# ZERO PHASE SPEECH REPRESENTATION FOR ROBUST FORMANT TRACKING

*Dayana Ribas González*[1,2], *Eduardo Lleida Solano*[2], *José R. Calvo de Lara*[1]

[1]Advanced Technologies Application Center (CENATAV), La Habana, Cuba
[2]Communications Technology Group (GTC), Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain

## ABSTRACT

In this paper we present a speech representation based on the Linear Predictive Coding of the Zero Phase version of the signal (ZP-LPC) and its robustness in presence of additive noise for robust formant estimation. Two representations are proposed for using in the frequency candidate proposition stage of the formant tracking algorithm: 1) the roots of ZP-LPC and 2) the peaks of its group delay function (GDF). Both of them are studied and evaluated in noisy environments with a synthetic dataset to demonstrate their robustness. Proposed representations are then used in a formant tracking experiment with a speech database. A beam search algorithm is used for selecting the best candidates as formant. Results show that our method outperforms related techniques in noisy test configurations and is a good fit for use in applications that have to work in noisy environments.

***Index Terms***— zero phase, linear predictive coding, group delay function, formant tracking.

## 1. INTRODUCTION

Formant tracking has been of interest to the scientific community for a long time mainly due to its application in various areas of speech processing. Formants have been used in speech phoneme recognition and speech recognition [1], speech synthesis and as discriminative features for speaker recognition [2]. In real life, speech processing applications need to perform in noisy conditions. Therefore, the current challenge lies in developing algorithms that can achieve accuracy in the presence of noise.

Traditionally, formant candidates are obtained by solving the polynomial roots of Linear Predictive Coding (LPC) [3] of the speech signal. LPC is able to detect the poles in the estimated spectral envelope with great accuracy. However, under noisy conditions it is difficult for LPC to obtain the correct solution since the noise signal has it own roots which are different from the roots of the signal and both of these roots combine nonlinearly to give rise to new roots in Region of Convergence (ROC). For broadband stationary noise, it is less likely that the roots of the noise lie where the formants are located, close to unit circle. On the other hand, narrow band noise can produce roots that are close to unit circle. These roots can be confused with formants or may combine with formants to produce new shifted roots.

Previous works have addressed this issue of robustness to noise using autocorrelation, with some interesting findings. McGinn and Johnson [4] showed that in the presence of Additive White Gaussian Noise (AWGN) the autocorrelation sequence is more robust compared to the noisy signal in preserving the pole positions of the original (non-noisy) signal. Later, Mansour and Juang [5] generalized the autocorrelation method to a speech production model and showed that similar properties hold under wide band noise. They proposed a spectral representation of all pole sequences called Short-time Modified Coherence (SMC) and applied it to the problem of isolated word recognition. Similarly, other authors [6], [7] have used autocorrelation sequence for representing speech in a noisy speech recognition system. All of these works show that in the context of formant tracking, the limitations of LPC in noisy environments can be overcome by taking advantage of robustness and pole-preserving properties of the autocorrelation domain. To this end, we look into Zero Phase (ZP) representation as a type of transformation in autocorrelation domain. ZP has previously been used for wide-band noise reduction [8] and applying Chirp Z-transform combined with Group Delay Function (GDF) for speech representation in speech recognition [9] [10] and glottal flow estimation [11]. GDF have been used in formant tracking too [12], [13], [14] with interesting results.

In this work, we propose two new representations based on ZP-LPC: a phase spectral representation using GDF and a roots representation. Both representations are then evaluated in terms of robustness for the problem under consideration: formant tracking for obtaining robust formant candidates in noisy environments. This work presents the continuation of our previous works [15], [16].

The paper is organized as follows: Section 2 presents the proposed representations along with mathematical background and motivation for using ZP. Some experiments on synthetic datasets are also presented in this section for performance evaluation of LPC and ZP-LPC. In Section 3, the

This work has been partially funded by Spanish goverment and European Union (FEDER) under project INNPACTO-2011-1696-390000.

proposed representations are used in a formant tracking task and experiments are presented in both noisy and clean conditions. Finally Section 4 provides discussions and conclusions.

## 2. ZERO PHASE VERSION OF THE SIGNAL

First we introduce the zero phase version of the signal. Consider a speech production model with an all pole filter that models the vocal tract. The output of the model is the speech signal $x(n)$ and the resonance frequencies of the vocal tract are the speech formants.

