

# Maximum Phase Modeling for Sparse Linear Prediction of Speech

Thomas Drugman, *Member, IEEE*

**Abstract**—Linear prediction (LP) is an ubiquitous analysis method in speech processing. Various studies have focused on sparse LP algorithms by introducing sparsity constraints into the LP framework. Sparse LP has been shown to be effective in several issues related to speech modeling and coding. However, all existing approaches assume the speech signal to be minimum-phase. Because speech is known to be mixed-phase, the resulting residual signal contains a persistent maximum-phase component. The aim of this paper is to propose a novel technique which incorporates a modeling of the maximum-phase contribution of speech, and can be applied to any filter representation. The proposed method is shown to significantly increase the sparsity of the LP residual signal and to be effective in two illustrative applications: speech polarity detection and excitation modeling.

**Index Terms**—Speech Processing, Linear Prediction, Maximum Phase, Sparsity, Residual Excitation

## I. INTRODUCTION

Linear prediction (LP) is an omnipresent analysis technique in speech processing. It has been successfully applied in several voice technology applications such as speech coding, synthesis, analysis or recognition [1]. LP analysis relies on a source-filter model in which the speech signal is obtained by passing an excitation through an all-pole filter. In the traditional LP analysis approach, the prediction coefficients are determined such that the  $l_2$ -norm of the residual signal (i.e. the difference between the observed and predicted signals) is minimized [2]. This is known to work rather well for unvoiced sounds where the excitation signal can be assumed to be Gaussian and independently and identically distributed [2]. Nonetheless, for voiced sounds where the excitation signal exhibits quasi-periodic strong peaks, this assumption does not hold. In this case, the hypothesized excitation source is a quasi-periodic pulse train. Minimizing the variance of the residual signal then turns out to not be an appropriate criterion, as this approach is known to suffer from problems such as overemphasis on peaks and cancellation of errors [2].

A better criterion would be to maximize the sparsity of the residual signal. In that perspective, its  $l_0$ -norm should be ideally minimized. This however yields a combinatorial optimization problem. Instead, various approaches [3] have investigated the minimization of the  $l_1$ -norm which is a convex relaxation of the  $l_0$ -norm problem and which can be solved using convex programming methods.

Across all aforementioned techniques, poles are expected to lie within the unit circle in the z-plane, otherwise the filter

is considered to be unstable [4], [3]. In the  $l_2$ -norm problem, this is guaranteed but is however not true with the  $l_1$ -norm [4]. Generally, when poles are found outside the unit circle, pole reflection is applied [3]. However, during the production of voiced sounds, the glottal flow is known to exhibit a maximum-phase (i.e. anticausal) component [1], [5], which is therefore generally not modeled in conventional LP analysis.

The goal of this paper is to propose a solution to incorporate a maximum-phase modeling in the LP analysis of speech. The proposed method can be applied to any LP-based method and is shown to significantly improve the sparsity of the residual signal. The paper is structured as follows. Section II first establishes the fundamentals of linear prediction. The existence of a maximum-phase component in speech is explained in Section III, where the motivations of this work are given. The proposed method is described in Section IV. Its efficiency is then confirmed in Section V through a comprehensive evaluation. Section VI finally concludes the paper.

## II. LINEAR PREDICTION: PROBLEM FORMULATION

The auto-regressive (AR) model of speech assumes that a speech sample  $s(n)$  can be written as a linear combination of its  $K$  past samples:  $s(n) = \sum_{k=1}^K a_k s(n-k) + r(n)$ , where  $K$  is the prediction order,  $a_k$  are the prediction coefficients and  $r(n)$  is the prediction error, also called residual signal or residue. Based on the observation of a sequence of speech samples, the optimization problem aims to find an estimate of the prediction coefficient vector  $\hat{\mathbf{a}} \in \mathbb{R}^K$  such that the prediction error is minimized. The LP analysis problem can then be written as the minimization of the  $l_p$ -norm of the residual signal.

