EPOCH ESTIMATION FROM A SPEECH SIGNAL USING GAMMATONE WAVELETS IN A SCATTERING NETWORK

Pavan Kulkarni[1]*, Jishnu Sadasivan*[1]*, Aniruddha Adiga*[2]*, and Chandra Sekhar Seelamantula*[1]

[1]Department of Electrical Engineering, Indian Institute of Science, Bengaluru - 560012, India
[2]Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, USA
{pavankulkarni, sadasivan, css}@iisc.ac.in, aniruddha@virginia.edu

ABSTRACT

In speech production, epochs are glottal closure instants where significant energy is released from the lungs. Extracting an epoch accurately is important in speech synthesis, analysis, and pitch oriented studies. The time-varying characteristics of the source and the system, and channel attenuation of low-frequency components by telephone channels make estimation of epoch from a speech signal a challenging task. In this paper, we propose a new technique that employs a Gammatone wavelet filterbank and compute a scattering sequence whose local maxima define the candidate epochs in the speech signal. Results are presented for both normal and telephone channel speech by considering the differential electroglottograph from CMU-Arctic database as the ground-truth. The proposed method gives significant improvements with respect to multiple performance metrics when compared with state-of-the-art techniques for epoch estimation.

Index Terms— Glottal closure instants, Epoch estimation, Gammatone wavelets, Gammatone wavelet filterbank, Scattering network.

1. INTRODUCTION

Speech is generated by the coordination of various anatomical organs known as the articulators. Air expelled from the lungs through vocal organs in the form of puffs at the glottis causes the vocal folds to vibrate. The rate at which the vocal folds vibrate is termed as the fundamental frequency F0. During speech production, significant excitation takes place around the closing phase of the glottal waveform. The glottal closure instants (GCIs) are referred to as *epochs*.

Speech signal analysis involves determining the frequency response of the vocal-tract system and the excitation. In characterizing the excitation, accurate estimation of the epochs is of paramount importance. Measurements of F0 disturbance, jitter and shimmer, are useful in describing the voice characteristics [1]. In applications such as time- and pitch-scale modification [2–5], voice conversion and text-to-speech synthesis, epochs are used as pitch markers. Epochs are also employed in estimating the time-delay between speech signals collected using a microphone array [6]. In speaker verification applications, the excitation features derived from the regions around the epoch locations serve as discriminative features [7].

1.1. Related Literature

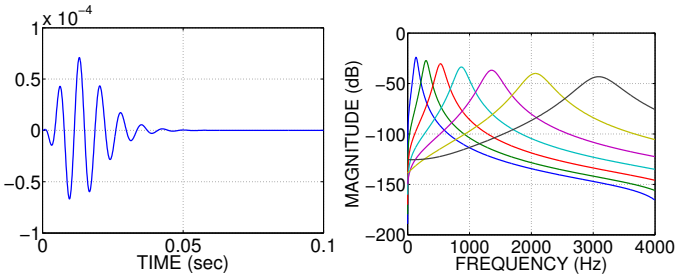
Most of the approaches for epoch estimation available in the literature are based on the source-filter model of speech production and rely on linear prediction (LP) based inverse filtering. In such methods, strong peaks in the inverse filter output are considered as epochs. Algorithms such as dynamic programming phase slope algorithm (DYPSA) [8], yet another GCI algorithm (YAGA) [9] and dynamic plosion index (DPI) [10] use LP analysis whose performance depends on the modeling capability of the vocal-tract system. Apart from inverse-filter-based approaches, alternative algorithms based on filtering the speech signal have been proposed in [11–13]. In these approaches, the speech signal is filtered at selected frequencies.

In [11], the Hilbert envelope (HE) of highpass filtered speech is used for the estimation of epochs. In [12], the speech signal is passed through a second-order zero-frequency resonator (ZFR) and then the local average in the output is removed. The positive zero-crossings of the resulting signal are choosen as the epochs. In [13], speech event detection using residual excitation and mean-based signal (SEDREAMS) algorithm has been proposed. It uses a Blackman-Tukey window to smooth the signal and computes a mean-based signal. Zero-crossings in the mean-based signal and LP residual are used to estimate the epochs. Both ZFR and SEDREAMS require a priori knowledge of the pitch period for window selection and are robust to noise, but their performance degrades in telephone channels due to the attenuation of the glottal excitation energy around zero frequency.

