INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Epoch-Based Spectrum Estimation for Speech

Jon Gudnason[1], *Guolin Fang*[1], *Mike Brookes*[2]

[1]Language and Voice Lab, Reykjavik University, Iceland
[2]Communications and Signal Processing Group, Imperial College London, UK
jg@ru.is, guolin19@ru.is, mike.brookes@imperial.ac.uk

Abstract

An implicit assumption when using the discrete Fourier transform for spectrum estimation is that the time signal is periodic. This assumption clashes with the quasi-periodicity of voiced speech when the traditional short-time Fourier transform (STFT) is applied to it. This causes distortion and leads to a performance handicap in downstream processing. This work proposes a remedy to this by using epochs in the signal to determine better frame boundaries for the Fourier transform. The epochs are the estimated glottal closure instants in voiced speech and significant peaks in the unvoiced speech signal. The resulting coefficients are compared to the traditional STFT coefficients using copy-synthesis. An improvement of 15 dB signal-to-noise ratio and a PESQ score of 2.5 to 3.5 is achieved for copy-synthesis using 20 mel-filters. The results demonstrate that there is a great potential in improving down stream speech processing applications using this approach to spectrum estimation.

Index Terms: speech signal processing, Fourier analysis, copy-synthesis, vocoding.

1. Introduction

Spectrum estimation is fundamental to speech processing and has been described in textbooks for decades [1, 2]. These textbooks typically describe the short term Fourier analysis of speech in terms of fixed time frames formed as a product of the (long) speech signal and a (typically rectangular) window function (see, for example [1, Ch. 6]) and [2, Ch. 4]). Common speech signal processing toolboxes such as Voicebox and Librosa [3, 4] implement the short-time Fourier transform (STFT) and mel-frequency cepstrum coefficient (MFCC) extraction according to the fixed frame approach where the frame size is kept constant while the frame is shifted in equal hops along the signal. This method has certain benefits as it is simple and easy to implement with no prior signal processing that is specific to the frame boundary determination required.

A problem with this approach, however, is that it ignores the implicit periodicity assumption of the discrete Fourier transform (DFT). An analysis frame that has a fixed duration will normally contain a non-integer number of fundamental periods in the voiced speech signal. As an example, an analysis frame of 25 ms containing voiced speech with a period of 10 ms will contain 2.5 periods. This fractional period will distort the DFT coefficients as they will sample the continuous discrete-time Fourier transform (DTFT) of an infinitely long signal that repeats with a period of 25 ms. The DFT phase will also be hard to interpret since the analysis frame has an arbitrary and variable time-shift with respect to the signal. This has been a serious obstacle in using phase information in speech processing [5].

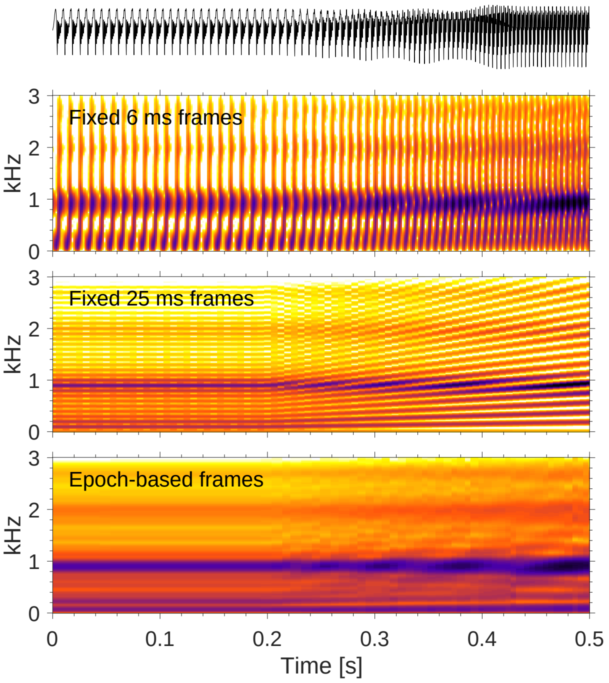

Figure 1: *The upper panel shows a* 0.5 s *sample of synthetic speech whose pitch equals* 100 Hz *for the first* 0.2 s *before rising linearly to* 190 Hz. *The next two panels show wideband and narrowband spectrograms which use fixed analysis windows of* 6 ms *and* 25 ms *respectively. The lower panel shows a spectrogram that uses an adaptive analysis window whose length equals the estimated pitch period.*