The conventional approach considers the minimization of the  $l_2$ -norm problem which can be solved in a rapid way by exploiting the Toeplitz structure of the correlation matrix, as in the widely-used Levinson-Durbin algorithm [6]. Nonetheless, it is known that the  $l_2$ -norm criterion is highly sensitive to outliers. As a consequence, this approach will favor solutions with many small non-zero values rather than a sparse solution containing a limited number of non-zero values [3]. Sparse solutions should be preferred as the target excitation source in voiced sounds is expected to be a quasi-period impulse train.

The ideal solution maximizing sparsity involves the minimization of the  $l_0$ -norm of the residual signal. Unfortunately this leads to a combinatorial problem which cannot be solved in polynomial time. Instead, several studies have addressed solving the  $l_1$ -norm, moving closer to the original  $l_0$ -norm problem [3]. This is possible thanks to the improvements in

convex optimization algorithms (e.g. using interior point methods [7]). Other approaches have proposed a weighted  $l_2$ -norm LP analysis in which a weighting function is used to give less emphasis to the samples around a strong excitation [8], [9], [10]. Compared to the conventional  $l_2$ -norm, aforementioned methods have proved to be more efficient for coding and to provide better estimates of the spectral envelope.

### III. MOTIVATION OF THIS WORK

According to the mixed-phase model [5], speech is composed of both minimum-phase (i.e causal) and maximum-phase (i.e anticausal) components. While the vocal tract impulse response and the *return phase* of the glottal component are minimum-phase signals, the *open phase* of the glottal flow is known to be maximum-phase [5]. In [11], it was proved that the use of an anticausal all-pole filter for the glottal pulse is necessary to resolve magnitude and phase information correctly. Unfortunately, deconvolving the minimum and maximum-phase components of speech is a complex problem which suffers from robustness issues [12], [13].

Despite the mixed-phase nature of speech, existing LP analysis approaches generally apply pole reflection when poles are found outside the unit circle. The resulting residue consequently contains a maximum-phase component which has not been captured in the AR modeling. The top plot of Fig. 1 shows an example of residue obtained using the conventional  $l_2$ -norm LP analysis. Glottal Closure Instants (GCIs, [14]) appear as strong quasi-periodic discontinuities in this signal. It can be noticed that the segments preceding GCIs follow a systematic shape with clearly non-zero valued samples. This is because the residue exhibits a persistent maximum-phase component which has not been completely removed by the standard LP analysis. An intuitive way to explain this would be to reverse the time axis: it can be understood that this remaining component could be modeled by a simple AR filter with a limited number of poles. The operation of reversing the time axis is equivalent to inverting the causality of the signal. As a consequence, modeling the maximum-phase component of speech could potentially engender sparser residual signals. This is the precise goal of the proposed method. The bottom plot of Fig. 1 shows the residue when applying the proposed technique to the  $l_2$ -norm LP analysis. It can be observed that in this case the maximum-phase component has been removed almost completely.

### IV. PROPOSED METHOD

#### A. Description

The workflow of the proposed method is given in Fig. 2. Note that a Matlab implementation of this technique can be found at [tcts.fpms.ac.be/~drugman/Toolbox](http://tcts.fpms.ac.be/~drugman/Toolbox). The speech signal is first standardly framed using overlapping windows. Because mixed-phase decomposition is a challenging problem suffering from robustness issues in realistic recording conditions [12], the proposed approach circumvents this hindrance by relying on steps of preemphasis and causality inversion. The residue is obtained by two successive inverse filtering operations: in the first one, we aim at removing the minimum-phase

