A Quantitative Comparison of Glottal Closure Instant Estimation Algorithms on a Large Variety of Singing Sounds

Onur Babacan[1], *Thomas Drugman*[1], *Nicolas d'Alessandro*[1], *Nathalie Henrich*[2],*Thierry Dutoit*[1]

[1]Circuit Theory and Signal Processing Laboratory, University of Mons, Belgium
[2]Speech and Cognition Department, GIPSA-lab, Grenoble, France

onur.babacan@umons.ac.be, thomas.drugman@umons.ac.be, nda@numediart.org,
nathalie.henrich@gipsa-lab.grenoble-inp.fr, thierry.dutoit@umons.ac.be

Abstract

Glottal closure instant (GCI) estimation is a well-studied topic that plays a critical role in several speech processing applications. Many GCI estimation algorithms have been proposed in the literature and shown to provide excellent results on the speech signal. Nonetheless the efficiency of these algorithms for the analysis of the singing voice is still unknown. The goal of this paper is to assess the performance of existing GCI estimation methods on the singing voice with a quantitative comparison. A second goal is to provide a starting point for the adaptation of these algorithms to the singing voice by identifying weaknesses and strengths under different conditions. This study is carried out on a large database of singing sounds with synchronous electroglottography (EGG) recordings, containing a variety of singer categories and singing techniques. The evaluated algorithms are Dynamic Programming Phase Slope Algorithm (DYPSA), Hilbert Envelope-based detection (HE), Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS), Yet Another GCI Algorithm (YAGA) and Zero Frequency Resonator-based method (ZFR). The algorithms are evaluated in terms of both reliability and accuracy, over different singing categories, laryngeal mechanisms, and voice qualities. Additionally, the robustness of the algorithms to reverberation is analyzed.

Index Terms: singing analysis/synthesis, GCI estimation, glottal closure instant

1. Introduction

The field of speech processing has seen a leaping development over the last decades, creating a variety of techniques for the analysis, modeling and synthesis of speech signals. The related field of singing processing has also developed, but the extent reached by the breadth and depth of the speech processing techniques has not been directly reflected in it. While both singing and speech is achieved by the same vocal apparatus, applying the speech approaches to singing may not be straightforward [1]. Some key differences of singing from speech are the wider pitch range, more controlled pitch variations, greater dynamic range and prolonged voiced sounds. Source-filter interaction also has more of an impact on singing and cannot be neglected as easily. Additionally, the large variety of singing techniques and phenomena has made it more difficult to formalize and generalize the singing voice.

As a consequence, existing singing synthesizers have limited their scope, generally focusing on one singer category or one singing technique. These limitations create a wide gap between the synthesizers and the expressive range of human singers, as well as the performative requirements of musicians wishing to use these tools.

Among existing systems, Harmonic plus Noise Modeling (HNM) has been used extensively [2]. In the SMS [3] and Vocaloid [4] systems, a degree of control is obtained over a unit concatenation technique by integrating HNM [5], though the synthesis results are still confined in singing space to the range of the pre-recorded samples. In the CHANT [6] and FOF [7] systems, rule-based descriptions characterizing some operatic voices are integrated, yielding remarkable results for soprano voices. Meron obtained convincing results for lower registers of singing by applying the non-uniform unit selection technique to singing synthesis [8]. Similar strategies have been applied to formant synthesis, articulatory synthesis [9] and HMM-based synthesis methods [10], but the limitations in vocal expression range have been quite similarly limited.

In this study, we continue our efforts to establish a foundation for an analysis framework targeting a wide range of singing techniques and singer categories, with the long-term aim of a wide-range synthesizer. We build upon our previous work on pitch analysis for singing [11]. The Glottal Closure Instants (GCIs) are significant excitations in the voice and many analysis, synthesis, modeling and decomposition algorithms rely on the accurate estimation of their location, including a variety of methods that employ Time-Domain Pitch-Synchronous Overlap Add (TD-PSOLA) analysis/resynthesis [12][13], mixed-phase decomposition [14] and Deterministic Plus Stochastic Model (DSM) [15]. While the behaviors of GCI estimation techniques are very well studied and understood on speech signals [16], they are still unkown on the singing voice. The aim of this study is to investigate the performance of existing GCI estimation techniques on a wide variety of singing signals and to identify strengths and weaknesses for particular conditions. This investigation is also intended to serve as a starting point for possible contributions to the field by identifying challenges to overcome.

