Robust pitch estimation in noisy speech using ZTW and group delay function

RaviShankar Prasad, B. Yegnanarayana

Speech and Vision Laboratory,
International Institute of Information Technology, Hyderabad, India
ravishankar.prasad@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

Identification of pitch for speech signals recorded in noisy environments is a fundamental and long persistent problem in speech research. Several time domain based techniques attempt to exploit the periodic nature of the waveform using autocorrelation function and its variants. Other set of techniques utilize the harmonic structure in the spectral domain to identify pitch values. Either of these techniques suffer significant degradation in their performance in cases of noisy speech signals with low SNRs. The paper presents a robust technique to identify pitch values for speech signals. The proposed algorithm utilizes a speech analysis method called zero-time windowing (ZTW) where the signal is processed using a heavily decaying window, and the spectral characteristics are highlighted using the numerator of the group delay function. The amplitude contour of dominant resonances in the spectra are extracted, and processed further using a Gaussian window. The resulting contour reflects the energy profile of the signal which is utilized for estimation of the pitch values. The proposed algorithm is robust to degradations, and has been tested on several utterances with added noises. The algorithm exhibits significant increment in performance when compared to existing techniques.

Index Terms: pitch, zero time windowing, numerator of group delay function

1. Introduction

Identification of pitch is a persistent and important problem in speech research. Pitch is a perceptual attribute of human speech, which proves to be an important element in characterizing the way humans perceive information embedded in the signals. A natural course of change in pitch values results in the variation of speech quality and emotional expressions. Pitch identification is an important pre-processing task in almost all speech engineering areas. Speech recognition, speech coding, speech enhancement, speech conversion, speaker identification, voice activity detection and speech synthesis are some of the leading applications, which require reliable algorithms to identify pitch values. Pitch identification in speech signals is a problem which deals with the computation of its acoustic correlate, which is known as the fundamental frequency (f_0). Speech signals are pseudo-periodic in nature which is attributed to the regular opening and closing of the vocal chords during the production of voiced speech segments. Most of the pitch estimation techniques usually try to identify this periodic phenomenon over the length of a segment.

Several algorithms have been proposed over a period of time to address the problem. Pitch tracking algorithms can be classified mostly in either spectral or temporal domain based methods. Methods relying on the temporal domain utilize the cyclic nature of the voiced speech waveform, and use correlation based techniques to identify f_0. Most commonly used methods are the autocorrelation function (ACF) and the average magnitude difference function (AMDF) [1]. Spectral domain based methods, on the other hand, use block processing and discrete Fourier transform to compute the spectrum for speech segments. The harmonics present in the spectrum are identified, mostly using the frequency histogram and the cepstrum method, to compute the pitch information [2]. Another group of techniques which utilizes the knowledge from both these domains are also very popular. For degraded speech, either of these, i.e., periodicity in temporal domain and harmonicity in the spectral domain, are usually not suitable. Previous studies have emphasized the use of time and frequency domains together to identify pitch in the presence of noise [3].

Pitch identification in cases of speech recorded in degraded environments is a challenging task. For these signals, the periodic structure in speech is lost due to smearing of the time and spectral domain correlates for pitch. Popular approaches towards robust pitch estimation in noisy speech use hybrid techniques, incorporating complementary information from both the domains. ACF and AMDF functions have together been used for pitch extraction in noisy speech [1]. Other approaches use several time-frequency analysis techniques apart from the Fourier analysis to identify the pitch in noisy speech. In another paper, the authors have utilized the localization ability of the Wigner-Ville distribution for pitch identification in noisy speech [2]. A robust time-domain representation based on harmonic modeling of the speech signals has also been proposed to detect pitch in noise [4].

