tracking is evaluated for a variety of conditions. Four different phonation types (creaky, modal, breathy and whis-

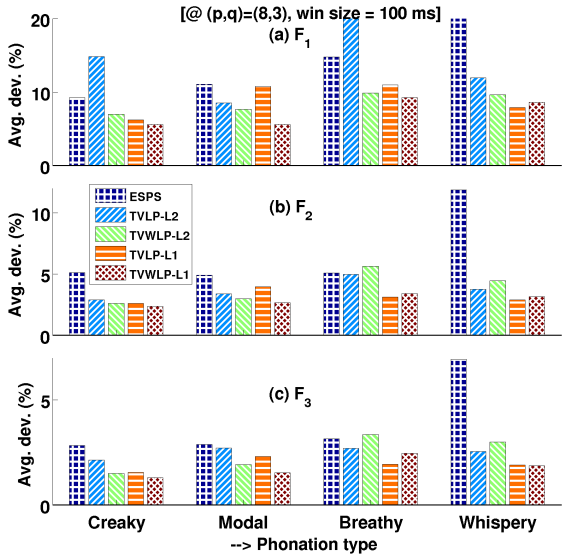

Fig. 4. Formant estimation error for different phonation types.

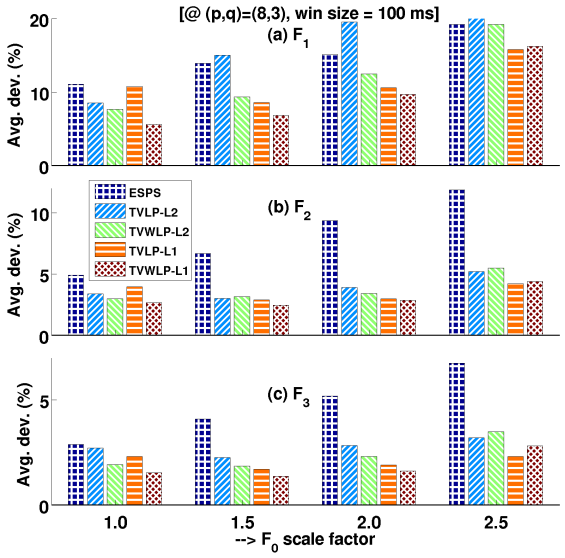

Fig. 5. Formant estimation error for different mean F_0 values.

pery) and four different ranges of fundamental frequencies (mean utterance F_0 scaled by factors 1.0, 1.5, 2.0 and 2.5) are considered. The phonation types and the F_0 range are controlled by using a parametric Liljencrants-Fant (LF) model for the glottal source [17]. Speech signals are synthesized by filtering the glottal flow derivative signal using an all-pole model with known formants and bandwidths from the VTR database [16]. Ten randomly selected utterances (5 male and 5 female) from the database are synthesized with four different phonation types and four different mean F_0. The analysis is carried out over non-overlapping window segments of 100 ms using an LP order of $p = 8$ and a polynomial order of $q = 3$. A comparison with the traditional LP covariance based method (popularly known as ESPS method [18]) used in the popular open source tool Wavesurfer [19] is also provided. It should be noted here that the ESPS method first performs a short-time (49 ms) LP analysis followed by a dynamic programming based tracking of formants.

4.3.1. Effect of phonation type

Performance of the TVLP and TVWLP methods for the four different phonation types is shown in Fig. 4. It can be seen that QCP based TVWLP method minimizing L_1 norm performs the best in most cases. In general, the L_1 norm minimization seem to perform better than using L_2 norm. Similarly, the TVWLP methods perform better than the TVLP methods. Performance of the both TVLP and TVWLP methods is better than the traditional ESPS method. The QCP based analysis improves the performance of both L_2 as well as L_1 norm based methods consistently for creaky and modal phonation types. However, QCP analysis seems to yield a mixed performance for breathy and whispery phonation types which exhibit high open quotient as well as high spectral tilt. A more careful investigation of these phonation types is essential to find out the reasons for mixed performance of the QCP methods.

