Modern pitch detection methods in singing voices analyzes

Mateusz Gawlik
Department of Mechanics and Vibroacoustics, AGH University of Science and Technology, Cracow, Poland.

Wiesław Wszołek
Department of Mechanics and Vibroacoustics, AGH University of Science and Technology, Cracow, Poland.

Summary
Voice analyzing problems, presented in the literature, concern mainly recognition, identification, verification or diagnostics of speakers. There has been a plethora of work focused only on speech surveys. Unfortunately, a range of studies of singing was not the subject of great interest by scientists. Results from presented research show that the most elementary parameter used in voice analyzing is fundamental frequency (*F0*). In this paper we presented pitch detection methods of singing signal in hope of choosing the best one. We decided to compare four detection methods, which according to the literature present promising results: *Zero Frequency Band Filtering* (*ZBF*), *SEDREAMS* algorithm, *Modified Zero Frequency Resonator* (*MZFR*) and *DYPSA* algorithm. The first conclusion from series of experiments is necessity of choosing probes to analyze in the proper way, because singing voice can change its basic frequency approximately by two octaves. This paper presents a precise analysis of the pitch detection by selected methods in singing voices analysis. As the most precisely method for singing voices analyzing after the study we classified *Zero Frequency Band Filtering* method.

Pacs no. 43.75.+a, 43.60.+d

1. Introduction

Though wide range of similarities during production process between speech and singing applying the speech processing techniques into singing acoustic signals analyzes is not straight forward. This is because in the speech voice the vocal tract filtering effect could be removed from output using an inverse-filtering method [1] [2] [3] and scientists are able to describe it by Linear Prediction of speech [4] [5]. Unfortunately simple linear source-filter theory, cannot be useful for singing voices applications as the result of broad range of fundamental frequency and timbre in various singing styles. Pitch for this signals could approach six or even eight octaves for the most talented singers and adjacent harmonics have higher distances between themselves in a frequency domain [6] [7] due to non-linear characteristic of these signals. Moreover difficulties in singing voice processing are connected with vibrato, high source filter interaction, changing dynamics of signal in short time frames and different tempos of the songs. Following further we can observe differences in singing categories and techniques [8] generating difficulties in the singing voice modeling process as a whole. Furthermore, humans generate voice in four laryngeal mechanisms, which are associated with glottis biomechanical configuration [9]:

- M0 – is the way to produce the lowest tones, mainly appears in speech. It could be found in vocal fry, pulse or strohbass voices.
- M1 – appears in chest and modal voices and male head registers. Sopranos can change laryngeal mechanism to M1 to sign lower pitches.
- M2 – natural for falsetto male loft and female head voices
- M3 – a method of voice production to reach the highest pitches: it appears in whistle, flageolet or flute registers.

In [10] authors proved an impact of laryngeal mechanism on *F0* detection efficiency. The best singers have the ability to produce tones from all their tessitura without any vocal tract shape transition. This fact will be used in the future authors' surveys.

This research is an introduction to a system, supporting singers' training process. As the datum point of our work we decided to investigate which modern method of pitch detection, common known from speech processing can be adopt in singing analyzes. Results of this survey will bused in future authors' studies.