The structure of the paper is as follows. Section 2 briefly describes the GCI estimators used in this study. Section 3 presents the singing database and details the experimental protocol, including establishment of the ground truth and definitions of the error metrics used. The results of the experiments in different groups and under different conditions are presented in detail in Section 4, along with their discussion. Lastly, Section 5 draws the conclusions of the study.

2. Methods for GCI Estimation

This section gives a brief overview of the state-of-the-art GCI estimation methods compared in this study:

- The Dynamic Programming Phase Slope Algorithm (DYPSA) [17] estimates GCIs by identifying peaks in the linear prediction (LP) [18] residual signal of speech (obtained by removing an auto-regressive model of the spectral envelope and whitening the resultant signal). It first generates GCI candidates using the group delay function of the LP residual signal, then selects the subset of GCI estimates using N-best dynamic programming.
- There are several methods in the literature that employ the Hilbert envelope (HE) [19]. In this study, a method based on the HE of the LP residual signal is evaluated. The HE exhibits large positive peaks when the residue presents discontinuities. The method employs a Center of Gravity (CoG)-based signal in conjunction with the HE to make GCI estimates as described in [16].
- Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) [20] is a recently proposed algorithm that first determines intervals of GCI presence using a mean-based signal (obtained by calculating the mean of a sliding window over the speech signal), then refines the GCI locations using the LP residual.
- Yet Another GCI Algorithm (YAGA) [21], similar to DYPSA, is an LP-based approach that uses N-best dynamic programming to select from GCI candidates. The GCI candidates are generated from an estimate of the voice source signal (time-derivative of glottal volume flow rate) instead of the residual signal.
- The Zero Frequency Resonator (ZFR)-based technique is rooted in the observation that the impulsive excitation at GCIs have an effect across all frequencies [22]. This method uses zero frequency resonators to minimize the influence of vocal tract resonances and isolate the excitation pulses.

Since several of these GCI estimation techniques are sensitive to an inversion of the signal, the polarity was detected and corrected using the RESKEW method [23].

3. Experimental Protocol

3.1. Database

For this study, the scope was constrained to vowels. Samples with verified reference pitch trajectories from our previous study were used [11]. Samples from different singers were taken from the LYRICS database recorded by [24, 25], for a total of 13 trained singers. The selection consisted of 7 bass-baritones (B1 to B7), 3 countertenors (CT1 to CT3), and 3 sopranos (S1 to S3). The recording sessions took place in a soundproof booth. Acoustic and electroglottographic signals were recorded simultaneously on the two channels of a DAT recorder. The acoustic signal was recorded using a condenser microphone (Brüel & Kjær 4165) placed 50 cm from the singer's mouth, a preamplifier (Brüel & Kjær 2669), and a conditioning amplifier (Brüel & Kjær NEXUS 2690). The electroglottographic signal was recorded using a two-channel electroglottograph (EG2, [26]). The selected samples contain a variety of singing tasks, such as sustained vowels, crescendos-decrescendos and arpeggios, and ascending and descending glissandos. Whenever possible, the singers were asked to sing in both laryngeal mechanisms M1 and M2 [27, 28]. Laryngeal mechanisms M1 and M2 are two biomechanical configurations of the laryngeal vibrator commonly used in speech and singing by both male and females. Basses and baritones mainly sing in M1. They may use M2 to sing high pitches. Countertenors commonly sing in M2. They may use M1 for artistic purposes, so their singing technique requires the ability to sing with similar voice qualities in both laryngeal mechanisms. Sopranos mainly sing in M2. They can choose to sing in M1 in the low to medium part of their tessitura.

3.2. Ground Truth

The ground truth is based on the differentiated EGG signal (dEGG). GCI marking was done on the dEGG signals by simple peak detection above an empirically-determined amplitude. This method yielded very accurate results. The same attempt was also made with the SIGMA [29] algorithm, but the results were much less accurate. The reliability of the ground truth was assessed by visual comparisons between the dEGG signals and the GCI marked positions. Due to time constraints, manual correction of reference GCIs was not possible, but the errors were manually counted. While errors of missing GCIs and false alarms existed in these marks, a practical acceptance threshold of 1% total error was set in order to strike a balance between the reliability and the scope of the ground truth. The mean total error across the 437 accepted singing samples is 0.14%, with 146 of them containing no errors.