The ZP version $x_{zp}(n)$ of the signal $x(n)$ is computed by first taking the absolute value of the Fourier Transform presentation of the signal and setting the phase to zero (effectively removing it). Let the Fourier Transform [17] coefficient $(X(e^{jw}))$ of $x(n)$ be given by:

$$X(e^{jw}) = |X(e^{jw})|e^{j\angle X(e^{jw})} \tag{1}$$

The inverse Fourier transform of the signal magnitude represents then the zero phase signal in the time domain. Mathematically, the ZP version of $x(n)$ is given by:

$$x_{zp}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{jw})|e^{jwn}dw \tag{2}$$

### 2.1. Pole preservation and robustness

The autocorrelation sequence is more robust to wide band noise than the original signal sequence. Previous works [4], [5] demostrated this by analyzing the Signal to Noise Ratio (SNR) variation in the autocorrelation of a noisy signal compare that of the original signal. A function is considered pole preserving if applying to an all pole sequence resulting in another sequence with the same poles.

### 2.2. From autocorrelation to Zero Phase sequence

As it is well known the autocorrelation sequence is an example of zero phase signal. So, the zero phase sequence could be generalized as in equation (3):

$$x_{zp}(n) = IDFT(|X(e^{jw})|^{\beta}) \tag{3}$$

where $\beta$ is an integer number. If $\beta = 2$, we have the autocorrelation function.

ZP has the same properties of robustness and pole-preservation as the autocorrelation since for real sequences the square root just changes the numerical value of the magnitude but the roots remain unaffected. Therefore, the frequency response of ZP will be the square root of the power spectral density and the dynamic range will be less. This is an advantage in the context of formant tracking because in the autocorrelation domain ($|X(e^{jw})|^2$) the pitch is very high, specially for female voices. Then LPC estimation can choose pitch peaks instead formants peaks. Then the smaller dynamic range allow more pronounced pitch and formant interaction [5].

### 2.3. ZP-LPC representation

As illustrated in the previous section, ZP can help prevent the shifting of roots due to noise. At the same time, LPC is very good at capturing the formant structure. Therefore we propose a method that takes the best from both. We use the ZP representation for making the signal robust to broadband noise while still preserving the poles, followed by LPC for formant tracking. From this representation, there are two ways of obtaining formant candidates: 1) by finding the peaks of the spectrum or 2) by calculating the roots of the prediction model. For the spectral representation, we use phase spectrum with its GDF since it provides a better resolution between formants as compared to magnitude spectrum. The steps from the raw noisy signal to extraction of format candidate are:

1. Segment speech signal $x(n)$ with framesize = 49 ms and overlap = 10 ms
2. Pre-emphasize each segment with a high pass filter of the form: $1 - 0.7z^{-1}$ and windowing with a Blackman window of 49 ms
3. Compute ZP using the procedure in equation (2)
4. Compute the LPC over $x_{zp}(n)$ to find the predictor coefficients
5. Obtain the roots of polynomial estimated in step 4
6. Compute the GDF [17] from parameters obtained in 4

### 2.4. Effects on Spectral representation: LPC vs ZP-LPC

This section investigates the observed effects of noise on the spectral representation using LPC and the proposed ZP-LPC. We evaluate this by comparing the variation in the GDF of LPC as well as ZP-LPC for a speech signal corrupted with additive white noise at SNR = 0, 5, 15, 25 dB. The resulting GDF are given in Fig. 1. For reference, the LPC and ZP-LPC computed for the clean signal are also given (in blue) in Fig. 1.

The figure shows that the GDF calculated with LPC has greater variation than the one computed with ZP-LPC. At SNR of 5 db, LPC already shows high variation while ZP-LPC is comparatively stable. As the signal degrades, both representation show variations due to added noise but ZP-LPC is still more stable compared to LPC at every noise level.

### 2.5. Effects on Roots: LPC vs ZP-LPC

This section evaluates how noise affects the roots estimation. Evaluation is carried out by comparing the mean distance to root for both LPC and ZP-LPC with a known clean reference.