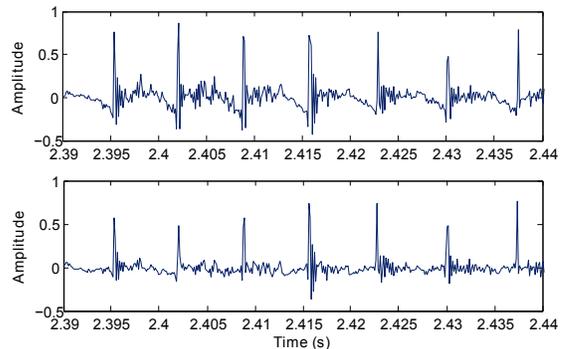


Fig. 1. Example of residual signals obtained using: *top panel*: the conventional  $l_2$ -norm LP analysis; *bottom panel*: the proposed  $l_2$ -norm LP analysis.

contribution, while the second targets removing the maximum-phase component. The coefficients of the 2 filters are estimated by LP analysis, as explained below.

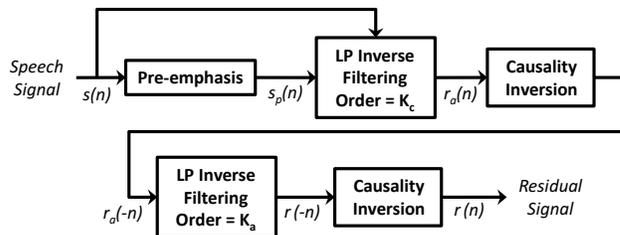


Fig. 2. Workflow of the proposed method.

The minimum-phase component of speech is mostly related to the vocal tract [5], whose dominant poles are due to the first formant  $F1$ . On the opposite, the maximum-phase contribution is due to the glottal open phase, which is characterized by the *glottal formant*  $Fg$ , whose range is known to cover  $[F_0, 3 * F_0]$  [15]. The goal of the first LP analysis (of order  $K_c$ ) is to model the minimum-phase component. In order to minimize the effect of the anticausal contribution, and maximize the impact of the causal component, the coefficients of the first LP analysis are estimated on a pre-emphasized version  $s_p(n)$  of the speech signal  $s(n)$ . Pre-emphasis is conventionally achieved by using a single real zero in  $\alpha$ . Preemphasis therefore balances the energy in the speech spectrum such that low frequencies do not dominate during the first LP analysis. The choice of  $\alpha$  used for pre-emphasis will then result from a trade-off, as it will be discussed in Section IV-B.

After the first LP inverse filtering, the resulting signal  $r_a(n)$  is mostly dominated by the anticausal component of speech, as the causal contribution has been substantially removed. The second LP analysis precisely aims at modeling this maximum-phase component. To force this, we apply causality inversion by reversing the time axis. After causality inversion, the original anticausal component is now seen as causal, and can be modeled by a standard LP analysis of order  $K_a$ . The signal  $r_a(-n)$  is then inverse filtered to get  $r(-n)$ , and the final residual signal  $r(n)$  is simply obtained by reversing the time axis in its original direction.

The success of the proposed method lies in two key concepts: *i)* since conventional LP analysis assumes the signal to be minimum-phase, reversing causality is a solution to

force the modeling of the maximum-phase component of speech; *ii*) preemphasis is essential as it guides the two successive LP analyses; its goal is to minimize the effects of the maximum-phase contribution, such that the first LP analysis is mostly driven by the minimum-phase component of speech. An advantage of the proposed method is that it can be applied to any LP analysis technique, and even transposed to other filter representations such as MFCCs or Mel-Generalized Cepstral (MGC) coefficients [16]. The experiments led in this paper however only focus on its usefulness for LP analysis.

### B. Parameter Settings

The proposed method makes use of 3 parameters: the total prediction order  $K = K_c + K_a$ , the prediction order  $K_a$  for the maximum-phase modeling, and the preemphasis coefficient  $\alpha$ .  $K$  can be fixed as done by standard LP methods. The influence of the two other parameters on the engendered sparsity is now studied on a development set containing 1000 sentences from the TIMIT corpus [17] (balanced across genders). As sparsity metrics, we use the Gini index [18] as it was the only sparsity metrics in [19] to meet the six attributes one can expect from a sparsity measure. Higher values of the Gini index imply a higher level of sparsity. Throughout our development experiments, we observed that the Gini index of the residual signal  $r(n)$  reaches higher values when  $K_a$  is fixed to 2 or 3. This goes in line with the speech production model according to which the maximum-phase component is due to the glottal open phase which can be modeled by two anticausal poles [5].  $K_a$  is therefore fixed to 2 in the remainder of this paper.