The approaches such as DPI [10], spectral zero crossing rate (SZCR) [14] and all-pass (AP) modeling of phase spectrum [15] show robust performance even with telephone channel speech compared to ZFR and SDREAMS. Recently, a method has been proposed using a single-pole filter (SPF) [16] to compute a time-frequency representation, the time marginal of which peaks at the epochs. Several of the proposed methods work well for clean speech and are reasonably robust to noise. However, robustness to telephone channel effects still remains a challenging problem.

1.2. This Paper

In this paper, we consider time-frequency coefficients of a speech signal obtained by using a Gammatone wavelet filterbank (GWFB). The corresponding time-frequency representation is processed using a convolution layer followed by max-pooling. The local maxima in the max-pooling layer corresponds to epochs. The Gammatone function was introduced by [17] and an auditory toolbox for gammatone filter design was implemented by [18]. The resulting representation after maxpooling is similar to a scattering representation. The scattering representation was introduced by [19, 20], and implemented in [21] for audio classification. Sparse deep scattering network (SDSN) based on scattering coefficients was used in [22] for acoustic scene analysis.



(a) (b)

Fig. 1. Gammatone wavelet $\psi^{(1)}(t)$ for $g_q(t) = t^4 e^{-54\pi t + \mathrm{j}20\pi t} u(t)$: (a) $\psi^{(1)}(t)$ and (b) magnitude spectrum at different scales (red represents the prototype) [23].

2. GAMMATONE WAVELET AND CONTINUOUS WAVELET TRANSFORM IMPLEMENTATION

2.1. Gammatone wavelet

Gammatone wavelet introduced in [23] is a complex-valued wavelet whose construction is based on the properties of the mammalian auditory system, which make it well-suited for classification of acoustic features. The gammatone function is defined in the time domain as

$$g(t) = t^{N-1} e^{-\alpha t} \cos(\omega_o t) u(t), \tag{1}$$

where α is the bandwidth parameter, ω_0 is the center frequency, $u(t)$ denotes the unit-step function, and N is the order of the wavelet, which affects the rise and decay of $g(t)$. We consider the quadrature approximation of $g_q(t)$ [23] given by

$$g_q(t) = t^{N-1} e^{-\alpha t} e^{\mathrm{j}\omega_o t} u(t). \tag{2}$$

The Fourier transform of $g_q(t)$ is

$$\hat{g}_q(\omega) = \frac{(N-1)!}{(\alpha + \mathrm{j}(\omega - \omega_0))^N}, \tag{3}$$

where $N!$ denotes the factorial of N.

The Gammatone wavelet is constructed by taking the derivative of the Gammatone function, and its Fourier transform is given by

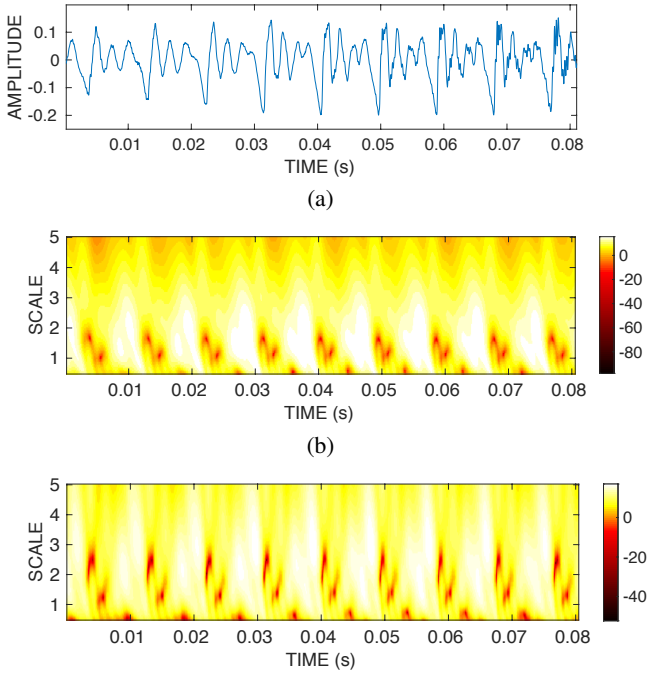
$$\widehat{\psi}^{(1)}(\omega) = \mathrm{j}\omega\, \hat{g}_q(\omega) = \frac{\mathrm{j}\omega(N-1)!}{(\alpha + \mathrm{j}(\omega - \omega_0))^N}. \tag{4}$$

