Implementation of cepstrum based voiced / unvoiced Classification

Bhumika Nirmalkar[1], Dr. Sandeep Kumar[2]

Department of Electronics and Telecommunication
Rungta College of Engineering and Technology Bhilai, India
bhuminirmalkar@gmail.com

Department of Electronics and Telecommunication
Rungta College of Engineering and Technology Bhilai, India
skumardr20@gmail.com

Abstract: ***In this paper, an enhanced cepstrum based method for the voiced/un-voiced (V/UV) classification has been simulated using MATLAB. Voicing choices are made utilizing a feature voiced/unvoiced characterization calculation taking into account factual investigation of cepstral peak. Execution investigation on an extensive database shows in terms of % classification accuracy. This algorithm is likewise appeared to be robust to additive noise as compared to traditional cepstrum based classifiers.***

Keywords: cepstrum, voiced/unvoiced classification, cepstral peak, TIMIT database.

1. Introduction

Voiced/un-voiced (V/UV) detection refers to identification of regions within the speech signal with sturdy vocal band activity. Throughout the assembly of voiced speech, the vocalism is excited by the vibration of vocal folds, leading to a quasi-periodic speech signal. The unvoiced speech is created once the air is tried and trues a slim constriction within the wind pipe, generating a noise like random signal. Silence happens within the absence of any excitation to the vocal tract system and contains solely background. Non-voiced speech includes unvoiced speech and silence.

It is therefore natural to combine voiced/voiceless segmentation. There are several algorithms exploring voiced/voiceless segmentation as a by-product of pitch detection. It looks that the mix would greatly improve the process potency. However, the particular results are typically increase of system complication and reduce of system operate. Particularly at the boundary of voiced and voiceless speech, those algorithms are going to be during a perplexity to correctly classify the sound. the problem originates from the different needs of those two tasks.

Voiced/voiceless segmentation would demand comparatively short analysis frame. An extended frame at the boundary is probably going to hide each voiced and voiceless sounds, and will be tough to be classified. The contradiction in terms of analysis frame becomes additional serious in continuous speech recognition as a result of short initials like b, d and g would probably be hid by adjacent vowels in a very long process window. Also, it's documented that voiced sounds square measure solely quasi-periodic. Factors throughout vocalization, as well as disturbance from outside, may result in irregular waveforms. If the segmentation between voiced and voiceless sounds depends only on the degree of speech regularity, the popularity rate will then be greatly attenuate.
While speech signal process applications like language identification [1], multi-rate speech coders [2,3], speech signal modeling [4] need classification of the speech signal into voiced, unvoiced and silence (V-UV-S) regions, there are some prominent speech signal analysis applications like identification of the vocal organ closure instants (GCIs) [5], pitch frequency estimation [6,7], that need information of solely the voiced regions of the speech signal. The requirement of boundaries of voiced regions of those applications will be catered by a V/NV detection technique requiring a lot of less process complexity than V–UV–S classification ways. Detection of voiced regions from the speech signal within the presence of noise finds use in automatic speech recognition (ASR) [8].Applications like speech sweetening [9], diagnosing of pathological voice disorders [10, 11], emotion recognition [12,13] have confidence the estimation of pitch frequency and detection of GCIs from buzzing speech signals. A noise resilient V/NV detection technique will offer reliable detection of voiced regions for pitch frequency determination and extraction of GCIs from speech signals distorted by noise.

In this paper an enhanced cepstrum based voiced/unvoiced speech classifier proposed in [14], has been simulated using MATLAB. Performance of this scheme has been compared with conventional cepstrum based speech classifier in terms of % voiced/unvoiced classification accuracy.

The rest of the paper is organized as follows. In Section II, a close description of the implementation of enhanced cepstrum based voiced/unvoiced speech classifier is given. In Section III, the results of the performance analysis are presented. Concluding remarks are given in Section IV

2. Implementation of cepstrum method

The cepstrum, outlined because the real a part of the inverse Fourier remodel of the log-power spectrum, includes a sturdy peak reminiscent of the pitch amount of the voiced speech section being analyzed [15]. A 512-point quick Fourier remodel (FFT) was found enough for correct computation of the cepstrum.