The setting of  $\alpha$  is linked with the fact that minimum and maximum-phase components are assumed to affect different spectral bands, and results from a tradeoff. It must be such that, during the first LP analysis: *i*) the minimum-phase component (whose dominant poles are due to  $F_1$ ) will be properly modeled; *ii*) the effects of the first harmonics and of the maximum-phase component (dominated by  $F_0$ ) are minimized.  $\alpha$  is then expected to be dependent in a certain extent upon  $F_0$ , and consequently upon the speaker gender.

Fig. 3 exhibits the distribution of the optimal  $\alpha$  values across the development set. For both genders, the distribution is bimodal with two clear modes in  $\alpha = -1$  and in  $\alpha = -0.7$  (this latter being more spread). As a consequence, the setting of  $\alpha$  is not univocal. The approach we adopt in the rest of this paper is then to investigate, for each frame to be analyzed, two possible values for  $\alpha$  (-1 and -0.7) and ultimately select the one engendering the greater sparsity, i.e. maximizing the Gini index of the resulting residue  $r(n)$ .

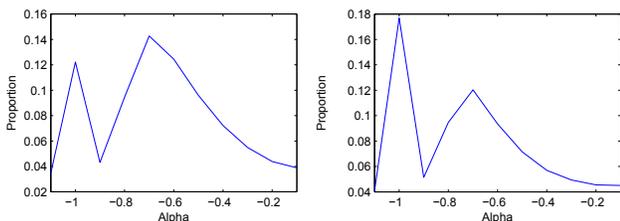


Fig. 3. Distribution of the optimal coefficient  $\alpha$  for male (left panel) and female (right panel) speakers.

## V. EXPERIMENTS

Our experiments are divided into three parts. Section V-A investigates the sparsity and computational complexity engendered by the proposed technique. Sections V-B and V-C then address its efficiency in two illustrative applications: speech polarity detection and excitation modeling. Note that a common way to assess the efficiency of a LP analysis technique is to compute the spectral distortion (SD) between a reference envelope and the estimated predictive model [3]. This would be meaningless here since the improvement brought by the proposed method relies on the exploitation of phase properties. SD calculation indeed only involves the amplitude component of the Fourier spectrum, and discards its phase counterpart.

### A. Sparsity and computational complexity

The proposed approach is here applied to 3 techniques: the conventional  $l_2$ -norm (LP2), the weighted  $l_2$ -norm (WLP2) proposed in [10] and the  $l_1$ -norm (LPI) LP analyses. WLP2 applies a weighting function to give less emphasis to the samples around GCIs [10]. GCIs are here determined using the SRH (for F0 tracking [20]) and SEDREAMS algorithms [21]. For the minimization of the  $l_1$ -norm, we use the publicly available  $l_1$ -magic toolbox [22] based on a primal-dual interior points optimization [7]. In our experiments, we compare the conventional implementation of these 3 techniques to their declined version based on the proposed MaxP (standing for maximum-phase) method. Across all techniques, framing is achieved by applying a 25ms-long Hanning window shifted every 5 ms. The 3000 longest sentences of the TIMIT corpus [17] (balanced across genders and not included in the development set) are used for the evaluation. As in [10], sentences are resampled at 8 kHz and  $K$  is fixed to 13.