Some of the recently proposed algorithms for pitch extraction, namely pitch estimation filter with amplitude compression (PEFAC), YIN, robust algorithm for pitch tracking (RAPT) and TAPS, show high tolerance to elevated levels of noise. PEFAC is a recently developed algorithm which combines nonlinear amplitude compression with application of a comb filter [5]. This filter is applied in the log-frequency power spectral domain, and is effective in attenuating noise components in the spectrum. RAPT is another algorithm which uses peaks in the normalized cross correlation function (NCCF) and dynamic programming to compute the best candidate for f_0 for speech in noisy conditions [6]. TAPS is a technique which exploits the quasi stationary characteristics of natural speech and performs sub-harmonic summation (SHS) in the log-spectral domain to combat noise [7]. An alternate technique to SHS was later proposed, where harmonic peaks are summed over consecutive time frames, to reflect upon a temporal similarity, and is expressed as a sparse linear combination of a large set of peak spectrum exemplars obtained $a\text{-}priori$ from clean speech [8].

In this paper, we propose a technique to estimate the pitch of voiced speech segments which is robust for signals added with high amount of degradation. The presented algorithm uses

a speech analysis technique called zero time windowing (ZTW) [10]. The ZTW method is effective in the representation of signal characteristics in time and frequency domains with high resolution. We compute the dominant resonances from the signal spectrum estimated using the numerator of group delay function. The dominant resonances are capable of reproducing the instantaneous energy of the windowed signal. This energy contour has high values around the GCI locations. The proposed algorithm uses the maxima to identify the pitch values. The rest of the paper is organized as follows: Section 2 discusses the ZTW analysis of speech, extraction of DRF amplitudes and the steps to compute the pitch values for speech. Section 3 describes the database used and discusses the results obtained. Section 4 presents concluding remarks.

2. ZTW and the proposed algorithm

2.1. ZTW technique and numerator of the group delay

In the paper, we use a signal processing technique, ZTW to extract spectral features from the speech signals [10]. The ZTW method processes the speech using a heavily decaying window in time, given by $h[n]$ in Eq(2). The instantaneous spectral characteristics of the windowed signal are obtained using the *Hilbert envelope of the Numerator of Group Delay function* (HNGD). Let

$$x[n] = s[n] * h[n], \tag{1}$$

where $s[n]$ is speech signal, and $h[n]$ is the window function given as [10],

$$h[n] = \frac{1}{8sin^4(\pi n/N)}, \quad n = 0, 1, 2...N-1. \tag{2}$$

The spectrum is estimated using the Hilbert envelope of the numerator of the group delay (HNGD) function given by

$$g(\omega) = X_I(\omega)X'_R(\omega) - X_R(\omega)X'_I(\omega) \tag{3}$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ is the discrete time fourier transform (DTFT) of $x[n]$ and $X'(\omega) = X'_R(\omega) + jX'_I(\omega)$ is the DTFT of $nx[n]$. The HNGD has the ability to selectively highlight the peaks in the spectrum representing the prominent resonances in the analysis segment [11]. The analysis window $h[n]$ is shifted by one sampling instant, resulting in a high resolution characterization of speech segment in the time-frequency plane. The parameter *Dominant Resonance Frequency* (DRF) is the frequency of the strongest resonance in the HNGD spectrum for each window location. The DRFs have been proposed as it correlates to the dimensions of the most dominant cavity of the vocal tract responsible for the production of the speech segment [12]. In this study we propose the use of strength of the dominant resonances to identify the instantaneous energy profile of speech signals. The strength of DRFs proves to be consistent and robust, as it uses the high SNR regions in speech signals even in heavy degradations. The following section discusses the method in detail.

2.2. DRFs and amplitudes

The DRFs are the frequency plane locations of the dominant resonances of the speech segment. The DRF locations are determined by the characteristics of the analysis segment under consideration. The DRF amplitudes, on the other hand, are mostly determined by the signal energy at the point of application of the window function.

The energy of the dominant resonance is computed from the HNGD spectra at every sampling instant. Speech excitation signal has an impulse like phenomenon at the glottal closure instants (GCIs). The energy of the excitation signal is subsequently reflected in the signal samples following the GCI. The DRF amplitude contour, which closely approximates the instantaneous energy profile of the windowed segment preserves the high SNR regions of speech signals, and thus helps in computing the pitch even in heavy degradations.

