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]
css@iisc.ac.in

[1]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

[2]Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, USA

May 4-8, 2020

45th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing

Contents

Introduction - Speech production

I The figure shows human speech production system.

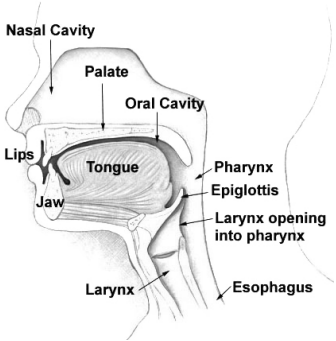


Figure 1: The human speech production system.

Picture credits: *Wikipedia*

Introduction - Epochs

Extracting epoch locations in a speech signal plays an important role in many applications.

- Epochs are glottal closure instants.
- It is used in describing the voice characteristics[1].
- Rao et al.[2] and Rudresh et al.[3] used epochs as pitch markers in time/pitch scaling.
- Epochs serve as pitch markers in applications such as voice conversion and text-to-speech synthesis.
- Yegnanarayana et al.[4] used epoch locations to estimate the time-delay between speech signals.

[1] Teixeira et al., *Procedia Technology*. 2013.
[2] Rao and Yegnanarayana, *IEEE TASLP,*. 2006.
[3] Rudresh et al., *arXiv preprint arXiv:1801.06492*. 2018.
[4] Yegnanarayana et al., *IEEE TASLP*. 2005.

Literature Review

- Murthy and Yagnanarayana[5] introduced a technique called zero-frequency resonator (ZFR) for epoch estimation.
- Drugman et al.[6] introduced an algorithm that uses residual excitation and mean-based signal (SEDREAMS).
- Both ZFR and SEDREAMS require prior knowledge of the pitch period for window selection and are robust to noise.

[5] Murty and Yegnanarayana, *IEEE TASLP*. 2008.
[6] Drugman et al., *IEEE TASLP*. 2012.

Literature

- Prathosh et al.[7] introduced a dynamic plosion index (DPI) to determine epcoh locations.
- Shenoy and Seelamantula[8] used spectral zero-crossing rate (SZCR) to determine epochs.
- Both the technique show robust performance even with telephone channel speech compared to ZFR and SDREAMS.

[7] Prathosh et al., *IEEE TASLP*. 2013.
[8] Shenoy and Seelamantula, *IEEE Transactions on Signal Processing*. 2015.

This Work

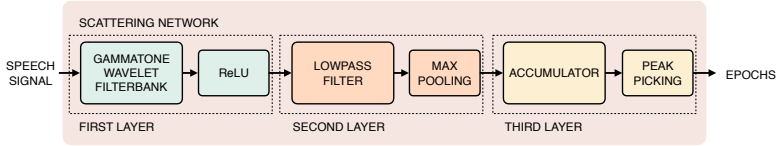


Figure 2: Block diagram of the proposed method.

- The figure shows overall block diagram of the proposed method in a scattering network[9] framework.
- In this work, 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 lowpass filter followed by max-pooling.
- The local maxima after max-pooling correspond to epochs.

[9] Bruna and Mallat, *IEEE TPAMI*. 2013.

Gammatone Wavelet and Continuous Wavelet Transform Implementation

- Johannesma et al.[10] introduced the gammatone function, and it 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.

- We consider the quadrature approximation $g_q(t)$

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

[10] Johannesma, *Proceedings of the Symposium on Hearing Theory*. 1972.

Gammatone Wavelet and Continuous Wavelet Transform Implementation

I The Gammatone wavelet[11] is constructed by taking the derivative of the Gammatone function, and its Fourier transform is given by

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

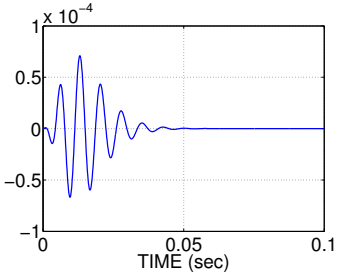
In the time domain,

$$\begin{aligned} \psi^{(1)}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{t^{N-1}e^{\beta t}u(t)\right\} \\ &= \left((N-1)t^{N-2} + \beta t^{N-1}\right)e^{\beta t}u(t), \end{aligned} \tag{4}$$

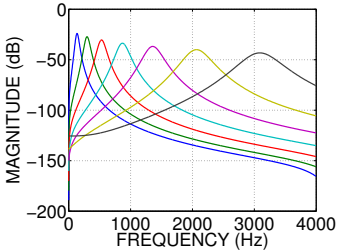
where $\beta = -\alpha + \mathrm{j}\omega_0$.

[11] Venkitaraman et al., *Signal Processing*. 2014.