3.3. Synchronization of EGG Recordings with Audio

Electroglottography is a non-invasive method for recording vocal-fold contact area by placing two contact electrodes on the singer's neck. While practically providing the true positions of GCIs, the EGG signal needs to be time-aligned with the audio signal, which is delayed by the duration of sound propagation between glottis and external microphone. Accuracy of GCI estimation depends on the synchrony of both signals. In this study, time-alignment was done separately for each individiual recording, since the mouth-to-microphone distance can be highly variable, even inside the same corpus. For each recording, a time delay between GCI estimates and nearest corresponding reference GCIs was calculated for all algorithms. The time-alignment lag was then given by the mode of the collection of all temporal distances. This method was employed after previous synchronization attempts, which used the cross-correlation of dEGG signals and LP residues, yielded highly inaccurate results.

3.4. Error Metrics

A group of standard error metrics were used to evaluate the performance of the algorithms both in reliability and accuracy by comparing their GCI estimates to the time-aligned reference GCIs. The first group, describing reliability, consists of:

- Identification Rate (IDR): the proportion of glottal cycles for which a unique GCI is detected,
- Miss Rate (MR): the proportion of glottal cycles for which no GCI is detected,
- False Alarm Rate (FAR): the proportion of glottal cycles for which more than one GCI is detected.

For each correctly-detected GCI (i.e. satisfying the IDR condition), a timing error is calculated.

The second group, describing accuracy, is derived from the distribution of these timing errors:

- Identification Accuracy (IDA): the standard deviation of the timing error distribution,

Table 1: *Performance of GCI estimation algorithms by singer type and laryngeal mechanism*

Singer Type	Algorithm	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to $\pm$ 0.25 ms.(%)
Baritone	DYPSA	98.67	0.85	0.47	0.19	67.31
	HE	97.41	1.57	1.02	0.26	10.83
	SEDREAMS	97.19	0.81	2.00	0.33	51.38
	YAGA	91.46	0.59	7.95	0.22	73.31
	ZFR	29.63	0.14	70.23	0.48	1.10
Counter-Tenor	DYPSA	87.83	11.17	0.99	0.22	76.98
	HE	89.41	8.10	2.49	0.28	30.12
	SEDREAMS	96.86	1.04	2.11	0.32	61.93
	YAGA	94.84	3.12	2.04	0.20	84.99
	ZFR	72.34	1.20	26.46	0.30	7.07
Soprano	DYPSA	64.30	35.50	0.20	0.31	65.29
	HE	77.47	21.85	0.68	0.35	42.70
	SEDREAMS	95.09	3.37	1.54	0.34	62.46
	YAGA	89.52	9.68	0.80	0.33	59.42
	ZFR	75.30	2.36	22.34	0.42	15.75
Laryngeal Mechanism	Algorithm	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to $\pm$ 0.25 ms. (%)
M1	DYPSA	98.33	0.65	1.01	0.17	71.78
	HE	97.96	0.79	1.25	0.25	11.26
	SEDREAMS	97.73	0.56	1.71	0.32	54.27
	YAGA	91.48	0.47	8.04	0.19	78.27
	ZFR	30.84	0.11	69.05	0.47	1.11
M2	DYPSA	74.03	25.76	0.21	0.29	67.91
	HE	83.05	15.93	1.02	0.33	38.03
	SEDREAMS	96.49	2.45	1.06	0.34	60.88
	YAGA	92.27	7.02	0.71	0.29	69.40
	ZFR	79.18	1.83	18.99	0.38	12.04

- Accuracy to $\pm$0.25 ms: the proportion of detections for which the timing error is smaller than a threshold of 0.25 ms in either direction.

It is worth noting for clarity that a high-accuracy algorithm generates lower values for the former and higher values for the latter metric. The threshold in the latter ($\pm$0.25 ms) is chosen very small since techniques that rely on GCI locations are very sensitive to errors therein.

4. Results

To examine the effect of different dimensions of singing, the error metrics were calculated for several different subsets of the database. The obtained results are presented in this section.