In order to carry out the evaluation, 150 instances of white noise signal were generated and filtered with an all-pole filter with 3 roots located at 500, 1500 and 2500 Hz, imitating

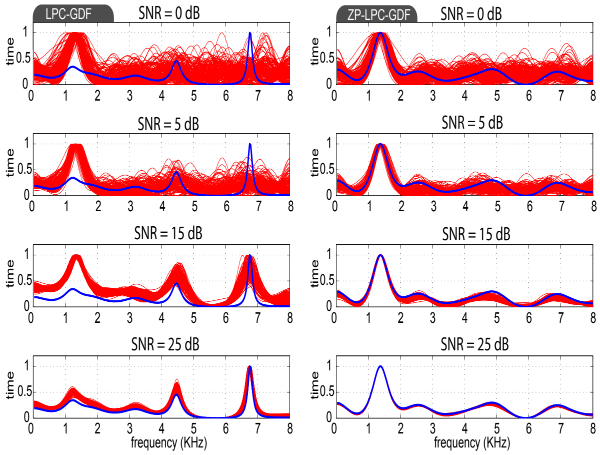

**Fig. 1**. *Group delay spectrum of the LPC and ZP-LPC for a speech signal corrupted with 150 instances of additive white noise at SNR = 0, 5, 15, 25 dB.*

the vocal tract. These serve as the reference signals. To each such reference signal three types of additive noise (stationary broadband noise, bandpass noise centered at 1500 Hz and lowpass pink noise) at 5 different SNR levels (0-20 with steps of 5 db) were added, simulating very noise to almost clean environments. This gave us 15 different noise configuration for each reference signal. For each signal in a given configuration, the roots for LPC and ZP-LPC were found and the distance to the root of the reference signal on the z-plane were computed. The Fig. 2 shows the mean for each configuration. A $12th$ order filter was used in both cases for the estimation of the three roots.

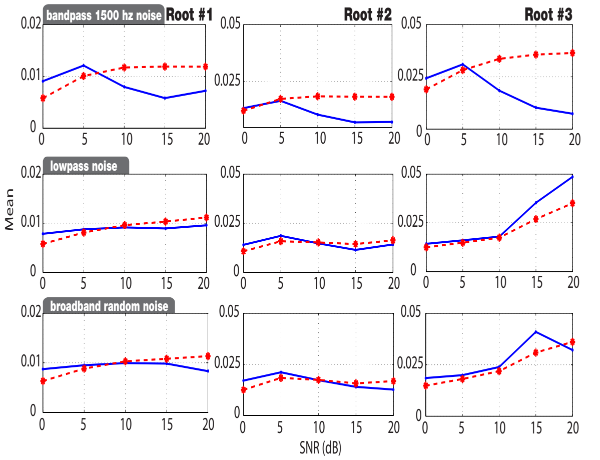

**Fig. 2**. *Mean distances between reference roots and signal roots in varying noise conditions (LPC: blue solid line, ZP-LPC: red dotted line).*

As can be seen from Fig. 2, for lowpass and broadband noise ZP-LPC outperforms LPC especially in regions of high noise (SNR $<$ 10 db). In the case for bandpass noise, Fig. 2 shows the same performance for high noise regions (SNR $<$ 5 dB). This demonstrates that this technique is as good as LPC in case of narrow band noise and can replace LPC without any loss of performance.

## 3. FORMANT TRACKING EXPERIMENTS

Having established the performance of the proposed method, we now apply it to our application of interest i.e. formant tracking. The experiments were carried out with the VTR-database [18]. For measuring the performance, absolute error between estimated and labeled formants for each signal was computed. Then, in order to capture a representative measure of dataset error avoiding outliers, the median of all the absolute errors were obtained (Median of Absolute Errors: MeAE) in voiced frames. For that purpose the pitch extractor ESPS algorithm from Snack toolkit [1] was used.

The tests were conducted in two stages: first with the clean signals from the database for selecting the ZP-LPC configuration i.e whether GDF performs better on real signals or finding the roots. In the second stage, we carry out the task of formant tracking with noise added to the clean signals. The noise types used were: stationary white noise and a type of non-stationary noise consisting of speech called babble noise, both of them from NOISEX 92 [19], pseudo-stationary street noise from AURORA 2 database [20], nonstationary music noise from a highly harmonic segment of the song November Rain of Guns and Roses band. All noise types were added at SNRs from -15 to 20 dB, making a total of 32 noise configurations. In order to compare the formant candidate stage we use our previous approach based on LPC and beam-searching algorithm (LPCiber) [16]. In that previous work LPCiber reached the best performance compared with known formant tracking methods: Wavesurfer from Snack Toolkit, Welling-Ney [21] and Mustafa [22]. Then we inserted the new formant candidate proposals, ZP-LPC with GDF as well as roots, in that formant tracking system for using the same formant selection algorithm (beam-search).

### 3.1. Results and discussions

In the following, results for the first three manually labeled formants are presented. The database does not label the fourth formant and therefore it can not be considered ground truth. The results for formant tracking with both the proposed methods on clean signals in MeAE(Hz) ar in Table 1.