This problem is illustrated in Fig. 1 with spectrograms of a synthetic speech segment. The segment is generated using Liljencrants-Fant source parameters [6, 3] and a time-invariant auto-regressive filter derived from a real speech segment. The pitch of the source is constant at 100 Hz for the first 0.2 s and then increases linearly to 190 Hz. The first spectrogram, which uses 6 ms fixed analysis frames, varies with time since the analysis frames are shorter than the pitch periods. In an attempt to remedy this, the analysis frame is increased to 25 ms in the second spectrogram ensuring that more than one pitch period is included in each frame. The vertical stripes visible in this spectrogram show, however, that the Fourier coefficients change from one frame to the next even during first 0.2 s when the signal is

The code is available at:
https://github.com/cadia-lvl/ebs/tree/interspeech2023

10.21437/Interspeech.2023-407

exactly periodic.

This work proposes an alternative way of determining the analysis frame boundaries for the STFT. For voiced speech the analysis frame captures a single fundamental pitch period in a speech signal by calculating the frame boundaries based on glottal closure instants (GCIs) [7]. For clean speech, GCIs can be accurately identified from, for example, discontinuities in derived signals such as the mean-based signal or the multi-scalar product [8, 9]. For unvoiced speech, the analysis frames are calculated similarly, albeit without the periodicity interpretation in the voiced speech. These time frames are however synchronised with respect to epochs that represent an underlying time structure in the unvoiced speech. Fig. 1 demonstrates the effectiveness of this approach where the third spectrogram demonstrates how little the discrete Fourier coefficients vary over the synthesized segment. Here, the Fourier coefficients and the frame duration vary consistently from one frame to the next and encode all the information needed for perfect reconstruction.

A brief overview of related work is given below before the proposed approach is described in Sec. 2. Section 3 describes a copy-synthesis method that is used to compare the proposed approached with traditional STFT of speech and the results of these experiments are presented and discussed in Sec. 4. The implications of the approach and the results are presented and the work concluded in Sec. 5.

1.1. Related work

Detection of GCIs has been of great interest to researchers [10, 7]. Much focus has been on developing accurate detectors evaluated on ground-truth data derived from the electroglottogram [9, 11] but they have also been used for glottal inverse filtering [12, 13, 14], voice quality assessment [15], cognitive workload monitoring [16], artificial bandwidth extension [17] and speaker identification [18]. Most of the applications revolve around improving the computation of linear prediction coefficients by achieving closed-phase analysis or weighted covariance estimation [19] by knowing the location of the closures. This line of development leads to an improved modern spectrum estimation, but GCIs have to the best of our knowledge, never been directly used to estimate Fourier coefficients.

The importance of pitch-synchronous speech processing has been known for a long time within the speech synthesis community. Pitch synchronous overlap add (PSOLA) was used to improve concatenate diphone [20] and unit selection [21] synthesis with the latter displaying an early GCI detection version using the group-delay [22]. The importance of pitch synchrony was also demonstrated [23] with the importance of avoiding a non-integer number of pitch periods in the analysis windows when estimating linear prediction parameters. This method was developed into the STRAIGHT [24] vocoder and widely used in statistical parametric speech synthesis. Other approaches to vocoding that also recognise the importance of estimating parameters over an integer number of pitch periods include the Harmonic plus Noise Model (HNP) vocoder [25], the Vocaine vocoder [26] and Magnitude and Phase Spectra (MagPhase) vocoder [27].

In recent years, the main attention of TTS research has coalesced around neural-vocoding [28, 29] while spectral based vocoding seems to have been abandoned. The early version of Tacotron [30], for example, used mel-filter energies for vocoding and used the Griffin-Lim algorithm for resynthesis [31]. With better spectrograms, there might not have been the need to move away from this kind of vocoding.