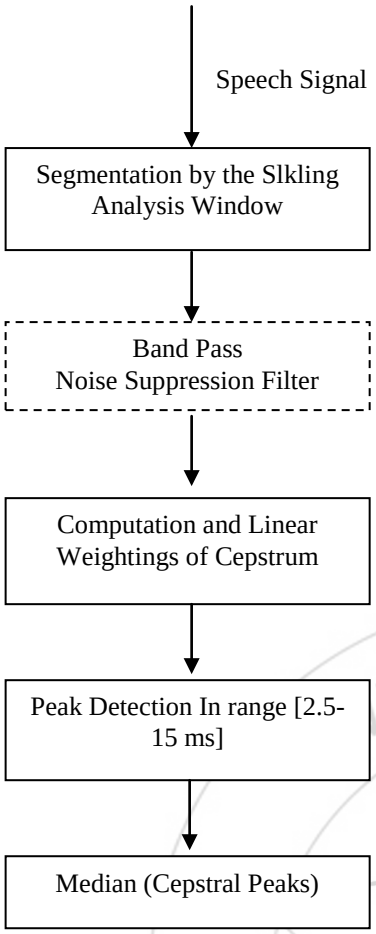

Fig. 1. Flowchart of the algorithm.

The cepstral peaks reminiscent of the voiced segments square measure clearly resolved and quite sharp. Hence, the height selecting theme is to determine the cepstral peak within the interval [2.5–15 ms], corresponding to pitch frequencies between 60–400 cycle per second, that exceeds some specified threshold. Since the cepstral peaks decrease in amplitude with increasing quefrency, a linear cepstral weight is applied over the 2.5 to fifteen ms vary. The linear cepstral weight with vary of one to eight was found by trial and error by victimization periodic pulse trains with variable periods because the input to the pitch determination program. The strength and existence of a cepstral peak for voiced speech is dependent on a spread of things, together with the length of the analysis window applied to the signal and therefore the formant structure of the input signal. The window length and therefore the relative positions of the window and the speech signal can have tidy result on the peak of the cepstral peaks [16]. If the window length is a smaller amount than 2 pitch period long, a powerful indication of cyclist cannot be expected. The longer the window, the larger the variation of the speech signals from the start to the tip. Therefore, considering the tapering result of the analysis window, the window length was set to forty ms to capture at least two clearly outlined periods within the windowed speech section.

The extraction of the cepstral peaks may be a settled downside. However, to determine if a cepstral peak represents a voiced section requires a choice level (i.e., the threshold) that's not settled and powerfully depends on the characteristics of the input speech. A plot of the histograms of the cepstral peaks cherish four different male and feminine utterances are shown in Fig. 1. In order to determine the optimum threshold, applied mathematics distributions of the cepstral peaks cherish the voiced and unvoiced segments of speech should be glorious beforehand. This a priori info isn't generally provided. If such info were accessible, a most a posteriori likelihood (MAP) estimate of the initial-threshold might be obtained by finding the worth of that the subsequent value function was minimized:

$$\eta(\theta) = P_v \int_{-\infty}^{0} f_x(x)dx + P_{uv} \int_{0}^{\infty} f_{uv}(x)dx \qquad (1)$$

Where P_v and P_{uv} denote the probabilities that speech is voiced or unvoiced, respectively. The functions $f_v(x)$ and $f_{uv}(x)$ represent the statistical distributions of the cepstral peaks associated with voiced and unvoiced segments of the speech signal, respectively. Similar expressions can be used to determine the optimum thresholds corresponding to the other features.

It is a widely known undeniable fact that the cepstral peaks reminiscent of the unvoiced segments has smaller magnitudes than those associated with the voiced segments. However, the regions that contain voiced and unvoiced cepstral peaks overlap and an absolute discrimination is not potential. It should be noted that, albeit the particular applied math distributions were renowned, the initial-threshold obtained in (1) may not strictly discriminate between voiced and unvoiced cases as a result of the inescapable overlapping between the regions.

A sensible approach is to seek a value that minimizes some important error criteria. Supported applied mathematics analysis of the observations and the properties mentioned on top of, it had been found that the median of the cepstral peaks is comparatively an honest criterion to be used as the initial-threshold.

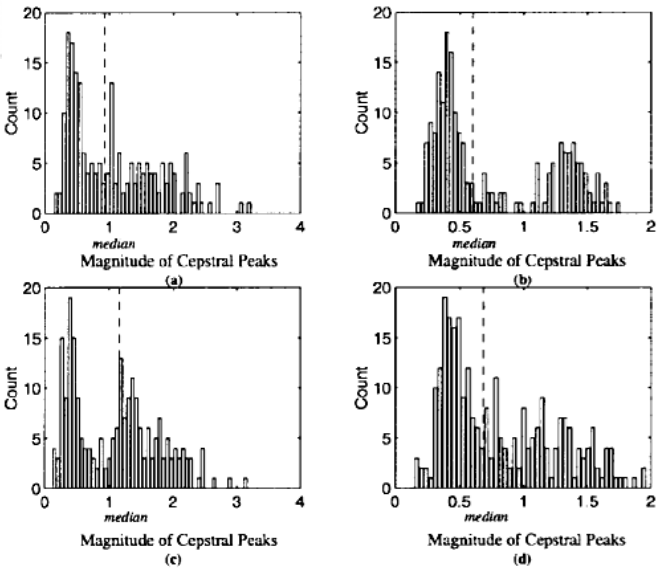

Fig. 2. Histograms of cepstral peaks [14] (a), (c) Distributions for two different male speakers. (b), (d) Distributions for two different female speakers