Three sparsity metrics are here used to assess the performance of the LP techniques: the kurtosis, Hoyer measure [23] (which is a normalized version of the  $l_2/l_1$  measure) and the Gini index [18] of the residual signal. These 3 metrics were shown in [19] to be the 3 most appropriate measures to reflect the sparsity of a signal, as they respectively meet 3, 5 and 6 out of the 6 essential attributes of a sparsity metrics. More precisely, we consider in the following the relative improvement in sparsity of the residue over the speech signal. In other words, considering a given sparsity measure  $SM(x)$  (which can be any of the three aforementioned metrics), we evaluate the sparsity improvement as:  $\frac{SM(r(n)) - SM(s(n))}{SM(s(n))}$ , where  $s(n)$  and  $r(n)$  are the original speech signal and its residue.

Metrics	Method	LP2	WLP2	LPI
Kurtosis	Conventional	250	403	392
	MaxP	364	432	468
Hoyer	Conventional	17.6	26.0	29.4
	MaxP	24.0	29.3	30.5
Gini	Conventional	3.48	4.71	7.54
	MaxP	5.64	6.82	7.36

TABLE I  
RELATIVE IMPROVEMENT IN SPARSITY (IN %) OVER THE SPEECH SIGNAL,  
USING THE COMPARED METHODS.

Results are summarized in Table I. It can be observed that the proposed MaxP method leads to a considerable increase of

sparsity. This was reflected across 8 out of the 9 configurations (3 techniques and 3 metrics). However, for the unfavorable case (using the LP1-based techniques and the Gini index), a paired t-test revealed no significant differences, while all other results showed a statistically highly significant improvement ( $p < 0.001$ ). Finally it is worth noting that we did not observe any gender dependency through our experiments.

The computational complexity of the methods is now assessed by the Relative Computation Time (RCT), defined as the ratio between the computation time over the sound duration. Table II shows averaged RCTs obtained for our Matlab implementations and with a Intel Core i7 3.0 GHz CPU with 16GB of RAM. The proposed MaxP method results in an increase of complexity by a factor varying between 2.8 and 3.9. This is because 2 LP analyses of order  $K_a$  and 2 of order  $K_c$  are achieved, instead of only one of order  $K$ . Note also that WLP2-based techniques require in addition the estimation of GCIs, which is performed in a RCT of 7.3%.

LP2	MaxP-LP2	WLP2	MaxP-WLP2	LP1	MaxP-LP1
1.6	4.5	2.2	7.4	41.2	160

TABLE II

RELATIVE COMPUTATION TIME (IN %) FOR THE COMPARED METHODS.

### B. Application to Speech Polarity Detection

The origin of a polarity in the speech signal stems from the asymmetric glottal waveform exciting the vocal tract resonances. Detecting the speech polarity is required in various applications such as concatenative synthesis, glottal source processing or in the great majority of pitch-synchronous techniques. In [24], the Residual Excitation Skewness (RESKEW) approach has been proposed to automatically determine the speech polarity. RESKEW exploits the statistical skewness of two excitation signals: the LP residue, and a rough approximation of the glottal source. Since the skewness is known to be a measure of the asymmetry of a probability density function, it is used here as an estimator of the asymmetry of the glottal excitation. As the LP residue and the glottal source are known to have an opposite polarity, the sign of their differenced skewness indicates the speech polarity [24].

Fig. 4 displays the distribution of the differenced skewness using either the traditional LP2 or the proposed MaxP-LP2 residual signal. This is done for the 10 corpora considered in [24], covering in total 7.5 hours of speech. Note that we here used MaxP-LP2 for the fact that it does not require GCIs. This is because polarity detection is the very first step in an analysis workflow and should then be as simple as possible. The advantage of using MaxP-LP2 is clearly noticed in Fig. 4. The distribution of the differenced skewness is indeed observed to move off from zero, reducing therefore considerably the risk of confusion in the polarity determination.