It can be seen the ZP-LPC-roots outperforms the GDF representation in every case (lower MeAE for all formants). GDF exhibits larger errors because in presence of small bandwidth poles mainly in frames with two very close formants, GDF is not able to separate them despite its high resolution properties. It ends up detecting a mixed peak for both the

[1] Snack toolkit: http://www.speech.kth.se/wavesurfer

**Table 1**. *MeAE(Hz) for formant tracking in clean database using both representations from the proposal ZP-LPC: GDF (ZP-LPC-GDF) and roots (ZP-LPC-roots).*

| Methods | F1 | F2 | F3 |
|---|---|---|---|
| ZP-LPC-GDF | 21.4129 | 50.4828 | 75.5479 |
| ZP-LPC-roots | 17.0628 | 31.2180 | 52.1793 |

formants. Roots on the other hand are able to handle this particular case very well.

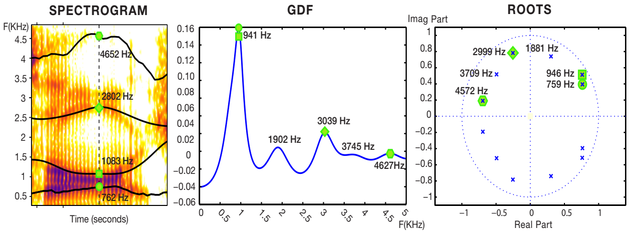

**Fig. 3**. *Spectrogram of a signal segment labeled with the ground truth and formant candidates from ZP-LPC-GDF and ZP-LPC-roots. Green figures forms are for signalize which formant candidate match with the ground truth formant.*

An example of this case is shown in Fig. 3. There we show the spectrogram of a VTRdatabase signal segment focusing on a frame with F1 and F2 very close and labeled with the ground truth, plus the formant candidates obtained from ZP-LPC-GDF and ZP-LPC-roots in that frame. Green marks there signify which formant candidate match with the the ground truth. ZP-LPC-GDF proposal is mistaken in the first candidate (it is a mixture of the real F1 and F2), ZP-LPC-roots achieves two right candidates for F1 and F2. In noisy environments, this situation is made even worse by new peaks formed due to the added noise which can be mistaken to be a formant. Due to the better performance of ZP-LPC with roots, we choose this configuration for the final set of experiments.

The main findings of this work are presented in Fig. 4. It shows a comparison of formant tracking algorithms with the noise configurations discussed earlier. It can be seen that the proposed method (ZP-LPC) achieves significant improvement over the previous approach (LPCiber). However the behavior of ZP-LPC is affected by the type of noise. Notice that the maximum MeAE for the proposal is around 100 Hz, while LPC reaches up to 240 Hz. Considering the overall results, ZP-LPC is a better algorithm.

For stationary white noise ZP-LPC-roots always has better performance compared to LPCiber. The most notable is the improved achieved in the third formant (F3) of about 140 Hz. This is in agreement with the theoretical findings of McGinn and Johnson [4] (see Section 1). Similarly, for narrow band noises (street and music) ZP-LPC outperforming LPC. These results indicate that robust property of autocorrelation sequence holds in narrow band noise as well.

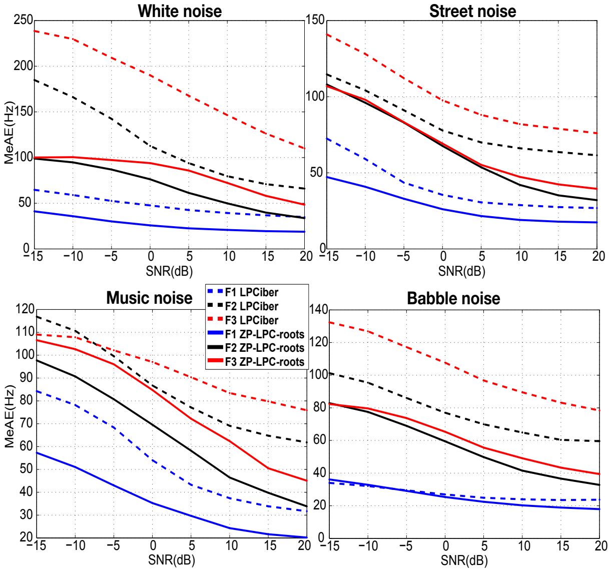

**Fig. 4**. *Median Absolute Error (MeAE) in Hz for the formants proposed by LPC, ZP-LPC-roots versus the reference in VTR-Database.*

Finally for the babble noise the proposal has a slightly lower performance in F1 for low SNR. Notice that babble noise is composed of speech, thus there are other formants frequencies acting as several narrow band noises interfering with the target roots. Under such conditions, the task of finding the correct formants is increasingly complex. Despite that the proposed method achieves improved performance for F2 and F3.