This alternative of the edge divides the set of observations into two subsets of equal range of entries. These regions are outlined as follows:

$$R_C^L = \{C_i \mid \min(C) \le C_i < median(C)\};$$
$$R_C^R = \{C_i \mid median(C) \le C_i < \max(C)\};$$
(2)

Where $C = \{C_m\}_{m=1}^{M}$ represents the set of all cepstral peaks, M is the total number of speech segments used in the experiment, and C_i denotes the i_{th} cepstral peak. In practice, the parameter M is equal to the number of segments in the speech file being analyzed.

It must be noted that the selection of median of a feature because the initial-threshold for preliminary classification of that feature doesn't constrain the number of voiced and unvoiced frames in an auditory communication. At the end of the primary pass, the median of the cepstral peaks is computed and used because the initial-threshold for the second pass. Alternative values for the threshold like mean and a proportion of the utmost price of the corresponding feature, also as a constant-threshold were conjointly investigated. These values were either signal-independent or powerfully affected by extreme values measured for the corresponding feature.

3. Result and discussion

The performance of the algorithm is evaluated on speech data taken from TIMIT database. The speech material used in our experiments contained 186 speech files, equivalent to approximately 50 000 speech frames at 10 ms frame update rate, with lengths starting from 2 to 15 s and covered a variety of speakers. An equal number of male and female speakers from various dialect regions were utilized. The following objective error measures. Voiced-to-unvoiced (V-UV) and unvoiced-to voiced (UV-V) error rates denote the accuracy in properly classifying voiced and unvoiced intervals, severally. A UV-V error happens when AN unvoiced frame is assessed mistakenly as voiced. On the other hand, a V-UV error happens if a voiced frame is detected as unvoiced by the formula. These errors square measure computed by averaging the per-frame UV-V and V-UV errors over all frames within the information.

The weighted gross pitch error (GPE) [17, 18] represents a properly Classified voiced frame wherever the reference and therefore the calculable pitch frequency tracks take issue in harmonic. This is often outlined as follows:

$$GPE = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{E_k}{E_{\max}}\right)^{1/2}\frac{f_k - \hat{f}_k}{\hat{f}_k}$$
(3)

Where K denotes the number of elements in the set of all correctly classified voiced indices in the database, E_{max} represents the maximum short-time energy, and f_k and $\hat{f}_k$ are the reference and estimated pitch frequencies for the k_{th} frame, respectively. It is obvious that a standard and perfectly labeled database does not exist. A labeled reference database was generated using 186 speech files taken from the TIMIT database.

Table 1 : Performance of the cepstral algorithm at different segmental snr's for male speakers, where the additive noise is a zero-mean white Gaussian noise. Gpe, v-uv, and uv-v denote gross pitch error, voiced-to-unvoiced error rate, and unvoiced-to-voiced error rate, respectively

Method used	SSNR (dB)	GPE (%)	V-UV (%)	UV-V (%)
Conventional Cepstrum Method	10	0.85	1.49	0.95
	5	1.51	1.96	1.58
	0	2.43	2.62	2.29
Enhanced Cepstrum Method	10	0.64	1.27	0.88
	5	1.32	1.66	1.42
	0	2.14	2.23	2.18

As already mentioned, the median of the options, on the typical, provides a lot of acceptable values for the thresholds to roughly distinguish between voiced and unvoiced regions in preliminary classification. Nevertheless, this selection doesn't prohibit the ultimate classification of voiced associate degreed unvoiced speech segments in a vocalization. In fact, the output results for several better-known pitch tracks were rigorously examined, and also the final results failed to show any restriction on the number of voiced and unvoiced frames. The projected algorithmic rule was applied to many cases wherever the share of voiced and unvoiced frames was completely different from five hundredth, and sensible results were obtained. It should be noted that the initial-thresholds area unit set per file and they area unit passionate about the characteristics of the input speech file. Moreover, the initial price obtained for the cepstral threshold is adapted in consecutive voiced segments.