In the time domain,

$$\begin{aligned} \psi^{(1)}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{ t^{N-1} e^{\beta t} u(t) \right\} \\ &= \left((N-1) t^{N-2} + \beta t^{N-1} \right) e^{\beta t} u(t), \end{aligned} \tag{5}$$

where $\beta = -\alpha + \mathrm{j}\omega_0$. The Gammatone wavelet $\psi^{(1)}(t)$ and its frequency response $\hat{\psi}^{(1)}(\omega)$ are shown in Fig. 1(a) and Fig. 1(b), respectively.

A family of Gammatone wavelets can be obtained by differentiating the Gammatone to produce wavelets up to a certain order:



(a) (b) (c)

Fig. 2. CWT analysis using wavelets $\psi^{(1)}(t)$ and $\psi^{(2)}(t)$. (a) Speech signal; (b) scalogram using $\psi^{(1)}(t)$; and (c) scalogram using $\psi^{(2)}(t)$.

$$\psi^{(n)}(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}(t^{N-1} e^{\beta t} u(t)). \tag{6}$$

Differentiating the Gammatone imparts the zero-average property while retaining finite energy. Both of these properties are crucial for wavelet analysis.

2.2. Implementation of CWT

We determine the continuous wavelet transform (CWT) using Gammatone wavelets as described in [23]. The continuous wavelet transform is an infinite collection of inner products measured between a given function $f \in \mathbf{L^2}(\mathbb{R})$ and a shifted and dilated wavelet $\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right)$, where $a \in \mathbb{R}^+$ is the scale/dilation parameter that controls the spread of the function, and $b \in \mathbb{R}$ is the translation parameter. The CWT of $f(t)$ is defined as

$$W_f(a,b) = \langle f, \psi_{a,b} \rangle = \int_{-\infty}^{+\infty} f(t) \psi^*_{a,b}(t)\, \mathrm{d}t, \tag{7}$$

where $\psi^*_{a,b}(t)$ is the complex conjugate of $\psi_{a,b}(t)$ and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbf{L^2}(\mathbb{R})$. In practice, we use the discrete-time approximation

$$W_f[a,n] = \sum_m f[m] \frac{1}{\sqrt{a}} \psi\left(\frac{m-n}{a}\right), \tag{8}$$

where $f[m]$ denotes the speech signal, $\psi[n]$ denotes the real part of the Gammatone mother wavelet, $a \in \mathbb{R}^+$ is the scale parameter and $n \in \mathbb{Z}$ is the translation parameter. The CWT analysis on speech signal and its corresponding scalograms are shown in Fig. 2.

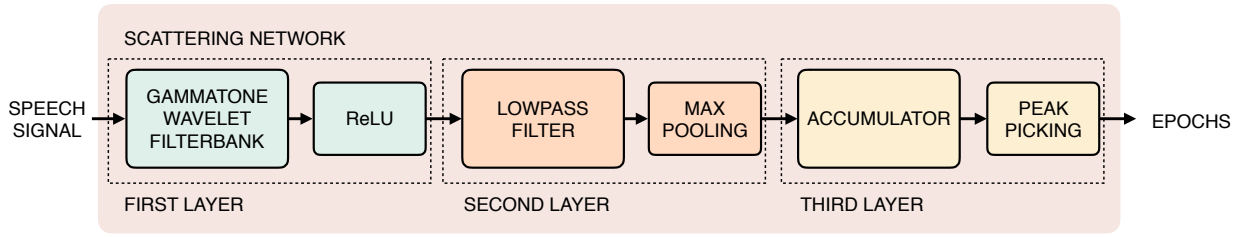


Fig. 3. Block diagram of the proposed method.

