INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Spectro-Temporal Demodulation Technique for Pitch Estimation

Jitendra Kumar Dhiman, Nagaraj Adiga, and Chandra Sekhar Seelamantula

Department of Electrical Engineering, Indian Institute of Science, Bangalore - 560 012, India
jkdiith@gmail.com, nvadigauvce@gmail.com, chandra.sekhar@ieee.org

Abstract

We consider a two-dimensional demodulation framework for spectro-temporal analysis of the speech signal. We construct narrowband (NB) speech spectrograms, and demodulate them using the Riesz transform, which is a two-dimensional extension of the Hilbert transform. The demodulation results in time-frequency envelope (amplitude modulation or AM) and time-frequency carrier (frequency modulation or FM). The AM corresponds to the vocal tract and is referred to as the vocal tract spectrogram. The FM corresponds to the underlying excitation and is referred to as the carrier spectrogram. The carrier spectrogram exhibits a high degree of time-frequency consistency for voiced sounds. For unvoiced sounds, such a structure is lacking. In addition, the carrier spectrogram reflects the fundamental frequency (F0) variation of the speech signal. We develop a technique to determine the F0 from the carrier spectrogram. The time-frequency consistency is used to determine which time-frequency regions correspond to voiced segments. Comparisons with the state-of-the-art F0 estimation algorithms show that the proposed F0 estimator has high accuracy for telephone channel speech and is robust to noise.

Index Terms: spectro-temporal demodulation, time-frequency consistency, fundamental frequency (F0) estimation, glottal excitation, Riesz transform.

1. Introduction

Pitch is the perceptual correlate of the fundamental frequency (F0) of speech signal [1], as perceived by humans. In speech signal processing, the definition of pitch is often motivated from the speech production perspective where the voiced speech sounds are produced by exciting the vocal-tract with the airflow pressure from the lungs and simultaneous quasi-periodic vibration of vocal folds. Pitch is related to the rate of vibration of vocal folds over a short segment of voiced speech sound [2]. Over the past few decades in speech processing research, accurate pitch estimation and tracking have been of considerable interest for many speech applications such as speech synthesis [3], prosody modifications [4], and speaker/speech recognition [2].

1.1. Prior art

There are many techniques in the literature to estimate pitch based on analysis carried out in the time domain [5], spectral domain [6, 7], or spectro-temporal domain [8]. Techniques based on the properties in the time domain operate on the speech signal [5]. To estimate the pitch, the autocorrelation of the periodic segment is computed, which shows a strong peak at the pitch period and the reciprocal of it gives the pitch frequency. Frequency domain techniques estimate pitch by utilizing the frequency spectrum of the signal which contains pitch frequency and its harmonics. In this class, the cepstrum method

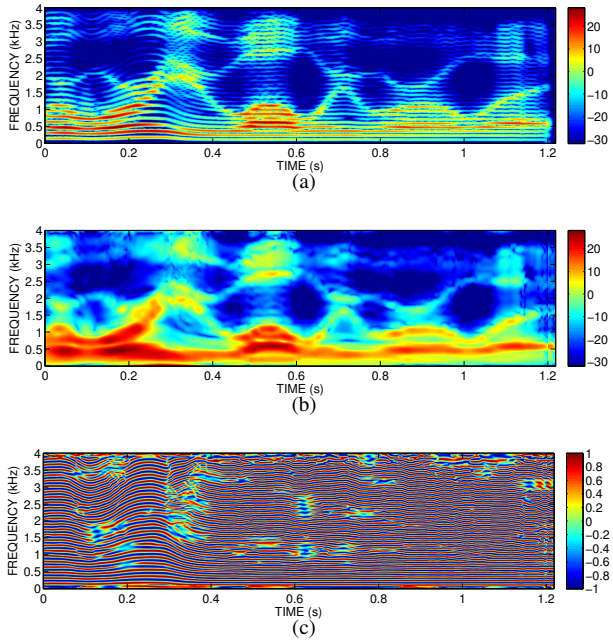

Figure 1: (a) *Narrowband spectrogram (NB) for a voiced sound (dB), (b) vocal-tract spectrogram (dB), and (c) carrier spectrogram. The utterance is "Where were you while we were away?"*