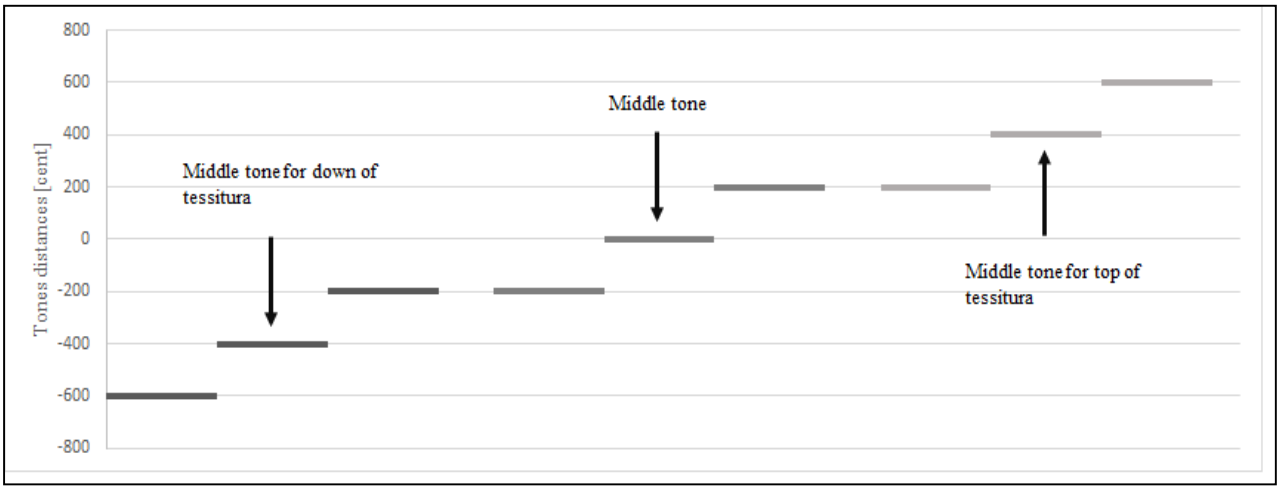

Figure 1. Distances between tones during exercising for soprano voice

The organization of this paper is as follows: the second chapter describes the theory and literature research of the paper's topic. Third part of the article consist of data analyzing way and error metrics explanation. In the last parts the results of survey were discussed and an additional analysis, containing S*ignal to Noise Ratio* was done. In the last chapter conclusions were described and the view for future surveys.

2. Theoretical basics

In this chapter we introduced a few modern glottal closure detection methods and experiment protocol. We chose methods used for glottal closure instants, because of its' close connection with voice signal pitch.

2.1. Existing methods

Voice analyzing topic has been present in the literature since 1950s [11]. Among existing solutions a lot of works describe only pitch detection algorithms [12] [13] [14] [15] [16]. Some of them were tested in singing signals cases, where instantaneous fundamental frequency was estimated by epochs locations, connected with Glottal Closure Instants (*GCIs*) [17] [18] [19]. In this paper we studied four methods, developed in the recent years, as follows:

- *Modified ZFF Method for Singing Voice* [18] – this method is dedicated for singing signals, based on *GCIs* isolation by passing the signal through cascade of three ideal zero frequency resonators
- *DYPSA* algorithm [20] – estimates *GCIs* by linear prediction algorithm
- *SEDREAMS* algorithm [21] – the most robust method for singing signals according to [22]. This method uses the LP residual to find *GCIs* locations from mean-based signal, counted for specific window function.
- *Zero Band Filtering* [14] – is a supplement method of *Zero Frequency Resonator*, proposed in [17]. The main modification in comparison to *ZFR* is including length of radius of unit circle in the z-plane. It has been tested for singing signals in the past, but there is no information about its robustness in terms low, medium and high voices.

2.2. Experiment description

Especially for this purpose dedicated dataset was created. We registered both acoustic as well as electroglottographic wavegrams for total number of 15 singers. *EGG* signal was registered as the reference to evaluate - by designated error metrics – performance of pitch detection methods and laryngeal mechanism classification [9]. Database consists of wide range of sounds – from *baritone* to *soprano* voices. During single session we asked singers to do simple singing exercises, consisting of sustained, single notes generation from the middle, beginning and the end of their vocal tessitura as the first point. The second task was to sing triads, distant from each other by 1 for low voices and 2 for high voices thirds of music in different tempos: *grave*, *andante* and *allegro*. During this exercise

extreme pitches in the range of single triad were distant from main triads' tone by 1 third (200 *cents*) for all singers. Figure 1 shows exactly dependencies between tones during recordings. As the reference tones in this exercise we use three middle notes from first task. Hence we obtained both basic laryngeal mechanisms *M1* and *M2*. In the last exercise singers were asked to sing simple melody, that is useful during singing training process [23].