3. SCATTERING NETWORK APPROACH FOR EPOCH ESTIMATION

We propose a three-layer scattering network (SN) architecture based on a perceptually motivated Gammatone wavelet filterbank (GWFB). A schematic representation is given in Fig. 3. The time-frequency coefficients obtained by filtering the speech signal using GWFB are essentially Gammatone wavelet coefficients (GWC). The first layer consists of the Gammatone wavelet filterbank followed by a rectified linear unit (ReLU) activation function. We consider a 91-channel filterbank since the sampling frequency of the speech signal used is 32 kHz. The output of the first layer is

$$x_{\text{HR}}[a,n] = \begin{cases} W_f[a,n], & \text{if } W_f[a,n] \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The second layer implements a low-pass filtering operation followed by max-pooling. The low-pass filtered signal in each channel is given by

$$x_{\text{LP}}[a,n] = x_{\text{HR}}[a,n] * h_{\text{LP}}[n], \tag{10}$$

where $h_{\text{LP}}[n]$ is a Gaussian lowpass filter with $\sigma = a$, i.e., the spread of the Gaussian increases with scale. The max-pool operation along time is represented as follows:

$$\hat{x}[a,n] = \begin{cases} x_{\text{LP}}[a, n = l_k], & \text{if} \quad l_k = \arg\max_{n \in \mathbf{I}_k} x_{\text{LP}}[a,n], \\ 0, & \text{if} \quad n \in \mathbf{I}_k \setminus l_k, \end{cases} \tag{11}$$

where $\mathbf{I}_k = \{n : (k-1)M \leq n \leq kM\}$ denotes the indices of k^{th} max-pooling window (along the time) and M is the width of the window. We fix M to be the average pitch period of 20 ms. The final layer computes the feature waveform :

$$\tilde{x}[n] = \sum_a \hat{x}[a,n]. \tag{12}$$

The local maxima in the feature waveform corresponds to the epochs. The peak locations are computed over a window length equal to average pitch period of 20 ms. Layer-wise output of the SN-GWFB and the estimated epoch locations for a given speech signal are shown in Fig. 4.

4. EXPERIMENTAL VALIDATION AND COMPARISON

We consider the CMU-ARCTIC database [24], [25] for performance evaluation on clean and telephonic channel speech and comparisons with the existing methods. We consider two corpora from the database viz., BDL, JMK, and SLT, where BDL and JMK have recordings by a male speaker and SLT has recordings by a female speaker. Each corpus has 1132 speech recordings spoken by a single

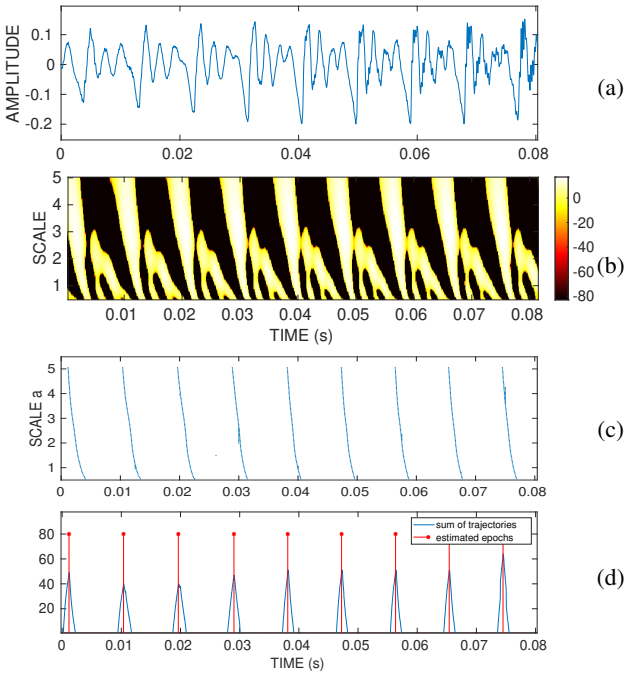


Fig. 4. Estimated epoch locations and layer-wise outputs of SN-GWFB for a given speech signal: (a) input speech signal; (b) half-rectified scattering representation; (c) output of the max-pooling layer; and (d) output after accumulation.

speaker and recorded at 32 kHz sampling rate. Each utterance in a corpus is of approximately 3 s duration. We considered 50 such utterances from each corpus for the analysis.

CMU-ARCTIC database also contains electroglottograph (EGG) recordings. The epoch reference locations are extracted from the EGG signal by finding prominent peaks in the differenced EGG signal. The epoch estimation error is computed with respect to these reference locations. Since we do not have a database that contains telephone quality speech and the corresponding EGG, we employ the technique described in [14] to simulate a telephone channel. We design a bandpass filter with passband edges at 300 Hz and 3400 Hz, and stopband edges at 20 Hz and 4000 Hz, respectively.