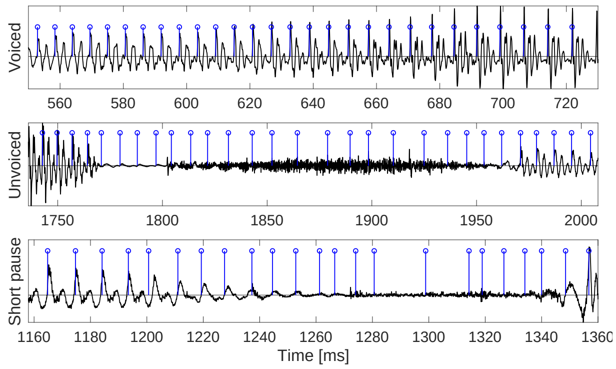

Figure 2: *The epochs extracted from three different speech segments. The epochs are the GCIs during voiced segments whereas they correspond to discontinuities in the speech during unvoiced and short-pause segments.*

2. Epoch-Based Spectrum

2.1. Epochs

Epochs are glottal closure instants (GCIs) in voiced speech and time instants that correspond to discontinuities in unvoiced speech. They are extracted using the Yet Another GCI Algorithm (YAGA) [9]. The algorithm also returns epochs during non voiced segments (i.e. unvoiced and silence). These epochs may have some significance as they represent interesting time instants that correspond to discontinuities in the inverse-filtered speech signal. Fig. 2 shows epochs extracted during a voiced, an unvoiced, and a short-pause segment of speech. Pseudo-epochs are introduced if the gap between consecutive extracted epochs exceeds $25\,\mathrm{ms}$; this corresponds to a $40\,\mathrm{Hz}$ pitch period which is therefore the lowest pitch that the algorithm recognises. A gap larger than $25\,\mathrm{ms}$ arises where no interesting time instants are detected; in this case, enough evenly-spaced pseudo-epochs are introduced into the gap to reduce the time interval between consecutive epochs to below $20\,\mathrm{ms}$.

Figure 2 shows the three types of epoch that are used in the work. The first panel shows the epochs that correspond to GCIs in voiced speech. The second panel shows the epochs extracted during a fricative where pronounced discontinuities are picked out by the algorithm. The third panel shows how pseudo-epochs have been inserted to fill in a gap where no significant discontinuity has been detected. The epochs are labelled as $i \in \{1, 2, \dots\}$ using their order of occurrence, irrespective of their types (voiced, unvoiced, pseudo/silence). The sample number of the i-th epoch is n_i so a speech signal, $s(n)$, sampled at a frequency $f_s = 16$ kHz might have its $i = 84$-th epoch occurring at $n_{84} = 8864$ when $552.9\,\mathrm{ms}$ have lapsed of the signal (this example corresponds to the first GCI in the first panel of Fig. 2).

2.2. Frame boundaries

Analysis frames which have a meaningful time location are determined from the epochs in the speech signal. The aim of the encoding step is to create an invertible projection so that the signal can be decoded with perfect reconstruction. We therefore use non-overlapping analysis frames with the epochs placed at a predetermined position within the window. The first sample of the frame is therefore,

$$o_i = n_i - \lfloor pN_i \rfloor \tag{1}$$

where $N_i = n_i - n_{i-1}$ is the number of samples between the epoch and its previous epoch and is the pitch period in samples during voiced speech. The final sample in the i-th frame is simply $o_{i+1} - 1$ and the design parameter $p \in [0, 1[$ allows us to control the time shift of the frame with respect to the epochs. If $p = 0$, the epochs themselves become the frame boundaries whereas if $p = 0.5$ the epochs will be close to the middle of the frames (so long as $N_i \approx N_{i+1}$). For this work the value of $p = 0.3$ is set to avoid having the frame boundaries close to the epochs and to place the frame boundaries in the open phase of the glottal cycle which tends to produce low values in the speech signal. The first panel of Fig. 3 shows a short segment of voiced speech. The positive stems show the frame boundaries chosen with this method resulting in frames that are synchronised with the pitch period. A single fixed frame of 25 ms is also shown in this panel as vertical broken lines. The duration of the frame is equivalent to three pitch cycles plus approximately 2.5 ms that will distort the Fourier coefficients as described in Sec. 1.