## 4. CONCLUSIONS

In this paper the study of a spectral speech representation based on the ZP transformation of the signal was presented. The proposal was applied to the candidates proposal stage of a formant tracking system in order to increase its robustness. The proposed approach offers interesting results. For stationary white noise the proposed ZP-LPC outperforms LPC representation. Performace is improved for the narrow band nonstationary noise case as well.

This work together with the previously mentioned [5] [6] have demonstrated the value of the autocorrelation domain operations in order to achieve robustness. However in previous works they have been used for handling stationary noises. Results in section 3 showed how this transformation provide robustness for nonstationary noises too. This fact makes ZP transformation a useful tool for noise compensation methods in cases when the type of noise is unknown. Finally we want to point out that the proposed representation is not useful exclusively for formant tracking task, it can also be used as robust features in another speech processing task, such as speaker or speech recognition.

## REFERENCES

[1] France Mihelic and Janez Zibert, *Speech Recognition. Technologies and Applications*, In-teh, 2008.

[2] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, *Springer Handbook of Speech Processing*, Springer-Verlag, Berlin, Germany, 2008.

[3] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE TASSP*, vol. 22, pp. 135–141, 1974.

[4] D.P. McGinn and D.H. Johnson, "Reduction of all-pole parameter estimation bias by successive autocorrelation," in *Proc. Int. Conf. Acoust., Speech, Signal Process (ICASSP)*, 1983, pp. 1088–1091.

[5] D. Mansour and B.H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Transaction on Audio, Speech and Signal Proc..*, vol. 37, no. 6, pp. 795–804, 1989.

[6] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 5, no. 1, pp. 80–84, 1997.

[7] J.A. Cadzow, "Spectral estimation: An overdetermined rational model equation approach," *Proc. IEEE*, vol. 70, pp. 907–939, 1982.

[8] W. Thanhikam, A. Kawamura, and Y. Iiguni, "Noise suppression based on replacement of zero phase signal," in *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2011.

[9] B. Bozkurt, *Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and lter characteristics of speech signals*, Ph.D. thesis, Facult Polytechnique de Mons, 2005.

[10] Baris Bozkurt, Laurent Couvreur, and Thierry Dutoit, "Chirp group delay analysis of speech signals," *Speech Communication*, vol. 49, pp. 159176, 2007.

[11] T. Drugman, Baris Bozkurt, and Thierry Dutoit, "Glottal source estimation using an automatic chirp decomposition," *Journal of NOLISP, LNAI*, vol. 5933, pp. 3542, 2010.

[12] Joseph M. Anand, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. ICSLP*, 2006.

[13] H.A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech," Tech. Rep., Indian Academy of Sciences, 2011.

[14] D. Gowda, J. Pohjalainen, M. Kurimo, and P. Alku, "Robust formant detection using group delay function and stabilized weighted linear prediction," in *Interspeech*, 2013.

[15] Jose Enrique Garcia Lainez, Dayana Ribas Gonzalez, Antonio Miguel Artiaga, Eduardo Lleida Solano, and Jose Ramon Calvo De Lara, "Beam-search formant tracking algorithm based on trajectory functions for continuous speech," in *CIARP*. 2012, pp. 749–756, Springer-Verlag Berlin Heidelberg.

[16] Dayana Ribas Gonzalez, Jose Enrique Garcia Lainez, Antonio Miguel, Alfonso Ortega Gimenez, Eduardo Lleida, and Jose Ramon Calvo de Lara, "Evaluation of a new beamsearch formant tracking algorithm in noisy environments," *CCIS*, 2012.

[17] Alan V. Oppenheim and Ronald Schafer, *Discrete Signal Processing*, Wiley, 1989.

[18] 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 *Proc. Int. Conf. Acoust., Speech, Signal Process (ICASSP)*, 2006, pp. 369–372.

[19] A. Varga and H.J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, 1992.

[20] Hans gunter Hirsch and David Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *6th International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, 2000, vol. 2000, pp. 16–19.

[21] Lutz Welling and Hermann Ney, "Formant estimation for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, 1998.

[22] K. Mustafa and I.C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Transactions on Speech and Audio Processing*, 2006.