It was impossible to use a standard method of evaluation, popular in the literature and based on comparison differenced *EGG* signal. From that reason we derived signal and counted error metrics on short, 25 milliseconds frames. Recording sessions took place in the anechoic chamber at University of Science and Technology in Cracow. We registered both acoustic and electroglottographic signals on parallel channels. To achieve this goal we used KayPentax EGG recorder, model 6103 to register EGG signal and G.R.A.S 40 AF high-precision condenser microphone to record acoustic wave graphs. Moreover we used G.R.A.S 12AA 2-channel power module and M-Audio PROFIRE 610 audio interface with Octanes preamplifiers.

3. Data analyzing

One of the main assumptions was to detect pitches in every frames independent. It was due to the necessity of applying it in future authors' surveys. We decided to detect *F0* as the mean value in the frame. It was necessarily to consider wider range of fundamental frequency in singing signals during windowing. We assumed extreme values of studying parameter as 60 *Hz* for the lowest value and 2100 *Hz* for the highest. We used 22ms frame length due to detect the lowest pitches, with 5,5ms overlapping for previous and next window. In order to minimalize zero-crossing coefficient we used low-pass filter with 2100 *Hz* cut-off frequency.

3.1. Statistics for musical measurements

As the musical scale is a logarithmic organization of pitch, grouped in octaves, ratios between two sounds form non-linear scale. It means, that octaves increase exponentially, when we use *Herz* units, as shown on Figure 2. From that reason we decided to use C*ent* as the standard metrics unit. It equalizes spaces between two tones, because it takes into account logarithmic nature of music tones [24]. As the cent we understand:

$$N = 1200 * \log_2(\frac{b}{a}) \quad (1)$$

3.2. Modified error metrics

In order to evaluate quality of analyzing methods, we modified popular error metrics, dedicated for epochs. This action was necessary, as the system with no EGG signal registration will be using in the future.

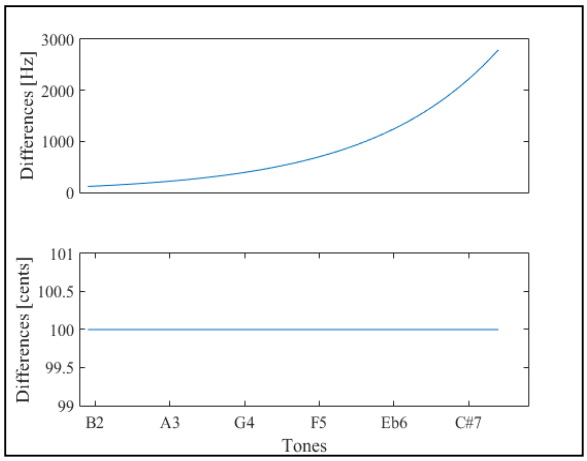

Figure 2 Logaritmic nature of notes in *Herz* and constant differences after transformation to *Cents*

Modified error metrics consists of:

- *Modified Identification Rate* (M*IDR*) – number of frames, with correctly pitch detection – standard *IDR* is the proportion of glottal cycles, with unique glottal cycle detected
- *Modified Miss Rate* (*MMR*) – number of frames, where detected *F0* is lower than expected – standard *MR* is the proportion of detected glottal cycles, for which no glottal cycles detected
- *Modified False Alarm Rate* (*MFAR*) – number of frames, where detected pitch was higher than given value – standard *FAR* is the proportion of detected glottal cycles, where more than one glottal cycle was detected

Additionally we use following group, describing accuracy:

- *Modified F0 Frame Error (MFFE)* – number of frames with improper detected pitches.
- *Modified Standard deviation (MSTD)* – deviation from the mean value of all studied frames in *cents.* An acceptable values are in the range 0-100 *cents* – it corresponds to the least differences between tones in the same octave, as shown on Figure 2.

The error tolerance threshold was described as 100 *cents*, where cent is a logarithmic unit of measure

Table I. Results for voice ranges analyzing.