2.3. Adjusted frame boundaries

The DFT of a finite length frame is equivalent to the DTFT of the infinite periodic signal obtained by concatenating multiple copies of the frame. Any discontinuities in the value or derivatives of this signal at the frame boundaries will introduce artefacts into the spectrum. We therefore adjust the frame boundaries by up to ±0.625 ms in order to minimize these artefacts. Dynamic programming is used to determine the adjustment to each frame boundary that minimizes the mean of the squared difference between the final samples of consecutive frames. The effect of this adjustment is illustrated in the first panel of Fig. 3 where the frame boundaries before and after adjustment are shown by the positive and negative stems respectively. The adjusted frame boundaries are denoted as $\tilde{o}_i$ and the corresponding frame duration as $\tilde{N}_i$ and can be used instead of o_i and N_i to calculate the Fourier coefficients. Although these adjustments are small, they result in a significant performance improvement to the copy-synthesis algorithm of Section 3.

2.4. Fourier coefficients

The i^{th} frame now contains N_i samples (or $\tilde{N}_i$ in the adjusted case). An N_i point discrete Fourier transform would produce coefficients at integer multiples of the fundamental frequency $f_0 = f_s/N_i$. This is the rate at which the vibration of the glottal folds samples the frequency response of the combined vocal tract and voice source system and represents a natural limit to frequency resolution in voiced speech. This limit is side-stepped by padding the frame before taking the Fourier transform to produce interpolation in the frequency domain. The time signal is padded with the constant $\alpha_i = \frac{1}{2}(s(o_i) - s(o_{i+1} - 1))$ to minimize the discontinuity of the frame boundaries,

$$s_i(m) = \begin{cases} s(o_i + m) & m = 0, 1, \dots, N_i - 1, \\ \alpha_i & m = N_i, \dots, K - 1. \end{cases} \tag{2}$$

and the resulting Fourier coefficients are

$$c_{i,k} = \mathcal{F}\{s_i(m)\} = \sum_{m=0}^{K-1} s_i(m) e^{-j2\pi km/K}, \tag{3}$$

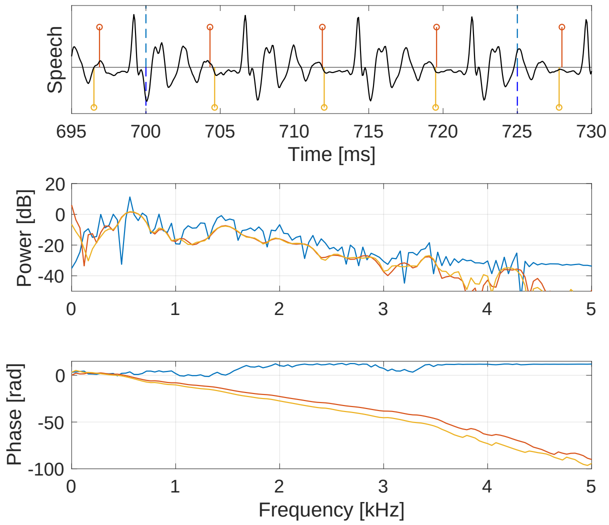

Figure 3: *One frame of speech analysed. The first panel shows one fixed-frame of speech demarcated with blue broken vertical lines, frame boundaries using the epoch-based method (red positive stems) and adjusted epoch-based method (yellow negative stems). The lower two panels show the power spectrum and the unwrapped phase of the three resulting Fourier analyses.*

for $k \in \{0, \dots, K - 1\}$. The longest period between epochs was chosen to correspond to 40 Hz (so $K = 400$ samples for f_s =16 kHz) giving an apparent frequency resolution of f_s/K (40 Hz) but the actual frequency resolution is f_s/N_i (e.g. if $N_i = 160$ the frequency resolution is f_0 =100 Hz).

Figure 3 shows the power spectrum and the phase based on fixed-frame (blue), epoch-based (red, with o_i and N_i) and adjusted epoch-based frames (yellow, with $\tilde{o}_i$ and $\tilde{N}_i$). The figure shows that the epoch-based Fourier coefficients produce a smoother spectrum since the underlying frequency resolution is coarser, but it also avoids the distortion due to the fractional glottal period that is included in the fixed-frame analysis. The phase of the epoch-based Fourier coefficients unwraps seamlessly whereas the phase of the fixed-frame analysis does not. The resulting spectrum for adjusted epochs is also shown for the synthetic segment in Fig. 1. The last panel demonstrates well how the proposed method gives consistent Fourier coefficients in the steady-state portion of the utterance, unlike either the wide-band or the narrow-band fixed frame methods.