is popular [6] — it separates the fundamental frequency component and its harmonic from the vocal-tract envelope by applying a logarithmic transformation. Most of these traditional methods work well for clean speech. However, their performance degrades severely in the presence of noise or other speech distortions. Further, pitch estimation under diverse conditions such as from noisy and telephone speech is a challenging task since the fundamental frequency (F0) component is filtered out by the telephone channel. In recent literature, there are some methods that have been proposed for the noisy and telephone speech scenarios. In [9], the pitch is estimated based on the autocorrelation technique with several modifications to prevent errors. Yegnanarayana *et al.* [10] proposed a zero-frequency filter (ZFF) based method, which gives instantaneous pitch estimates by exploiting impulse-like characteristics present in the excitation signal. ZFF method is robust to noise. However, for telephone speech, it does not give an accurate estimate of F0 since that spectral component is filtered out by the telephone channel. Kawahara *et al.* [3] proposed the TEMPO algorithm, which gives an instantaneous pitch estimate and is based on the wavelet analysis by finding out the pitch around a *fundamentalness measure*. Stephen *et al.* [11] proposed a technique by combining temporal and spectral methods, which showed significant improvements in accuracy for noisy and telephone speech.

http://dx.doi.org/10.21437/Interspeech.2017-1138

Quatieri *et al.* [8] proposed a 2-D approach for pitch estimation by computing the grating compression transform (GCT) over a narrowband spectrogram. The GCT was observed to coherently represent pitch information in a transformed 2-D space and used for pitch estimation. However, no performance evaluations were reported for noisy speech and telephone speech.

1.2. Our contribution

In this paper, we employ a recently proposed spectro-temporal demodulation technique to estimate the pitch [12]. Effectively, the technique operates in the joint time-frequency (t-f) plane and falls within the framework of two-dimensional (2-D) analysis of speech. Aragonda and Seelamantula showed that the demodulation of a narrowband (NB) spectrogram using complex Riesz transform (CRT) [13] yields smoothed time-frequency vocal-tract envelope (amplitude modulation or AM) and carrier spectrogram (frequency modulation or FM) [12] as shown in Figure 1. The carrier spectrogram represents joint spectro-temporal variations in the fundamental frequency and its harmonics. In this paper, we perform pitch estimation from the carrier spectrogram. The carrier spectrogram is devoid of interferences from the vocal-tract envelope. Hence, we hypothesize that accurate pitch estimates can be obtained using the carrier spectrogram. For instance, the pitch can be estimated from any time-frequency region of the carrier spectrogram, unlike the standard spectrogram, which gives reliable estimates only when one operates in the high-energy regions. We demonstrate that the proposed method is robust to noise. For clean speech signals, the proposed method performs on par with the state-of-the-art pitch estimation techniques, whereas, for telephone speech and degraded speech signals, the proposed method is superior and gave rise to accurate F0 estimates.

2. Spectrogram Demodulation Using the Riesz Transform

In this paper, we focus on narrowband speech spectrograms, and by default, a spectrogram refers to the narrowband flavor. The demodulation of speech spectrograms for voiced speech signals can be achieved using complex Riesz transform (CRT) [12], where small patches of the spectrogram, corresponding to voiced sounds, are modeled as 2-D AM-FM signals. A 2-D AM-FM spectrogram patch is modeled according to the following equation

$$S_W(\boldsymbol{\omega}) = V(\boldsymbol{\omega})(\alpha_0 + \cos\ \Phi(\boldsymbol{\omega})), \qquad (1)$$