4.1. The Effect of Singer Type

As can be observed in Table 1 and Figure 1, SEDREAMS and YAGA are the only algorithms that can be said to perform reliably in a consistent manner across different singer types. DYPSA and HE perform better for lower-pitch singers, and a clear trend of decreasing reliability is apparent as pitch increases. ZFR's high false alarm rates prohibit any practical usage. Overall, SEDREAMS offers the best reliability with consistency and lowest error rates. In terms of accuracy, none of the algorithms reach the level of performance that would be expected in speech. YAGA performs the best for baritones and counter-tenors, with DYPSA being a close second. For sopranos, SEDREAMS surpasses YAGA in accuracy. Considering the reliability and accuracy in conjunction, DYPSA and YAGA are good choices for baritones, latter of which is also the best choice for counter-tenors. Noting the low IDR of DYPSA for sopranos, SEDREAMS is the most suitable in that group. Additionally, for baritones, a trade-off between reliability and accuracy observed when choosing between DYPSA and YAGA.

A point worth noting is that even the most accurate methods do not perform nearly as well as they do on speech signals[16]. One of the reasons is that the discontinuities are less significant in the LP residual or glottal source estimate of singing signals, especially for higher-pitch voices. A potential improvement can be obtained by using an excitation signal that presents discontinuities in a more pronounced way, such as [30]. Another possible explanation is the source-filter interactions that are unaccounted for. All the techniques are based on the linear source-filter assumption, which has proved to be a good first approximation in normal speech. In singing, however, the linear source-filter theory is not as suitable, especially at high pitches. LP analysis works at low pitches but fails to estimate accurate inverse-filtered glottal flow at high pitches, which may account for the observed degradation in reliability and accuracy as vocal range gets higher.

4.2. The Effect of Laryngeal Mechanism

Similar trends as in singer types is observed for laryngeal mechanisms. In terms of reliability, SEDREAMS is the clear best performer, with YAGA providing consistent results but being outperformed by DYPSA and HE for M1. In accuracy, YAGA provides the best results, with DYPSA being comparable. It is worth noting that an overall accuracy increase is observed from

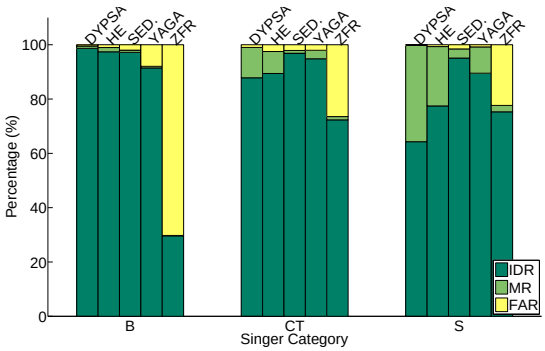

Figure 1: *Effect of singer type on reliability. IDR, MR and FAR values from bottom to top in each stack.*

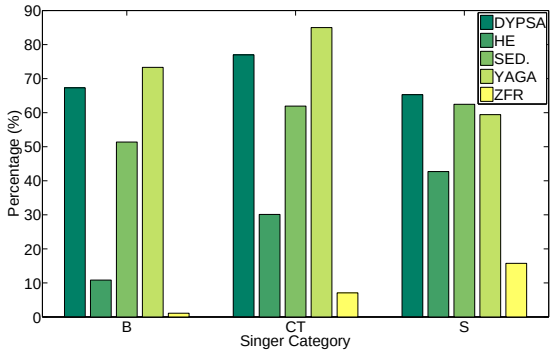

Figure 2: *Effect of singer type on accuracy to ±0.25 ms.*

M1 to M2. This could be attributed to the fact that as pitch increases, glottal cycles become shorter, and for the same level of relative timing error, the accuracy will increase.

4.3. The Effect of Reverberation

In many concrete cases, singers perform within large rooms or halls, where the microphone might capture replicas of the voice sound caused by reflections from the surrounding surfaces. To simulate such reverberant conditions, we considered the L-tap Room Impulse Response (RIR) of the acoustic channel between the source to the microphone. RIRs are characterized by the value T_{60}, defined as the time for the amplitude of the RIR to decay to -60dB of its initial value. A room measuring 3x4x5 m and T_{60} ranging $\{100, 200, \ldots, 500\}$ ms was simulated using the source-image method [31] and the simulated impulse responses convolved with the clean audio signals.