Gammatone wavelet and Continuous Wavelet Transform implementation



(a)



(b)

Figure 3: Gammatone wavelet $\psi^{(1)}(t)$ for $g_q(t) = t^4 e^{-54\pi t + \mathrm{j}20\pi t} u(t)$

Gammatone Wavelet and Continuous Wavelet Transform Implementation

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

$$\psi^{(n)}(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}(t^{N-1}e^{\beta t}u(t)). \tag{5}$$

Gammatone Wavelet and Continuous Wavelet Transform Implementation

- The continuous wavelet transform (CWT) of $f(t)$ is defined as

$$W_f(a,b) = \langle f, \psi_{a,b} \rangle = \int_{-\infty}^{+\infty} f(t)\psi_{a,b}(t)\,\mathrm{d}t, \tag{6}$$

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

$$W_f[a,n] = \sum_m f[m] \frac{1}{\sqrt{a}} \psi\left(\frac{m-n}{a}\right), \tag{7}$$

where $f[m]$ denotes the speech signal, $\psi[n]$ denotes the real part of the Gammatone mother wavelet, $a \in \mathbb{R}^+$ and $n \in \mathbb{Z}$.

Gammatone Wavelet and Continuous Wavelet Transform Implementation

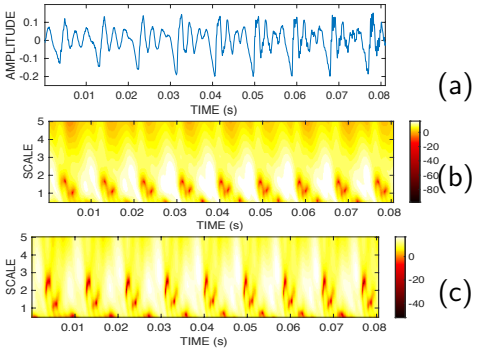


Figure 4: CWT analysis using wavelets $\psi^{(1)}(t)$ and $\psi^{(2)}(t)$.

Proposed Method

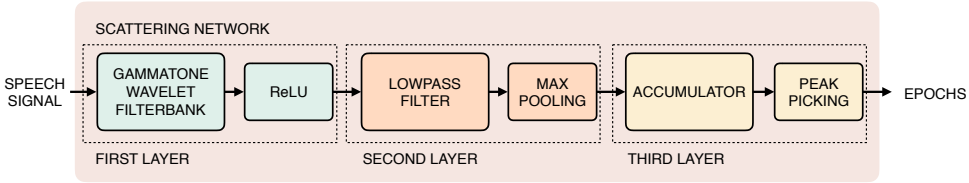


Figure 5: Block diagram of the proposed method.

- The output of the first layer is

$$x_{\mathrm{HR}}[a,n] = \begin{cases} W_f[a,n], & \text{if } W_f[a,n] \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Proposed Method

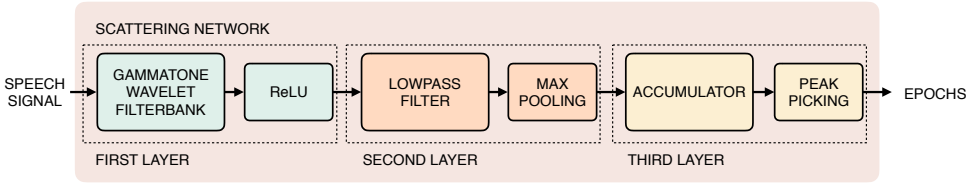


Figure 6: Block diagram of the proposed method.

- The low-pass filtered signal in each channel is given by

$$x_{\mathrm{LP}}[a,n] = x_{\mathrm{HR}}[a,n] \quad h_{\mathrm{LP}}[n], \tag{9}$$

where $h_{\mathrm{LP}}[n]$ is a Gaussian lowpass filter with $\sigma = a$.

Proposed Method

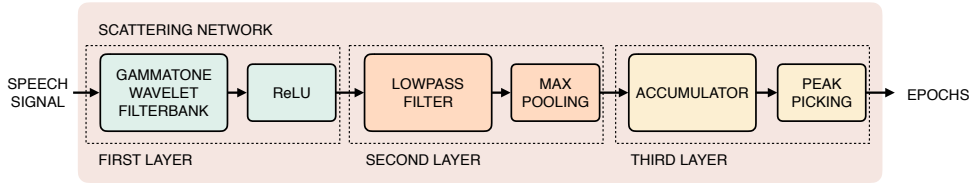


Figure 7: Block diagram of the proposed method.

I The max-pool operation along time is represented as follows:

$$\hat{x}[a,n] = \begin{cases} x_{\mathrm{LP}}[a, n = l_k], & \text{if} \quad l_k = \arg\max_{n \quad \mathbf{l}_k} x_{\mathrm{LP}}[a,n], \\ 0, & \text{if} \quad n \quad \mathbf{l}_k \setminus l_k, \end{cases} \tag{10}$$

where $\mathbf{l}_k = \{n : (k-1)M \leq n \leq kM\}$ M is the width of the window is fixed to the average pitch period of 20 ms.