where $\Phi(\boldsymbol{\omega}) = \Omega(\boldsymbol{\omega})(t\ \cos\ \beta(\boldsymbol{\omega}) + \omega\ \sin\ \beta(\boldsymbol{\omega}))$ denotes the argument of a 2-D cosine with spatial frequency $\Omega(\boldsymbol{\omega})$ and $\boldsymbol{\omega} = (t, \omega) \in \mathbb{R}^2$. The amplitude modulation, frequency modulation, and local orientation of the 2-D cosine are represented by $V(\boldsymbol{\omega})$, $\Phi(\boldsymbol{\omega})$, and $\beta(\boldsymbol{\omega})$, respectively. A suitable constant α_0 makes the spectrogram patch $S_W(\boldsymbol{\omega})$ a non-negative quantity. The AM-FM spectrogram patches are demodulated into 2-D AM and FM signals by applying the CRT. The full AM and FM component matrices are reconstructed by overlapping and adding the demodulated patches in the least-squares sense (OLA-LSE) [14]. Following this, we use the same technique to demodulate the spectrograms of the continuous speech signal. Figure 2 depicts AM and carrier spectrogram for a continuous speech signal, where AM represents the demodulated time varying smooth envelope corresponding to vocal-tract frequency response $V(\boldsymbol{\omega})$ and spectro-temporal frequency modulations are

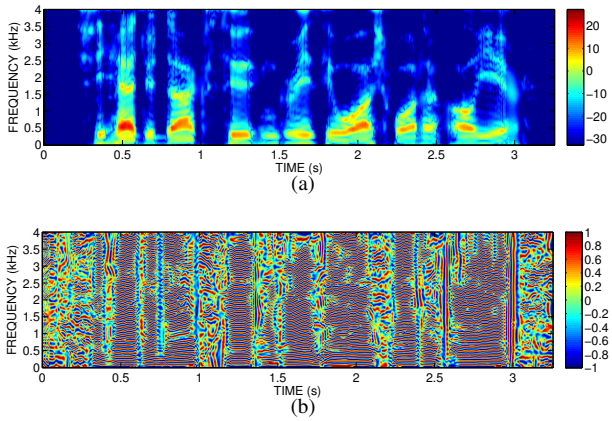

Figure 2: *AM and FM components for a continuous speech sentence, "She had your dark suit in greasy wash water all year." (a) vocal-tract spectrogram (dB), and (b) carrier spectrogram.*

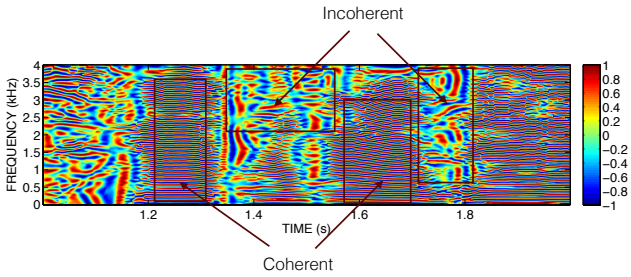

Figure 3: *Illustration of coherent and incoherent t-f regions in carrier spectrogram.*

captured by the estimated FM component of the underlying 2-D carrier $\cos\ \Phi(\boldsymbol{\omega})$. The goal of this paper is to estimate F0 by utilizing the carrier spectrogram.

2.1. Carrier analysis for a continuous speech signal

Carrier spectrogram shows spectro-temporal characteristics of speech sounds, such as quasi-harmonic nature of voiced speech and the time evolution of harmonics, referred as harmonic tracks. Further, visual inspection of carrier spectrogram (in Figure 3) highlights two distinct t-f patterns — coherent and incoherent t-f patterns. Coherent carrier patterns have a preferred orientation determined by the F0, and correspond to voiced speech sounds whereas such a structure is lacking in t-f regions that correspond to unvoiced sounds. The coherent and incoherent carrier patterns are determined by employing structure tensor matrix [13] on carrier spectrogram. This allows for achieving separation of coherent and incoherent t-f regions for voiced and unvoiced speech sounds, respectively.