4.3.2. Effect of fundamental frequency

The four different ranges of fundamental frequency are generated by scaling the original F_0 contour of an utterance by different factors before synthesizing the speech signal. A modal LF excitation is generated based on the new F_0 contour while retaining the original rate of formants and hence the speaking rate intact. Performance of the TVLP and TVWLP methods for all four ranges of F_0 values is shown in Fig. 5. It can be seen that the TVWLP methods provide consistent improvement over the TVLP methods for both L_2 as well as L_1 norm optimization up to a scale factor of 2.0. A mixed performance for the scale factor 2.5 may be due to the new F_0 values moving very close to the first formant. However, L_1 norm optimization seem to perform better than minimizing L_2 norm in most cases. The overall performance of all these methods averaged over all phonation types and F_0 ranges is given in Table 1. It can be seen that QCP based TVWLP-L1 method provides significant improvement over the traditional ESPS method.

Table 1. Overall formant tracking performance in terms of percentage deviation averaged over all phonation types and fundamental frequencies.

ESPS	TVLP-L2	TVWLP-L2	TVLP-L1	TVWLP-L1
8.74	7.44	5.54	5.18	4.42

5. CONCLUSIONS

A new method for accurate formant tracking using quasi closed phase analysis of speech signals using time varying weighted linear prediction was proposed. The proposed QCP-TVWLP method combines the accurate source-system separation capabilities of the QCP based analysis and the long-term continuity constraints on vocal tract resonances imposed by TVLP. Optimal parameters for analysis window size and polynomial order of the time varying LP coefficients were explored. A moderate window size of 100 to 200 ms and a polynomial order of 3 to 4 were found to be best suited. Experimental results show that the proposed QCP-TVWLP methods provide significant improvement in the formant tracking performance over TVLP methods, as well as the traditional ESPS method, for different phonation types and a wide range of fundamental frequency.

6. ACKNOWLEDGEMENTS

This work has been funded by the Academy of Finland (project no. 256961 and 284671).

7. REFERENCES

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[2] P. Alku, "Glottal inverse filtering analysis of human voice production a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[3] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 1, pp. 69 – 81, 1993.

[4] J. Pohjalainen, H. Kallasjoki, K. J. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis is noisy conditions," in *Proc. Interspeech*, Brighton, UK, September 2009.

[5] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[6] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilized weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401 – 411, 2009.

[7] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 3, pp. 596–607, March 2014.

[8] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear predictiona)," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, 2013.

[9] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Processing*, vol. 5, no. 3, pp. 267 – 285, 1983.

[10] K. Schnell and A. Lacroix, "Time-varying linear prediction for speech analysis and synthesis," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 3941–3944.

[11] S. Chetupalli and T. Sreenivas, "Time varying linear prediction using sparsity constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6290–6293.

[12] D. Wong, J. Markel, and J. Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 4, pp. 350–355, Aug 1979.

[13] S. M. Kay, *Modern Spectrum Estimation: Theory and Application*, Prentice Hall NJ, USA, 1988.

[14] D. Giacobello, M. Christensen, M. Murthi, S. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1644–1657, July 2012.

[15] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 317–329, April 2010.

[16] L. Deng, X. Cui, R. Pruvenok, J. Huang, and S. Momen, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Toulouse, France, 2006, pp. I–369–I–372.

[17] G. Fant, J. Liljencrants, and Q. G. Lin, "A four-parameter model of glottal flow," Q. Prog. Stat. Rep., Speech Trans. Lab., R. Inst. Technol., Stockholm, Sweden, vol. 4, pp. 1-17, 1985.

[18] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. Am.*, vol. 82, no. S1, 1987.

[19] K. Sjolander and J. Beskow, "Wavesurfer - An open source speech tool," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, October 2000, pp. 464–467.