### C. Application to Excitation Modeling

The usefulness of the proposed technique will now be studied in the context of excitation modeling. In [25], the Deterministic plus Stochastic Model (DSM) of the residual signal was proposed. DSM consists of two contributions acting in two distinct spectral bands delimited by a maximum voiced

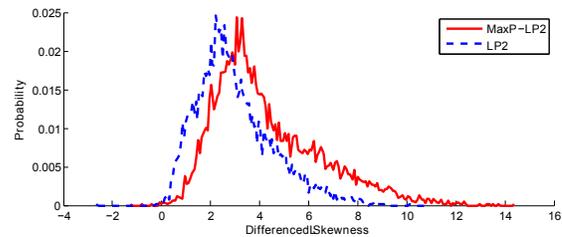


Fig. 4. Distribution of the differenced skewness used in the RESKEW method to determine the speech polarity.

frequency. The deterministic part models the periodicity in the low frequencies, while the stochastic component is a time-modulated high-frequency noise accounting for the glottal turbulences. Both components are extracted from an analysis performed on a dataset of GCI-synchronous windowed residual frames. The deterministic component arises from an orthonormal decomposition led on this dataset, which is achieved by Principal Component Analysis (PCA, [26]). It has been shown in [25] that the resulting first eigenvector can be assumed to model the deterministic component of the residue.

Fig. 5 shows the first eigenvector for the male speaker AWB from the CMU-ARTIC database [27], using the LP2 and MaxP-WLP2 techniques. WLP2 is here used since GCIs are already necessary for DSM to perform a pitch-synchronous analysis. In the case of LP2, the first eigenvector exhibits a waveform at the left of the GCI which is similar to what is described by models of the glottal source [28]. This is because the LP2 residue exhibits a persistent maximum-phase component which has not been eliminated. As a consequence, this component is reflected in the open phase of the first eigenvector. Contrastingly, the first eigenvector obtained with MaxP-WLP2 is very close to a Dirac pulse and its open phase is almost completely flat. When inspecting the eigenvalues, we observed that MaxP-WLP2 allows to cover a comparable dispersion with a reduced number of eigenvectors, which makes it interesting for speech coding.

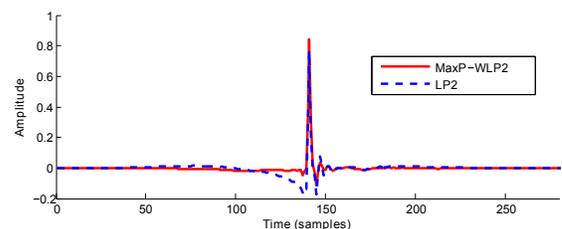


Fig. 5. First eigenvector for the male speaker AWB.

## VI. CONCLUSION

The goal of this paper was to propose a novel approach to the problem of sparse LP analysis. The proposed method aims to integrate a modeling of the maximum-phase component of speech, which is discarded in existing LP-based techniques. It has also the advantage of being applicable to any filter representation. The resulting technique was shown to dramatically increase the level of sparsity, independently of the LP method it is applied to. This was achieved with a minor augmentation of the computational complexity. Finally, its potentiality was confirmed in two illustrative applications: polarity detection and excitation modeling.