TABLE 2 : Performance of the cepstral algorithm at different segmental snr's for female speakers, where the additive noise is a zero-mean white Gaussian noise. Gpe, v-uv, and uv-v denote gross pitch error, voiced-to-unvoiced error rate, and unvoiced-to-voiced error rate, respectively

	SSNR (dB)	GPE (%)	V-UV (%)	UV-V (%)
Conventional Cepstrum Method	10	1.78	1.46	0.89
	5	4.20	1.99	1.65
	0	9.05	4.02	4.72
Enhanced Cepstrum Method	10	1.67	1.17	0.77
	5	3.99	1.78	1.46
	0	8.83	3.86	4.22

The results of the analysis for male and female speakers at totally different SSNR's square measure shown in Tables 1

and 2. The intuitive reasoning for maintaining the performance under noisy condition are often summarized as follows:

1) Noise samples area unit unrelated from one section to following segment;

2) Cepstral coefficient at high quefrency, that improves the detect ability of low-frequency pitch peaks;

3) The employment of a multi-feature classification algorithmic rule and applied math analysis of data;

4) The employment of trailing and correction algorithm;

5) The employment of median smoothing to get rid of single and double errors in readjustment and pitch frequency information.

4. Conclusion

An enhanced cepstrum based method for the voiced/un-voiced (V/UV) classification has been simulated using MATLAB. Results of voiced/unvoiced classification are presented for TIMIT database. Result of % voiced/unvoiced speech classification shows that, results obtained for the enhanced cepstrum based method is more accurate as compared to conventional cepstrum method.

References

[1] B. Yin, E. Ambikairajah, F. Chen, "Voiced/unvoiced pattern-based duration modeling for language identification," in: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 4341–4344.

[2] E. Paksoy, J. Carlos de Martin, A. McCree, C.G. Gerlach, A. Anandakumar, W.M. Lai, V. Viswanathan, "An adaptive multi-rate speech coder for digital cellular telephony", in: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Phoenix, USA, vol. 1, 1999, pp. 193–196.

[3] A.M. Kondoz, "Digital Speech: Coding for Low Bit Rate Communication Systems," Wiley, England, 2004.

[4] P. Sircar, R.K. Saini, "Parametric modeling of speech by complex AM and FM signals", Digital Signal Processing 17 (6) (2007) 1055–1064.

[5] P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," IEEE Transactions on Audio, Speech and Language Processing 15 (1) (2007) 34–43.

[6] B. Resch, M. Nilsson, A. Ekman, W.B. Kleijn, "Estimation of the instantaneous pitch of speech", IEEE Transactions on Audio, Speech and Language Processing 15 (3) (2007) 813–822.

[7] D. Joho, M. Bennewitz, S. Behnke, "Pitch estimation using models of voiced speech on three levels", in: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Honolulu, USA, vol. 4, 2007, pp. 1077–1080.

[8] P. Jancovic, M. Kokuler, "Voicing-character estimation of speech spectra: application to noise robust speech recognition, in: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Toulouse," France, vol. 1, 2006, pp. 257–260.

[9] Y. Kuroiwa, T. Shimamura, "An improvement of LPC based on noise reduction using pitch synchronous addition," in: IEEE Proceedings of the International Symposium on Circuits and Systems, Orlando, USA, vol. 3, 1999, pp. 122–125.

[10] S. Jang, S. Choi, H. Kim, H. Choi, Y. Yoon," Evaluation of performance of several established pitch detection algorithms in pathological voices", in: IEEE Proceedings of the International Conference on Engineering in Medicine and Biology Society, Lyon, France, 2007, pp. 620–623.

[11] C. Manfredi, M. D'Aniello, P. Bruscaglioni, A. Ismaelli," A comparative analysis of fundamental frequency estimation methods with application to pathological voices," Medical Engineering and Physics 22 (2) (2000) 135–147.

[12] A. Tawari, M.M. Trivedi, "Speech emotion analysis in noisy real-world environment", in: Proceedings of the International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 4605–4608.

[13] S.G. Koolagudi, R. Reddy, K.S. Rao," Emotion recognition from speech signal using epoch parameters," in: Proceedings of the International Conference on Signal Processing and Communications, Bangalore, India, 2010, pp. 1–5.

[14] A., S., A.S. Spanias. "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm". Speech and Audio Processing, IEEE Transactions y;7(3):333-8. May 1999.

[15] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, Feb. 1967.

[16] L. R. Rabiner and R. W. Schafer, "*Digital Processing of Speech Signals"*.Englewood Cliffs, NJ: Prentice-Hall, 1978.

[17] B. G. Secrest and G. R. Doddington, "Postprocessing techniques for voice pitch trackers," in *Proc. IEEE ICASSP'82*, pp. 172–175.

[18] V. R. Viswanathan and W. H. Russell, "New objective measures for the evaluation of pitch extractors," in *Proc. IEEE ICASSP'85*, pp. 11.10.1–11.10.4.