	ZFF_Modified			*SEDREAMS*			*DYPSA*			*ZBF*		
	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High
MIDR [%]	71,66	91,83	91,20	98,20	97,07	96,58	96,37	71,90	63,22	99,40	97,32	99,02
MMR [%]	24,89	4,39	8,80	0,15	-	3,42	0,84	18,84	31,47	-	-	0,98
MFAR [%]	3,45	3,78	-	1,65	2,93	-	2,79	9,25	5,30	0,60	2,68	-
MFFE [%]	28,34	8,17	8,80	1,80	2,93	3,42	3,63	28,09	36,77	0,60	2,68	0,98
MSTD [*cents*]	373,13	45,83	55,41	14,08	11,50	27,01	68,47	106,46	95,43	15,73	8,91	15,73

intervals in music. That level of threshold minimalizes likelihood of incorrect detection even between two semi tones.

4. Results

We registered *EGG* signals for two purposes. Firstly we used differenced *EGG* to detect *F0* reference in order to choose probes, where set point was deviated from expected value not more than 100 *cents*. Secondly we used it to laryngeal mechanism classification [9] [25]. The analysis was complicated and it was hard to set results together. From that reason we decided to do it in two steps: compare results for tones, classified into voice range: *bass*, *baritone*, *tenor*, *alto*, *soprano* and for laryngeal mechanisms *M1* and *M2*.

4.1. Analysis of voice range

To analyze voice ranges we chose 3 pitches from the middle of singers' tessitura, as shown in Figure 1. We classified probes into three groups: low – *bass* and *baritone*, medium – *tenor* and *alto* and high –*soprano*.

Table II. Results for laryngeal mechanisms analyzing.

	ZFF_Modified		SEDREAMS		DYPSA		ZBF	
	M1	M2.	M1	M2.	M1	M2.	M1	M2.
MIDR [%]	79,50	94,15	98,74	99,24	90,12	46,01	99,58	99,82
MMR [%]	18,09	5,09	0,11	0,76	4,94	45,94	-	-
MFAR [%]	2,42	0,76	1,16	-	4,94	8,05	0,42	0,18
MFFE [%]	20,51	5,85	1,27	0,76	9,88	53,99	0,42	0,18
MSTD [*cents*]	247,25	37,81	13,02	14,45	66,92	229,77	13,39	13,94

4.2. Analysis of laryngeal mechanism

The distance between main tones in Figure 2 depended on type of singing voice. For lower voices – bass, baritone – it was equal to 1 musics' third, which was equivocal 200 *cents*. For higher voices it was equal 1,5 – 2 thirds, which suited 300-400 cents. To analyze laryngeal mechanisms we additionally used probes from down of tessitura for highest voices and top of tessitura for lower voices in order to check generating of *M1* and *M2* registers for this singers. We checked the correctness of classified probes by analyze *EGG* signals, as said before.