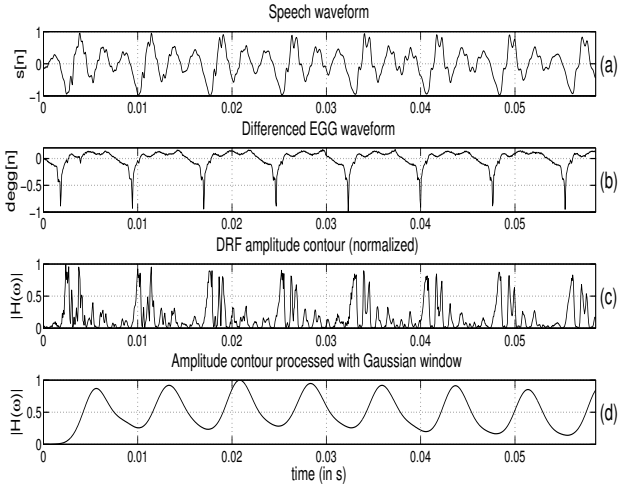

Figure 1: (a) Speech waveform. (b) Differenced EGG signal. (c)DRF amplitudes obtained using ZTW analysis with a window $N = 3ms$. (d) Amplitude contour convolved with a *Gaussian* window of $80ms$.

Figure 1 illustrates the behavior of the amplitude of the dominant resonance amplitudes obtained using the ZTW analysis. Figures 1(c) and (d) show the periodic behavior of the amplitude contour. The gross peaks in 1(c) occur around the GCI locations as can be seen in comparison in the differenced EGG waveform in figure 1(b). The DRF amplitude contour is convolved with a *Gaussian* window of length $5ms$ as shown in the figure 1(d). The peaks can be identified using a peak picking algorithm, which will eventually lead to f_0 detection.

2.3. Computation of pitch

The algorithm presented here is generic in nature, and is applicable for speech signals recorded in degraded environments. The voiced regions in the speech signals are assumed to be known *a priori*. The DRFs are computed from the HNGD spectrum obtained from a ZTW analysis with a small window size $(N = 3ms)$ for these segments. A small window of this duration will help to obtain the instantaneous signal characteristics, and therefore capture the signal energy profile in a precise manner. Following are the steps to obtain the pitch for voiced segments in speech signals.

- Compute the HNGD spectrum for the given speech signals using the ZTW analysis with a window size of $N = 3ms$ and a shift of each sampling instant.
- Compute the DRFs from the HNGD spectrum by identifying the frequency of the strongest resonance.
- Identify the amplitude for the dominant resonances from the HNGD spectra for every window location. Normalize the amplitude contour between 0 and 1.

- Smooth the amplitude contour of resonances, also known as the DRF amplitude contour, using a 5-point median filter to remove any outliers.
- Process the DRF amplitudes using a Gaussian filter of length $5ms$ to further highlight the peaks in the contour.
- Identify the peaks in the processed DRF amplitude contour using a peak picking algorithm.
- The pitch period value for an analysis segment is computed by calculating the distance between successive peaks.

Figure 2 illustrates the steps of the proposed algorithm to compute the pitch of a voiced speech segment for different cases of degraded speech. Figs. 2(a) and (b) show a voiced speech segment along with the DRF amplitude contour obtained using the ZTW analysis. The periodic structure of the amplitude peaks clearly highlights the pitch component of the analyzed signal. It is also evident from figs. 2(c) and (d), that the DRF amplitudes are robust to high levels of degradations (SNR = $0dB$). Figs. 2(b), (c) and (d) show the normalized DRFs amplitude contour for the given speech segment. In the case of real wideband noises such as *'babble'* and *'vehicle (volvo)'*, the DRF amplitudes serve as reliable cues to track the signal energy profile. This can further be inferred from fig. 2(e) which compares the pitch values obtained for all the 3 cases. The pitch values obtained are consistent over the length of the segment, reflecting the robustness of the proposed method. The following section discusses the results obtained by the proposed algorithm for clean and degraded speech signals.