Objective measures viz., identification accuracy (ID), false alarm rate (FAR), miss rate (MISS), and the standard deviation (SD), identification accuracy (IDA) up to $\pm$ 0.25 ms error as described in [12–14] are used. An illustration of these measures is provided in Fig. 5. Table 1 shows a performance comparison of the proposed method with state-of-the-art epoch estimation meth-

Table 1. Comparison of the performance of the proposed SN-GWFB method with state-of-the-art epoch estimation algorithms for clean and telephone channel speech.

		Clean speech					Telephone channel speech				
Speaker (Epochs)	**Technique**	**ID %**	**MISS %**	**FAR %**	**SD ms**	**Accuracy within 0.25 ms**	**ID %**	**MISS %**	**FAR %**	**SD ms**	**Accuracy within 0.25 ms**
BDL (10856)	ZFF [12]	98.08	0.03	1.89	0.30	71.75	86.51	0.01	13.48	0.29	77.44
	SEDREAMS [13]	97.85	1.10	1.05	0.30	84.42	98.21	0.23	1.56	0.38	69.63
	SZCR [14]	98.74	0.10	1.16	0.35	83.17	97.20	0.22	2.58	0.43	84.18
	DPI [10]	95.01	0.20	0.79	0.89	**86.26**	98.53	0.22	1.25	0.33	85.42
	Proposed	**99.71**	0.06	**0.23**	0.41	81.27	**99.20**	0.33	**0.47**	0.58	**88.88**
SLT (15099)	ZFF [12]	99.85	0.03	0.12	0.18	87.32	98.77	0.05	1.18	1.43	86.70
	SEDREAMS [13]	99.78	0.07	0.15	0.28	74.03	97.83	0.74	1.43	1.61	55.65
	SZCR [14]	99.73	0.13	0.14	0.21	87.84	97.12	0.99	1.89	1.91	79.19
	DPI [10]	98.97	0.69	0.34	0.44	89.74	88.24	5.54	6.22	2.36	79.57
	Proposed	**99.90**	0.01	**0.09**	0.33	**89.98**	**99.68**	0.21	**0.11**	0.73	**89.94**
JMK (17923)	ZFF [12]	99.36	0.03	0.61	0.69	57.32	97.82	1.92	0.26	0.81	68.70
	SEDREAMS [13]	99.00	0.95	0.05	0.44	81.03	99.28	0.34	0.38	0.49	62.67
	SZCR [14]	99.29	0.38	0.33	0.95	59.14	99.29	0.38	0.33	0.95	86.50
	DPI [10]	99.45	0.16	0.39	0.44	88.53	98.08	1.26	0.66	1.36	86.57
	Proposed	**99.92**	0.05	**0.03**	0.35	**89.98**	**99.78**	0.05	**0.17**	0.51	**89.04**

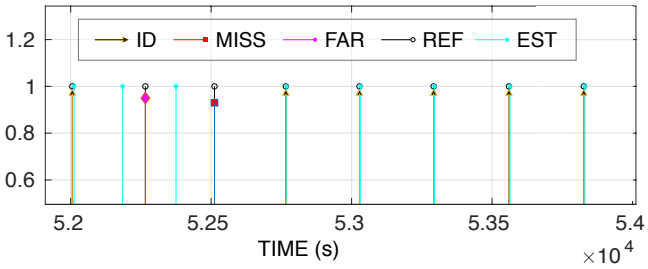


Fig. 5. Illustration of objective measures used to evaluate the performance of the proposed method.