3. Pitch Estimation Using the Carrier Spectrogram

We use carrier spectrogram for pitch estimation in two steps. First, we separate coherent t-f regions from the incoherent ones in the carrier spectrogram by employing the structure tensor matrix. Next, pitch is estimated by picking the dominant peaks from the carrier spectrogram in coherent t-f regions.

3.1. Structure tensor matrix and coherence map

Suppose $f(\boldsymbol{\omega}) : \mathbb{R}^2 \rightarrow \mathbb{R}$. A 2×2 structure tensor matrix is defined as follows:

$$J(\boldsymbol{\omega}) \triangleq \begin{bmatrix} (\psi * f_1^2)(\boldsymbol{\omega}) & (\psi * f_1 f_2)(\boldsymbol{\omega}) \\ (\psi * f_1 f_2)(\boldsymbol{\omega}) & (\psi * f_2^2)(\boldsymbol{\omega}) \end{bmatrix},$$

where $f_1(\boldsymbol{\omega})$ and $f_2(\boldsymbol{\omega})$ are the Riesz transform components of $f(\boldsymbol{\omega})$ along the time and frequency axes, respectively. A smoothing function ψ (typically a Gaussian) with standard deviation σ is applied locally to smooth out sudden variations of function values in the t-f plane. The eigenvalues and corresponding eigenvectors of structure tensor matrix give the distribution of gradient of underlying input function $f(\boldsymbol{\omega})$ in 2-D [13]. The relative discrepancy between the two eigenvalues of the structure tensor matrix is an indicator of the degree of uniformity present in the 2-D function, and is captured in the coherence $C(\boldsymbol{\omega})$ defined as follows

$$C(\boldsymbol{\omega}) \triangleq \begin{cases} \left(\frac{\lambda_1(\boldsymbol{\omega}) - \lambda_2(\boldsymbol{\omega})}{\lambda_1(\boldsymbol{\omega}) + \lambda_2(\boldsymbol{\omega})}\right)^2, & \lambda_1(\boldsymbol{\omega}) + \lambda_2(\boldsymbol{\omega}) \neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\lambda_1(\boldsymbol{\omega})$ and $\lambda_2(\boldsymbol{\omega})$ are the eigen values of structure tensor matrix $J(\boldsymbol{\omega})$. In order to exploit the structural properties present in the carrier spectrogram, we compute a 2-D time-frequency map from the carrier spectrogram, which is referred to as the coherence map and is shown in Figure 4(b). The coherence map is computed locally by dividing the carrier spectrogram into overlapping time-frequency patches. The structure tensor matrix for each t-f patch is obtained and coherence values are computed using (2). We used a rectangular 2-D window to divide the carrier spectrogram into patches of size 100 ms $\times$ 600 Hz. A full coherence map is obtained by using overlap-add procedure. The coherence map takes on continuous values between 0 and 1 according to (2). Values close to 0 represent unvoiced sounds, whereas values close to 1 correspond to voiced sounds in t-f plane. We use coherence map to separate highly coherent t-f regions in the carrier spectrogram from incoherent ones. Multiplication of carrier spectrogram by coherence map yields weighted carrier spectrogram (WCS), which retains coherent t-f regions of carrier spectrogram as shown in Figure 4(c). We use WCS for reliable pitch estimates.