3. Copy-Synthesis Methodology

In Sec. 1 it was argued that an epoch-based spectrum estimation is preferable to a fixed-frame approach and it was demonstrated in Sec. 2 that this is the case. Applying the inverse DFT to either to the epoch-based or the fixed-frame spectra will give a perfect reconstruction of the time signal. It is, however, a major objective of many down-stream speech processing algorithms to parameterise the speech signal, so that it can be, for example, recognised in an ASR system or regressed on to in a TTS-vocoder setting. It is beyond the scope of this paper to examine the full effect of the proposed method on such systems, but in order to get an experimental validation of the approach, a copy-synthesis system is set up and tested.

3.1. Mel filter copy-synthesis

A mel-filter bank is applied to the magnitude spectrum in order to parameterise it. The phase is kept the same for the purpose of the experiments avoiding the need to rebuild the phase using the Griffin-Lim algorithm [31]. This favours the fixed-frame method, but allows for a targeted assessment of the magnitude spectrum. The mel-filter energies are calculated for the ith frame using the triangular Mel-filters $\mathcal{M}_j(k)$ [32],

$$Y_{i,j} = \sum_{k=0}^{K-1} |c_{i,k}| \mathcal{M}_j(k). \quad j = 1, \ldots, M \tag{4}$$

The number of filters, M, is varied in the experiments but the typical choice in feature extraction is between 20 and 26 depending on the sampling frequency whereas a choice of 80 is common in TTS where better frequency resolution is needed.

The magnitude spectra $|\hat{c}_{i,k}|$ is rebuilt by interpolating the Mel-filter energies and the Fourier coefficients are then determined using the original phase,

$$\hat{c}_{i,k} = |\hat{c}_{i,k}| e^{\angle c_{i,k}}. \tag{5}$$

The synthesised time signal $\hat{s}(n)$ is then produced by concatenating the first N_i samples of the inverse DFT for each frame, i.e. $\hat{s}_i(m) = \mathcal{F}^{-1}\{\hat{c}_{i,k}\}$.

3.2. Performance assessment and data set

The signal-to-noise ratio (SNR) is evaluated for every speech utterance as

$$\text{SNR}_{db} = 10 \log_{10}(||s(n)||^2 / ||e(n)||^2) \tag{6}$$

where the noise is the error signal $e(n) = \hat{s}(n) - s(n)$ between the synthesised and the original signals. The study also includes the perceptual evaluation of speech quality (PESQ)[1] [33] which provides a comparison between $\hat{s}(n)$ and $s(n)$ that predicts subjective mean-opinion scores. The score ranges between 1 (bad) and 4.5 (no distortion). The copy-synthesis scheme is applied to the entire TIMIT dataset [34] and, since this is not a supervised learning set-up, the utterances from the training and test parts are treated the same.

4. Results and Discussion

Figure 4 shows the Mel-filter copy-synthesis SNR and PESQ scores averaged over all the utterances in the TIMIT database using fixed 25 ms frames (F, blue), epoch- (E, red) and adjusted epoch frames (A, yellow) for Mel-filter numbers ranging from 2 to 140. The error bars show the standard deviation of the SNR and PESQ scores over the utterances. In most speech processing it is desirable to have as few parameters (so in this case, Mel-filters) as possible. It is, for example, better to have as few vocoding parameters as possible when designing a TTS system since that means that the regressor has fewer outputs resulting in fewer model parameters (e.g. neural-network weights) to be trained. Similarly, shorter feature vectors for ASR systems result in a smaller input layer with fewer model parameters to train. The model design, therefore, presents the familiar trade-off between compactness and accuracy. The proposed approach is able to maintain accuracy with more compact representation (i.e. fewer Mel-filters), apparent in the difference between the yellow (A) and blue (F) curves of Fig. 4. The adjusted-epoch approach achieves superior performance for $M < 80$ Mel-filters but the fixed-frame approach does better as M is increased beyond that. The superior performance for lower M supports the main argument of the paper, in that epoch-based spectrum estimation is preferable to fixed-frame analysis. The performance ceiling that the epoch-based approaches display needs to be investigated further. The higher frequency resolution of the fixed-frame approach may explain its raised performance as M is increased. The difference between the red (E) and yellow (A) curves shows how important it is to adjust the frame boundaries in the proposed method so that the rectangular-windowing effect of the padding is reduced. The use of padding was a result of a difficult design choice as higher (apparent) frequency resolution was sought with the resulting interpolation in the frequency domain. This design choice will be further investigated in future work, but it is nevertheless interesting to see how small adjustments in the frame boundaries can achieve a big performance gain for the copy-synthesis.