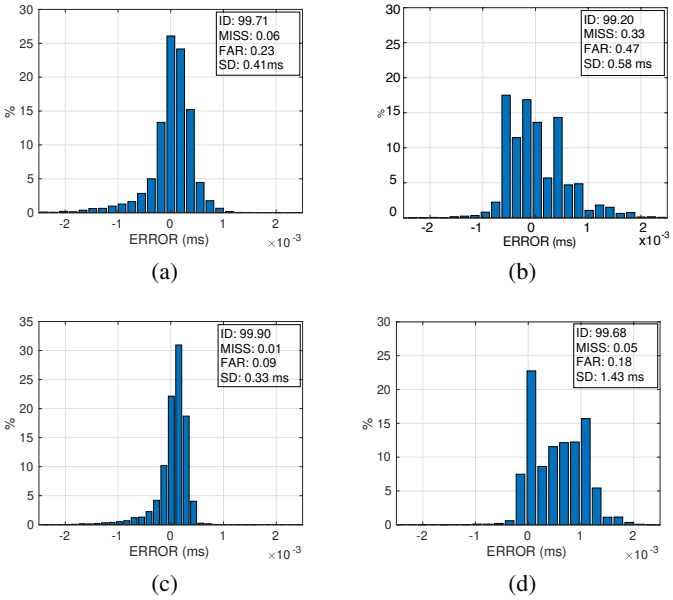


Fig. 6. Distribution of errors in the estimated epochs. (a) Clean speech (BDL); (b) telephone channel speech (BDL); (c) clean speech (SLT); and (d) telephone channel speech (SLT).

ods. Histograms of errors in the proposed method are shown in Fig. 6. The proposed method has an accuracy within 0.25 ms and is comparable with the other techniques. However, the identification rates of the proposed SN-GWFB method are higher than the other methods. There is a marginal degradation in the performance of all the methods for the simulated telephone channel speech of SLT dataset as the fundamental frequency of the female speaker is closer to the lower band-edge of the filter. However, the identification rate of the proposed method is higher even for telephone quality speech. The proposed method also has the lowest FAR for both clean and telephone quality speech.

5. CONCLUSIONS

We proposed a scattering network framework using the Gammatone wavelet for epoch estimation in a speech signal. The key feature of the proposed network is employing the perceptually motivated Gammatone wavelet filterbank in single layer. The discrete-time approximation of the continuous wavelet transform was employed in constructing the 91-channel Gammatone filterbank. The epoch locations are estimated as the peak of the accumulated local maxima of filterbank channels. The proposed method outperforms the state-of-the-art methods in term of identification accuracy and false alarm rate, for both clean and telephone quality speech.

6. REFERENCES

[1] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, pp. 1112 – 1122, 2013.

[2] S. Rudresh, A. Vasisht, K. Vijayan, and C. S. Seelamantula, "Epoch-synchronous overlap-add (ESOLA) for time-and pitch-scale modification of speech signals," *arXiv preprint arXiv:1801.06492*, 2018.

[3] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006.

[4] B. Yegnanarayana and S. V. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, no. 5, pp. 651–697, 2011.

[5] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–205, 1995.

[6] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1110–1118, 2005.

[7] A. Neocleous and P. A. Naylor, "Voice source parameters for speaker verification," *in Proceedings of 9th European Signal Processing Conference*, pp. 1–4, 1998.

[8] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[9] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[10] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013.

[11] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 6, pp. 562–570, 1975.

[12] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[13] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[14] R. R. Shenoy and C. S. Seelamantula, "Spectral zero-crossings: Localization properties and applications," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3177–3190, 2015.

[15] K. Vijayan and K. S. R. Murty, "Epoch extraction by phase modelling of speech signals," *Circuits, Systems, and Signal Processing*, vol. 35, no. 7, pp. 2584–2609, Jul 2016.

[16] C. M. Vikram and S. R. M. Prasanna, "Epoch extraction from telephone quality speech using single pole filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 624–636, 2017.

[17] P. I. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," *Proceedings of the Symposium on Hearing Theory, Institute of Perception Research, Eindhoven, Holland*, pp. 58–69, 1972.

[18] M. Slaney, "An efficient implementation of the patterson and holdworth auditory filter bank," *Apple Technical Report*, p. 35, 1993.

[19] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[20] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[21] J. Andén and S. Mallat, "Multiscale scattering for audio classification," *Proceedings of the* 12*th International Society for Music Information Retrieval Conference*, pp. 657–662, 01 2011.

[22] R. Cosentino, R. Balestriero, R. Baraniuk, and A. Patel, "Overcomplete frame thresholding for acoustic scene analysis," *arXiv preprint arXiv:1712.09117*, 2017.

[23] A. Venkitaraman, A. Adiga, and C. S. Seelamantula, "Auditory-motivated gammatone wavelet transform," *Signal Processing*, vol. 94, pp. 608 – 619, 2014.

[24] J. Kominek and A. Black, "The CMU Arctic speech databases," *5th ISCA Speech Synthesis Workshop*, pp. 223–224, 2004.

[25] *CMU-ARCTIC Speech Synthesis Databases*, `http://festvox.org/cmu_arctic/index.html`.