## REFERENCES

- [1] T. Quatieri: *Discrete-time speech signal processing*, Prentice Hall, 2002.
- [2] J. Makhoul: *Linear Prediction: A Tutorial Review*, Proc. IEEE, vol. 63, no. 4, pp. 561–580, April 1975.
- [3] D. Giacobello, M. Christensen, M. Murthi, S. Jensen, M. Moonen: *Sparse Linear Prediction and Its Applications to Speech Processing*, IEEE Trans. on Audio, Speech and Language Processing, vol. 20, Issue 5, pp. 1644–1657, 2012.
- [4] L. Knockaert: *Stability of linear predictors and numerical range of shift operators in normed spaces*, IEEE Trans. on Inf. Theory, vol. 38, no. 5, pp. 1483–1486, 1992.
- [5] T. Drugman, B. Bozkurt, T. Dutoit: *Causal-anticausal Decomposition of Speech using Complex Cepstrum for Glottal Source Estimation*, Speech Communication, vol. 53, Issue 6, pp. 855–866, 2011.
- [6] G. Cybenko: *The Numerical Stability of the Levinson-Durbin Algorithm for Toeplitz Systems of Equations*, SIAM J. Sci. and Stat. Comput., vol. 1, Issue 3, pp. 303–319, 1980.
- [7] S. Boyd, L. Vandenberghe: *Convex Optimization*, Cambridge University Press, 2004.
- [8] C. Ma, Y. Kamp, L. Willems: *Robust signal selection for linear prediction analysis of voiced speech*, Speech Communication, vol. 12, no. 1, pp. 69–81, 1993.
- [9] C. Magi, J. Pohjalainen, T. Backstrom, P. Alku: *Stabilized weighted linear prediction*, Speech Communication, vol. 51, no. 5, pp. 401–411, 2009.
- [10] V. Khanagha, K. Daoudi: *An efficient solution to sparse linear prediction analysis of speech*, EURASIP Journal on Audio, Speech, and Music Processing, vol. 3, 2013.
- [11] W. Gardner, B. Rao: *Noncausal All-pole Modeling of Voiced Speech*, IEEE Trans. on Audio and Speech Processing, vol. 5, issue 1, pp. 1–10, 1997.
- [12] T. Drugman, B. Bozkurt, T. Dutoit: *A Comparative Study of Glottal Source Estimation Techniques*, Computer Speech & Language, Elsevier, vol. 26, issue 1, pp. 20–34, 2012.
- [13] T. Drugman, B. Bozkurt, T. Dutoit: *Chirp decomposition of speech signals for glottal source estimation*, ISCA Workshop on Non-linear Speech Processing, 2009.
- [14] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit: *Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review*, IEEE Trans. on Audio, Speech and Language Processing, vol. 20, Issue 3, pp. 994–1006, 2012.
- [15] B. Bozkurt, B. Doval, C. d’Alessandro, T. Dutoit: *A method for glottal formant frequency estimation*, Proc. ICLSP, 2004.
- [16] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai: *Mel-generalized cepstral analysis - a unified approach to speech spectral estimation*, Proc. ICLSP, 1994.
- [17] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, V. Zue: *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. Tech. rep.*, U.S. Dept. of Commerce, NIST., 1993.
- [18] S. Rickard, M. Fallon: *The Gini index of speech*, Conf. on Information Sciences and Systems, 2004.
- [19] N. Hurley, S. Rickard: *Comparing measures of sparsity*, IEEE Trans. Inf. Theory, vol. 55, pp. 4723–4740, 2009.
- [20] T. Drugman, A. Alwan: *Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics*, Proc. Interspeech, 2011.
- [21] T. Drugman, T. Dutoit: *Glottal Closure and Opening Instant Detection from Speech Signals*, Proc. Interspeech, 2009.
- [22] E. Candes, J. Romberg: *LI-MAGIC: Recovery of sparse signals via convex programming*, California Institute of Technology, Pasadena, 2005.
- [23] P. Hoyer: *Non-negative Matrix Factorization with Sparseness Constraints*, Journal of Machine Learning Research, vol. 5, pp. 1457–1469, 2004.
- [24] T. Drugman: *Residual Excitation Skewness for Automatic Speech Polarity Detection*, IEEE Signal Processing Letters, vol. 20, issue 4, pp. 387–390, 2013.
- [25] T. Drugman, T. Dutoit: *The Deterministic plus Stochastic Model of the Residual Signal and its Applications*, IEEE Transactions on Audio, Speech and Language Processing, vol. 20, Issue 3, pp. 968–981, 2012.
- [26] I. Jolliffe: *Principal Component Analysis*, Springer Series in Statistics, 2005.
- [27] J. Kominek, A. Black: *The CMU Arctic Speech Databases*, SSW5, pp. 223–224, 2004.
- [28] G. Fant, J. Liljencrants, Q. Lin: *A four parameter model of glottal flow*, STL-QPSR4, pp. 1–13, 1985.