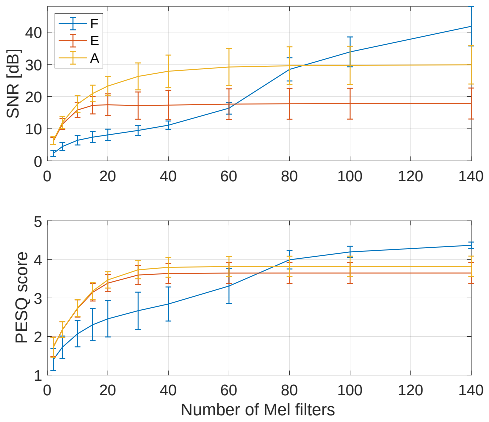

Figure 4: *The copy-synthesis SNR for fixed frame- (F), epoch- (E) and adjusted epoch (A) approaches to spectrum estimation as a function of the number of Mel-filters used in parameterisation.*

5. Conclusions

The work shows the importance of using glottal synchronous frame boundaries when calculating DFT coefficients for processing speech. The paper presents three arguments for this: (1) The mathematical argument, that it is not a good idea to calculate DFT coefficients over a fractional number of periods, (2) the demonstration using a synthetic speech segment (see. Fig. 1) and (3) the Mel-filter based copy-synthesis experiments where the proposed approach was compared to the conventional fixed-frame approach to calculating the STFT. The paper does not propose a fully fledged vocoder that is ready for use in down-stream applications. The copy-synthesis does not, for example, handle the angle of the Fourier coefficients, but the third panel of Fig. 3 shows that the proposed approach has a better chance of doing so than the conventional STFT. The method also gives an explicit description of the fundamental frequency allowing any vocoder built on this approach to model prosody separately from the short-term spectral content of the speech. Fixed-frame spectrum estimation does not achieve that.

[1] https://www.itu.int/rec/T-REC-P.862-200102-I/en

6. References

[1] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, NJ, 1979.

[2] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-time processing of speech signals*. Institute of Electrical and Electronics Engineers, 2000.

[3] M. Brookes, "Voicebox: Speech processing toolbox for matlab. world wide web," 2000.

[4] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[5] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital signal processing*, vol. 17, no. 3, pp. 578–616, 2007.

[6] G. Fant, J. Liljencrants, Q.-g. Lin *et al.*, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[7] 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.

[8] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech 2009*, 2009, pp. 2891–2894.

[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] 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.

[11] J. Matoušek and D. Tihelka, "A comparison of convolutional neural networks for glottal closure instant detection from raw speech," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6938–6942.

[12] J. Gudnason, M. R. Thomas, D. P. Ellis, and P. A. Naylor, "Data-driven voice source waveform analysis and synthesis," *Speech Communication*, vol. 54, no. 2, pp. 199–211, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639311001191

[13] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 596–607, 2013.

[14] Y.-R. Chien, D. D. Mehta, J. Gudnason, M. Zañartu, and T. F. Quatieri, "Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1718–1730, 2017.

[15] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and nonmodal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2281–2291, 2017.

[16] M. Meier, M. Borsky, E. H. Magnusdottir, K. R. Johannsdottir, and J. Gudnason, "Vocal tract and voice source features for monitoring cognitive workload," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2016, pp. 000 097–000 102.

[17] M. R. P. Thomas, J. Gudnason, P. A. Naylor, B. Geiser, and P. Vary, "Voice source estimation for artificial bandwidth extension of telephone speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4794–4797.

[18] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4821–4824.

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

[20] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[21] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 232–239, 2001.

[22] M. Brookes, P. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 456–466, 2006.

[23] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[24] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[25] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[26] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4230–4234.

[27] F. Espic, C. Valentini-Botinhao, and S. King, "Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis." in *INTERSPEECH*, 2017, pp. 1383–1387.

[28] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[29] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5679–5683.

[30] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[31] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[32] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.