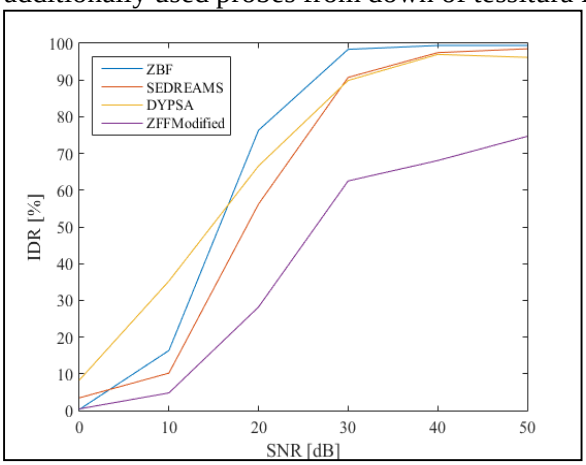

Figure 3 Signal to noise ratio impact on low voices

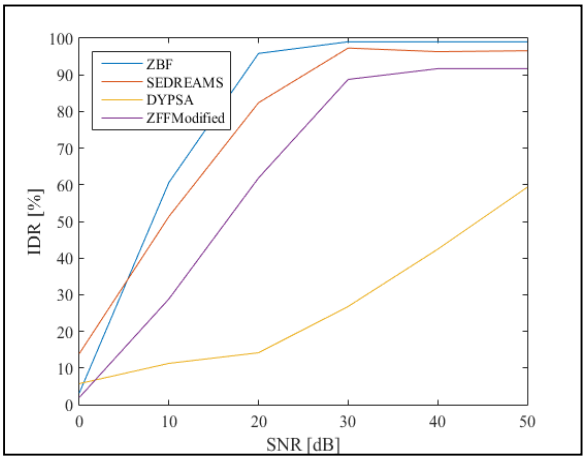

Figure 4 Signal to noise ratio impact on high voices

4.3. Signal to noise ratio

The last part of the experiment was to study impact of noising signals on error metrics for all methods. To achieve this aim, we decided to add white noise to analyzing probes with increasing *signal to noise ratio SNR*. We analyzed it in three steps, the same as in the voice range analysis. We expected different behavior of particular methods, when signal is more noisy. Results of that experiment was described in this section.

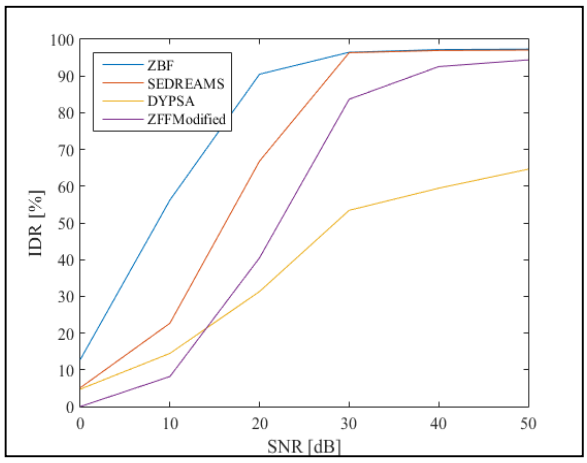

Figure 5 Signal to noise ratio impact on medium voices

4.4. Discussion

From registered results we can conclude that there is a relationship between type of singers, laryngeal configuration and observed *F0* for all analyzed methods. *ZFF_Modified* method can be used for high and medium voices with satisfying accuracy. Its' results for *M2* vocal production process also presents high percentage of correctly detection. Worth attention fact is modified standard deviation metric for this method. It presents acceptable values for high and medium voices and for *M2* mechanism. Unfortunately, this method is unsuitable for low voices and chest registers (*M1*). The opposite to this are results for *DYPSA* algorithm. It can be apply to analyzing *bass* and *baritone* vocal tracts but is not able to detect pitches for *sopranos* and medium voices. As can be expected this method detects *M1* mechanism in the proper way, because it is natural for low voices and shows high errors percentage and *MSTD* for *M2*. Both *SEDREAMS* and *ZBF* methods do not rely on voice tessitura and glottis configuration. The *MIDR* factor is stable and more than 95% in all cases. There is no significant differences in *MSTD* parameter results for these two methods. Both present maximum deviations from the mean value, which are lower than distance between two neighboring notes in the octave. Hence we conclude that there is very little probability to classify analyzing pitch as another one in the octave. We observed differences between counting time in all cases but did not measure it. This fact should be study in the future analyzes.

Analyzing *SNR* impact on methods' results proofed that the most robust one is *ZBF*, as shown on Figures 3-5. We can detect *F0* even, when the signal is noisy, with 20 *decibels SNR*. This fact is important during singing analyzes with the music background.

Almost all analyzed methods require *a priori* knowledge about expected *F0*. Only *ZBF* and *DYPSA* do not require this activity. This is undoubtedly great advantage of this methods during comparison all together. It is significant, when there is lack of knowledge about analyzing signal and singers' tessitura.