3.2. Pitch estimation

The proposed pitch estimation algorithm is based on finding the peaks in carrier spectrogram within a frequency band from 0 to 1000 Hz, which covers a broad range of F0 values for both male and female speakers, even high-pitched ones. For a given frame, peaks in the carrier occur at F0 and its prominent harmonics. Hence, we use a peak-picking based approach in the carrier spectrogram for estimating average pitch. The carrier shows dominant peaks under strong voicing conditions, whereas under weak voicing conditions, the carrier exhibits non-sinusoidal nature and there could be several spurious peaks [15]. In addition, spurious peaks in frequency domain also exist due to analysis window artifacts. The non-sinusoidal nature of carrier leads to unreliable pitch estimates. Hence, we use WCS for reliable pitch estimates by retaining only peaks in WCS above a threshold which can be set close to 0 (we choose 0.05). Figure 5 illustrates that the coherence weighting suppresses the spurious peaks which are further removed by using the specified thershold. The average pitch $F_0^{(i)}$ (in Hz) for the i^{th} speech frame is estimated according to the harmonic mean

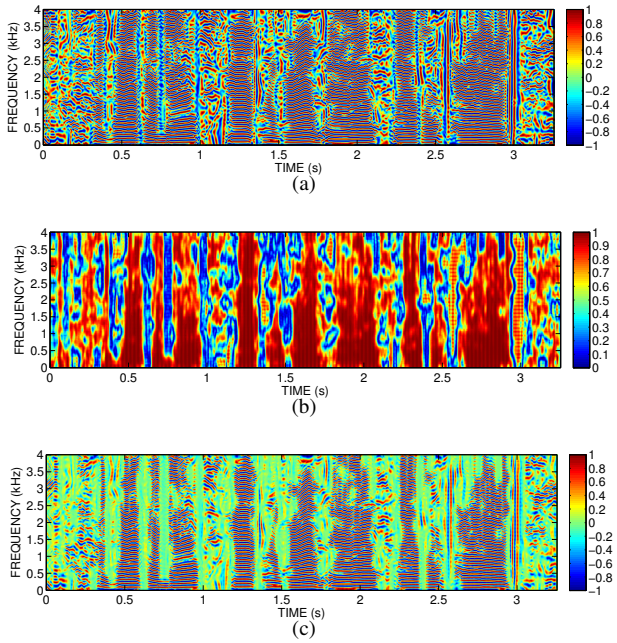

Figure 4: *(a) Carrier spectrogram, (b) coherence map, and (c) weighted carrier spectrogram (WCS), for the utterance, "She had your dark suit in greasy wash water all year."*

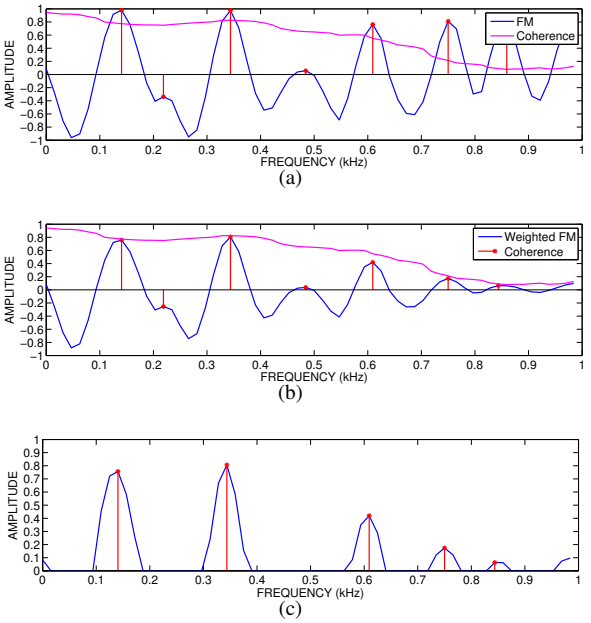

Figure 5: *Illustration of peak picking: (a) Actual carrier slice and coherence values, (b) WCS suppresses spurious peak around 0.5 kHz, and (c) Thresholding on WCS completely removes the spurious peaks.*

$$F_0^{(i)} = \frac{f_s}{N} \times \frac{1}{\frac{1}{K_i} \sum_{k=1}^{K_i} \left(\frac{1}{d_{k+1} - d_k} \right)}, \tag{3}$$

where N and f_s denote the FFT size and sampling frequency of speech signal, respectively. The number of peaks found in the