Proposed Method

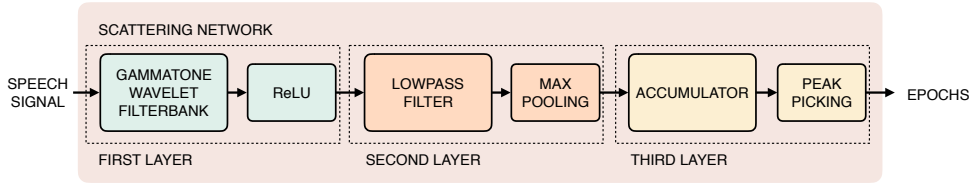


Figure 8: Block diagram of the proposed method.

- The third layer computes the pitch-specific feature waveform :

$$\tilde{x}[n] = \sum_{a} \hat{x}[a, n]\,. \tag{11}$$

Proposed Method

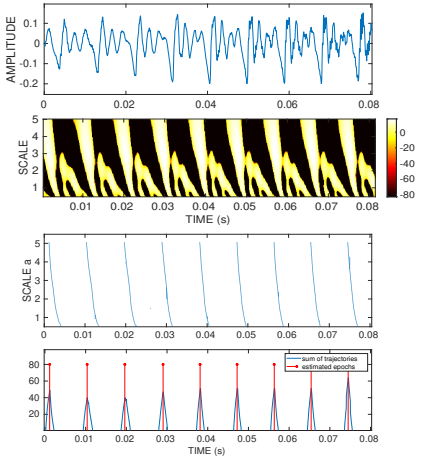


Figure 9: Layer-wise outputs of SN-GWFB for a given speech signal.

Experimental Validation

- We consider the CMU-ARCTIC database[12] for performance evaluation on clean and telephonic channel speech
- We consider three databases, viz., BDL, JMK, and SLT.
- Each corpus has 1132 speech recordings spoken by a single speaker and recorded at 32 kHz sampling rate.
- We considered 50 utterances from each corpus for the analysis.
- Corresponding telephonic quality speech is simulated by designing the bandpass filter with passband edge at 300 Hz and 3400 Hz and stopband edges at 20 Hz and 4000 Hz, repectively.

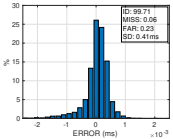
[12] Kominek and Black, *5th ISCA Speech Synthesis Workshop*, 2004.

Results

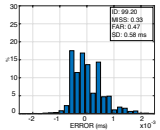
Table 1: Performance Comparison.

Speaker (Epochs)	Technique	Clean speech ID %	Clean speech MISS %	Clean speech FAR %	Clean speech SD ms	Clean speech Accuracy within 0.25 ms	Telephone channel speech ID %	Telephone channel speech MISS %	Telephone channel speech FAR %	Telephone channel speech SD ms	Telephone channel speech Accuracy within 0.25 ms
BDL (10856)	ZFF	98.08	0.03	1.89	0.30	71.75	86.51	0.01	13.48	0.29	77.44
	SEDREAMS	97.85	1.10	1.05	0.30	84.42	98.21	0.23	1.56	0.38	69.63
	SZCR	98.74	0.10	1.16	0.35	83.17	97.20	0.22	2.58	0.43	84.18
	DPI	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	99.85	0.03	0.12	0.18	87.32	98.77	0.05	1.18	1.43	86.70
	SEDREAMS	99.78	0.07	0.15	0.28	74.03	97.83	0.74	1.43	1.61	55.65
	SZCR	99.73	0.13	0.14	0.21	87.84	97.12	0.99	1.89	1.91	79.19
	DPI	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	99.36	0.03	0.61	0.69	57.32	97.82	1.92	0.26	0.81	68.70
	SEDREAMS	99.00	0.95	0.05	0.44	81.03	99.28	0.34	0.38	0.49	62.67
	SZCR	99.29	0.38	0.33	0.95	59.14	99.29	0.38	0.33	0.95	86.50
	DPI	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**

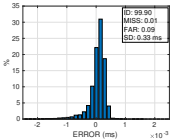
Results



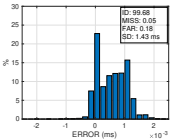
(a)



(b)



(c)



(d)

Figure 10: Distribution of errors in the estimated epochs.

Conclusion

- We proposed a scattering network framework using the Gammatone wavelet for epoch estimation in a speech signal.
- 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 terms of identification accuracy and false alarm rate, for both clean and telephone quality speech.

Acknowledgement

Funding Agency

- "Development of text-to-speech synthesis systems for Indian languages - Phase II", funded by the Department of Information Technology, Government of India.

Thank you