5. Conclusions

The scope of this paper was to analyze possibility of applying pitch detection methods, elaborated in the last few years to singers' vocal tracts surveys. We compered four techniques on a created for this purpose dataset including wide range of fundamental frequency and two singing techniques. We obtained that the most robust algorithm to this task is *Zero Band Filtering*, as the most usable for further surveys, focused on studying only acoustical signals.

References

[1] B. Doval, C. D'Alessandro, N. Henrich: The Spectrum of Glottal Flow Models. *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 1026-1041, 2006.

[2] M. Rothenberg: A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *The Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1632-1645, 1973.

[3] M. Airas: An environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocology, vol. 33, no. 1, pp. 49-64, 2009.

[4] T. Raito, A. Suni: HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 1, pp. 153-165, 2011.

[5] J. Makhoul: Linear prediction: A tutorial review. Proceedings of the IEEE, vol. 63, no. 4, pp. 561-580, 1975.

[6] M. Kob, N. Henrich: Analysing and Understanding the Singing Voice: Recent Progress and Open Questions. Current Bioinformatics, vol. 6, no. 3, pp. 362-374, 2011.

[7] J. Sundberg: The acoustic of singing voice. Scientific America, vol. 236, pp. 82-91, 1977.

[8] O. Babacan T. Drugman, T. Raito, D. Erro, T. Rutoit: Parametric representation for singing voice synthesis: A comparative evaluation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014.

[9] B. Roubeau, N. Henrich M. Castellengo: Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited. Journal of Voice, vol. 23, no. 4, pp. 425-438, 2009.

[10] O. Babacan, T. Drugman, N. D'Alessandro, N. Henrich, T. Dutoit: A comparative study of pitch extraction algorithms on a large variety of singing sounds. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, 2013.

[11] I. Pollac, J. M. Picket, W. H. Sumby: On the Identification of Speakers by Voice. The Journal of the Acoustical Society of America, vol. 26, no. 3, pp. 403-406, 1954.

[12] G. David: "Pitch Extraction and Fundamental Frequency: History and Current Techniques. Department of Computer Science University of Regina, Regina, Saskatchewan, 2003.

[13] W. Wszolek, M. Kłaczyński: Comparative study of the selected methods of laryngeal tone determination. Archives of Acoustics, vol. 31, no. 4, pp. 219-226, 2006.

[14] K. Deepak, S. Prasanna: Epoch Extraction Using Zero Band Filtering from Speech Signal. Circuits, Systems, and Signal Processing, vol. 34, no. 7, pp. 2309-2333, 2015.

[15] B. Yegnanarayana, S. Prasanna, S. Guruprasa: Study of robustness of zero frequency resonator method for extraction of fundamental frequency. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, 2011.

[16] S. Srinivas, K. Prahallad: An FIR Implementation of Zero Frequency Filtering of Speech Signals. IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 9, pp. 2613-2617, 2012.

[17] K. Murty, R. Sri, B.Yegnanarayana: Epoch Extraction From Speech Signals. IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 8, pp. 1602-1613, 2008.

[18] S. Kadiri, B. Yegnanarayana: Analysis of singing voice for epoch extraction using Zero Frequency Filtering method. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, 2015.

[19] B. Yegnanarayana, K. Murty: Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals. IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 4, pp. 614-624, 2009.

[20] P. A Naylor., A. Kounoudes: 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.

[21] T. Drugman: Advances in glottal analysis and its applic ations. PhD thesis, 2011.

[22] O. Babacan, T. Drugman, N. D'Alessandro, N. Henrich, T. Dutoit: A Quantitative Comparison of Glottal Closure Instant Estimation Algorithms on a Large Variety of Singing Sounds. 2013.

[23] J. K. Lasocki: Mały Solfeż. PWM, 2005.

[24] R.J. McNab, A Smith Lloyd., I. H. Witten, C. L. Henderson: Towards the digital music library: tune retrieval from acoustic input. Proceedings of the first ACM international conference on Digital libraries, New York, 1996.

[25] E. B. Lacerda, C. Mello: Automatic classification of laryngeal mechanisms in singing based on the audio signal. Procedia Computer Science, vol. 112, pp. 2204-2212, 2017.