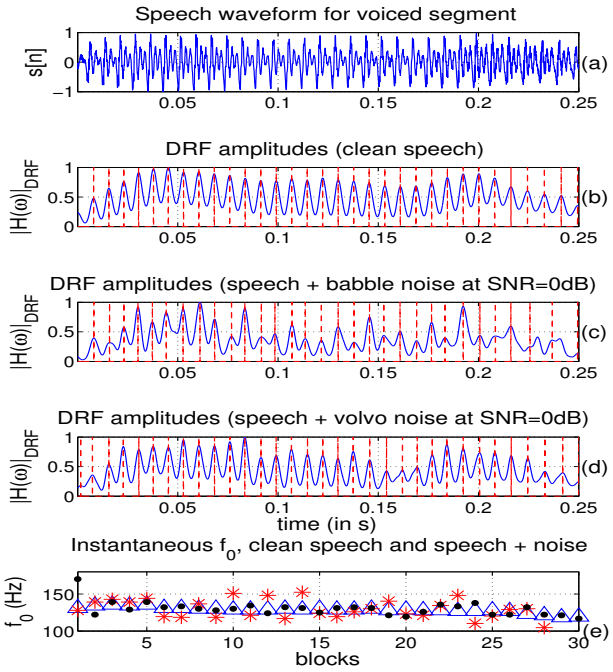

Figure 2: (a) Speech waveform. (b) DRF amplitudes obtained using ZTW analysis with a window $N = 3ms$. (c) and (d) are normalized and processed DRF amplitudes for speech corrupted with babble noise and volvo noise at SNR = $0dB$ respectively. (e) Instantaneous f_0 obtained from amplitude peak markers for all 3 cases, ($\triangle$) =clean speech, ($*$) = speech+babble noise and ($\cdot$) = speech+volvo noise.

3. Database and Results

The proposed algorithm has been tested for around 200 utterances from 3 different speakers in the CMU-Arctic database. The database also provides the corresponding EGG signals for these utterances, which help to establish the ground truth for the pitch values. The voiced regions in the speech signals are obtained by EGG signals by computing their short time energy and using a threshold. The GCI locations in the EGG signal are computed using the SIGMA algorithm [9]. The difference in locations for these GCIs is averaged over a block to compute the ground truth for the the pitch values. The algorithm is evaluated for clean speech segments and for 3 types of added degradation, *white*, *babble* and *volvo* noise. The degradations are obtained from the $NOISEX-92$ database [13], and each clean utterance was added with the noise at SNRs $20, 10, 5, 2, 0, -2, -5$ and $-10dB$. We choose the metric, *gross pitch error* (GPE) and *fine pitch error* (FPE), to report the performance of the algorithm. An error in the pitch measurement is regarded as GPE when the average pitch of the voiced segment deviates from the ground truth by 20%. FPE is computed as the root mean square (RMS) percentage of the standard deviation of relative f_0 error distribution, when the error is below 20%. The GPE results from the proposed algorithm are compared with two different methods discussed in Section 1, namely, PEFAC and TAPS-AutoC. The PEFAC is a frequency domain based algorithm, whereas TAPS-AutoC is a time-frequency domain based algorithm. Fig.3 shows the GPE results obtained for the given database from these algorithms. It can be seen in the figures that for the cases of clean and degraded speech with different degradations, the DRF amplitude based method is more robust than the other two techniques. In the case of high degradation with SNR between -5 to $-10dB$, the DRF amplitudes give reliable cues to recreate the signal energy profile. It is a general trend for all pitch identification algorithms to give high GPE for wideband noise such as babble. GPE values for the proposed technique are lower by less than 20%.

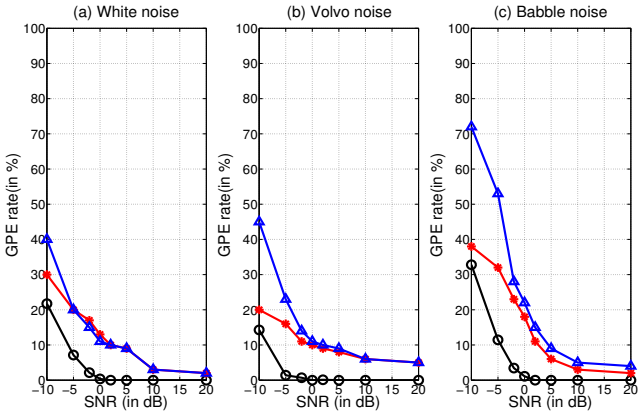

Figure 3: GPE values for the proposed algorithm (o) for speech degraded with noise at multiple SNR in cases of (a) white noise, (b) volvo noise and (c) babble noise. Also shown are results obtained from PEFAC ($*$) and TAPS-AutoC ($\triangle$)