Table 1: *Performance evaluation for CMU-ARCTIC database, based on GPE (%) and SD (Hz).*

	Clean speech				Noisy speech (Input SNR: 0 dB)				Telephone speech			
	Male		Female		Male		Female		Male		Female	
Method	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD
Proposed	1.97	0.77	1.01	0.56	**9.38**	0.99	9.44	1.54	**4.93**	1.51	**3.28**	1.07
TEMPO	10.93	0.56	3.84	0.89	53.76	0.51	13.96	0.76	99.43	**0.10**	26.68	1.02
SRH	4.04	0.64	1.93	0.81	9.90	0.63	**5.98**	0.85	7.68	1.12	5.08	0.99
ZFF	**1.37**	**0.49**	**0.90**	**0.43**	47.06	0.99	76.26	1.53	99.97	0.25	9.08	0.99
YAAPT	5.53	**0.49**	3.71	0.55	13.71	**0.47**	6.69	**0.59**	8.85	0.85	7.15	**0.69**

Table 2: *Performance evaluation for KEELE database, based on GPE (%) and SD (Hz).*

	Clean speech				Noisy speech (Input SNR: 0 dB)				Telephone speech			
	Male		Female		Male		Female		Male		Female	
Method	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD	GPE (%)	SD
Proposed	15.61	4.68	10.88	8.46	17.93	5.67	**11.92**	9.90	18.35	4.66	14.54	8.03
TEMPO	6.53	3.35	10.01	7.18	29.57	4.40	43.15	7.36	96.79	6.48	41.84	7.50
SRH	3.86	4.06	13.38	7.77	**5.17**	4.37	13.19	7.92	32.70	5.86	13.82	7.98
ZFF	5.47	**2.23**	4.55	6.38	21.36	**2.78**	35.05	11.39	95.77	7.93	50.16	10.37
YAAPT	**2.21**	3.08	**3.14**	**5.85**	13.58	3.19	18.10	**5.72**	**6.86**	**4.11**	**6.06**	**6.51**

i^{th} speech frame within the specified frequency band is denoted by K_i and d_k denotes the location of the k^{th} peak. Note that, in contrast, the arithmetic mean would only make use of the first and last harmonic peak locations.

4. Results and Evaluation

The proposed method is evaluated on the CMU-ARCTIC and KEELE databases [16, 17]. We chose 200 utterances from the ARCTIC database, which includes 100 utterances each from the male `bdl` speaker and female `slt` speaker. The ground-truth average pitch contour is obtained by using the method given in [18], which is based on deriving glottal closure instants from the electroglottograph (EGG) signal. The KEELE database consists of simultaneously recorded speech and reference pitch computed from the laryngograph. The database consists of 5 male speaker and 5 female utterances sampled at 20 kHz. The wave files are resampled to 8 kHz for comparison. For evaluation, two objective measures were used namely gross pitch error (GPE) rate and the standard deviation (SD). The scores are evaluated for only those voiced frames which are common to both proposed pitch tracker and the ground truth. GPE is defined as the percentage of voiced frames with an estimated pitch value that deviate from the reference value by more than 20%. SD is defined as the standard deviation of the absolute difference between reference pitch and the estimated pitch values within 20% deviation.