The techniques, PEFAC and TAPS-AutoC, perform well for bandpass noises. It is mainly because the methods exploit the harmonic behavior of the signal over a block of time. It is therefore possible to extract an accurate harmonic information from

the accumulated information from less affected sections of the spectrum. The added noise samples do not significantly smear the low frequency characteristics of small segments of speech. The speech segments with duration smaller than a pitch period therefore retain most of their spectral characteristics. On the other hand, the periodic behavior of a larger block, covering a few pitch periods, is significantly smeared by the addition of noise. This in turn affects the performance of the pitch identification algorithms which extract information from a relatively longer segment of speech signal. The advantages with using the ZTW analysis technique is that the analysis segment considered for spectrum estimation is of smaller duration. The HNGD technique highlights the spectral resonances from such small segment. The dominant resonances computed from the ZTW spectra, which usually occur in the low frequency bands for voiced segments of speech[12], are consistent to obtain the energy profile of the signal. The FPE values calculated over the database also lie in the range of 0.3%-1.9% as compared to 1%-5.6% for the other two methods.

4. Conclusions

The paper proposes a new and robust technique for identification of pitch from voiced segments of speech signals even in the cases of heavy signal degradation. The presented algorithm utilizes a signal processing technique ZTW which effectively highlights the spectral characteristics for small analysis segments in speech signals. The pitch values obtained using the proposed technique are accurate. The error rate for degraded speech signals is also very low as compared to the other state-of-art techniques. For cases of wideband noises like *babble* and *pink*, the algorithm produces accurate pitch values at low SNR values such as -5 and $-10dB$.

5. References

[1] Li Hui; Bei-qian Dai; Lu Wei, "A Pitch Detection Algorithm Based on AMDF and ACF," IEEE International Conference on Acoustics, Speech and Signal Processing, 2006.

[2] Noll, A. Michael, "Cepstrum pitch determination." The journal of the acoustical society of America, vol. 41, no. 2, pp. 293–309, 1967

[3] W.W. Zhao, T. Ogunfunmi, "Formant and pitch detection using timefrequency distribution," International Journal of Speech Technology, vol. 3, no. 1, pp. 35–49, 1999

[4] Shahnaz, Celia, W-P. Zhu, and M. Omair Ahmad, "A pitch extraction algorithm in noise based on temporal and spectral representations." IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.

[5] Gonzalez, Sira, and Mike Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)." In Proc. Euro. Sig. Process. Conf, pp. 451-455, 2011.

[6] Talkin, David, "A robust algorithm for pitch tracking (RAPT)." Speech coding and synthesis, vol. 495, 1995.

[7] F. Huang and T. Lee, Pitch estimation in noisy speech based on temporal accumulation of spectrum peaks, in Proc. Interspeech 10, Sept 2010, pp. 641–644

[8] Huang, Feng, and Tan Lee, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique." IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 1, pp. 99–109, 2013.

[9] Thomas, Mark RP, and Patrick A. Naylor. "The SIGMA algorithm: A glottal activity detector for electroglottographic signals." IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 8, pp. 1557-1566, 2009.

[10] Bayya, Yegnanarayana, and Dhananjaya N. Gowda. "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function." Speech Communication vol. 55, no. 6, pp. 782-795, 2013.

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

[12] Prasad, RaviShankar, and B. Yegnanarayana. "Acoustic segmentation of speech using zero time liftering (ZTL)." Proc. Interspeech 2013, pp. 2292–2296, Aug 2013.

[13] Varga, A., Steeneken, H. J., Tomlinson, M., & Jones, D. (1992). The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, Malvern, England.