We compare the performance of proposed method with four state-of-the-art methods: STRAIGHT [3], ZFF [10], SRH [19], and YAAPT [11]. Pitch estimation in STRAIGHT is based on the TEMPO algorithm, which uses wavelet transform. ZFF method performs epoch-based pitch estimation in every glottal cycle. The ZFF pitch estimator as given in [10] is followed in this paper. The SRH algorithm is based on the sum of the residual harmonics. YAAPT uses a combination of time and frequency domain pitch estimation methods followed by dynamic programming to get the final estimate of the pitch. In order to evaluate the robustness of the proposed method in noisy case, we use additive white Gaussian noise at 0 dB input signal-to-noise ratio (SNR). In the current work, telephone quality speech corresponding to each clean speech is simulated using the G.191 software tool provided by international telecommunication union (ITU) [20]. We have followed same steps as given by King *et al.* [21] to simulate the telephonic speech quality for both ARCTIC and KEELE databases. The results on the two databases are shown in Table 1 and Table 2. We observe that, on the CMU-ARCTIC database, the proposed method gives lower GPE scores for noisy and telephone speech than the other methods, whereas SD scores are comparable. The performance of the proposed method is better for telephone speech since it is able to estimate pitch by using any selected frequency band from carrier spectrogram. We selected the frequency range from 300 Hz to 1000 Hz from the telephone speech. We also observe that the method gives a consistent performance on CMU-ARCTIC database across male and female speakers for noisy speech signals, whereas on KEELE database, the proposed method showed comparable performance to the other methods except YAAPT.

5. Conclusions

We employed a spectro-temporal 2-D demodulation technique based on the Riesz transform for pitch estimation. Riesz transform demodulates the spectro-temporal narrowband spectrogram into carrier spectrogram consisting of harmonic/inharmonic patterns, which are free from any interferences due to vocal-tract AM envelope. The coherent/incoherent t-f regions in the carrier spectrogram are separated using the coherence map. Pitch is estimated from the coherence-weighted carrier spectrogram coherent patches using a peak picking-based approach. The proposed method allows for computing the pitch from any spectro-temporal patch. The resulting pitch estimate is robust to noise and gives higher accuracy for telephone-channel speech compared with other state-of-the-art algorithms. The future work is to explore multi-speaker pitch tracking and its application to speaker separation.

6. Acknowledgments

This work was supported by the Department of Electronics and Information Technology (DeitY) project on "Development of text-to-speech synthesis systems for Indian languages – Phase II."

7. References

[1] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, 2011.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[3] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Commun.*, vol. 27(3-4), pp. 187–207, 1999.

[4] D. T. Chappell and J. H. L. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 1998, pp. 885–888.

[5] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, no. 5, pp. 367–377, Dec 1972.

[6] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293–309, 1967.

[7] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 333–338, May 1999.

[8] T. F. Quatieri, "2-D processing of speech with application to pitch estimation." in *Proc. Interspeech*, 2002.

[9] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Opt. Soc. Am. A*, vol. 111, no. 4, pp. 1917–1930, 2002.

[10] B. Yegnanarayana and K. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 4, pp. 614–624, 2009.

[11] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 2002, pp. I–361.

[12] H. Aragonda and C. S. Seelamantula, "Demodulation of narrowband speech spectrograms using the Riesz transform," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 11, pp. 1824–1834, Nov 2015.

[13] C. S. Seelamantula, N. Pavillon, C. Depeursinge, and M. Unser, "Local demodulation of holograms using the Riesz transform with application to microscopy," *J. Opt. Soc. Am. A*, vol. 29, no. 10, pp. 2118–2129, Oct 2012.

[14] T. T. Wang and T. F. Quatieri, "Towards co-channel speaker separation by 2-D demodulation of spectrograms," in *Proc. IEEE Workshop on Applications of Signal Process to Audio and Acoustics*, Oct 2009, pp. 65–68.

[15] O. Fujimura, "An approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 1, pp. 68–72, Mar 1968.

[16] J. Kominek and A. W. Black, "The CMU-ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[17] F. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," *Children*, vol. 8, no. 12, pp. 30–50, 1995.

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

[19] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976.

[20] "ITU-T, recommendation G. 191, software tools for speech and audio coding standardization," *Int. Telecom. Union, Geneva, Switzerlan*, 2005. [Online]. Available: https://www.itu.int/rec/T-REC-G.191/en

[21] S. King and V. Karaiskos, "The Blizzard challenge 2009," in *Proc. Blizzard Challenge*, 2009.