The performance of the algorithms under simulated reverberant conditions are presented in Table 2 for the whole experimental database. DYPSA is excluded from this experiment because the large majority of recordings caused software errors in the Voicebox [32] implementation used in our experiments. The results from the remaining algorithms are presented in Table 2

Taking both reliability and accuracy into account, SEDREAMS is the most robust algorithm among the tested. Similarly, HE provides good reliability but the accuracy is significantly lower in comparison. While YAGA is comparably reliable, its accuracy is very sensitive to reverberation, and makes it a less desirable choice along with the worst performer, ZFR.

Table 2: *Performance of GCI estimation algorithms with increasing reverb levels (RL)*

IDR (%)					
Algorithm / T_{60} (ms.)	100	200	300	400	500
HE	90.99	90.66	90.58	90.55	90.58
SEDREAMS	90.81	91.81	90.74	90.37	89.68
YAGA	86.02	83.31	81.69	80.20	79.10
ZFR	43.80	43.01	41.89	41.09	40.62
Accuracy to ±0.25ms.					
Algorithm / T_{60} (ms.)	100	200	300	400	500
HE	25.63	28.07	28.97	28.71	29.58
SEDREAMS	52.52	49.03	47.34	45.27	43.52
YAGA	49.24	32.23	27.21	25.27	24.23
ZFR	16.02	13.88	15.05	15.91	18.02

5. Conclusion

In this study, we provided a comparative evaluation of GCI estimation methods as an effort toward establishing the foundation for the development of efficient wide-range singing synthesis algorithms. This problem has been studied extensively for the speech signal, and we aimed to answer the question of what the best method for estimating GCIs of the singing voice is. Five of the most representative state-of-the-art techniques were evaluated on a large dataset containing a rich variety of singing techniques. As we expected, the question does not have a single answer, and the best choice largely depends on the pitch range of the singing, as well as the requirements of the target application. The robustness of the GCI estimation methods to reverberation was also evaluated. Here, SEDREAMS was clearly the most robust. Another aim of this paper was to identify areas of improvement specific to the singing voice. Our results display that there is a clear lack and need of high-accuracy methods, and helps display the accuracy gap between speech and singing quantitatively.

6. Acknowledgements

Onur Babacan is supported by a PhD grant funded by UMONS and Acapela Group.

7. References

[1] M. Kob, N. Henrich, H. Herzel, D. Howard, I. Tokuda, and J. Wolfe, "Analysing and understanding the singing voice: Recent progress and open questions," *Current Bioinformatics*, vol. 6, no. 3, pp. 362–374, 2011.

[2] Y. Stylianou, "Modeling speech based on harmonic plus noise models," in *Nonlinear Speech Modeling and Applications*, 2005, pp. 244–260.

[3] X. Serra and J. O. Smith, "Spectral modeling synthesis: a sound analysis/synthesis based on a deterministic plus stochastic decomposition," vol. 14, p. 1224, 1990.

[4] "Vocaloid," http://www.vocaloid.com, [Online; accessed 12-December-2012].

[5] J. Bonada, O. Celma, A. Loscos, J. Ortol, X. Serra, Y. Yoshioka, H. Kayama, Y. Hisaminato, and H. Kenmochi, "Singing voice synthesis combining excitation plus resonance and sinusoidal plus residual models," in *Proceedings of International Computer Music Conference*, 2001.

[6] X. Rodet, Y. Potard, and J. B. Barriere, "The CHANT project: From the synthesis of the singing voice to synthesis in general," *Computer Music Journal*, vol. 8, no. 3, pp. 15–31, 1984.

[7] X. Rodet, "Time-domain formant wave function synthesis," vol. 8, pp. 9–14, 1984.

[8] Y. Meron, "Synthesis of vibrato singing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2000, pp. 745–748.

[9] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 873–876.

[10] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "HMM-based singing voice synthesis system," in *Proceedings of Interspeech*, 2006, pp. 1141–1144.

[11] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A Comparative Study of Pitch Extraction Algorithms on a Large Variety of SingingSounds," in *ICASSP 2013, accepted for publication*.

[12] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, Dec. 1990. [Online]. Available: http://dx.doi.org/10.1016/0167-6393(90)90021-Z

[13] T. Dutoit and H. Leich, "Mbr-psola: Text-to-speech synthesis based on an mbe re-synthesis of the segments database," *Speech Communication*, vol. 13, no. 3, pp. 435–440, 1993.

[14] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[15] T. Drugman and T. Dutoit, "The Deterministic Plus Stochastic Model of the Residual Signal and Its Applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 968–981, 2012.

[16] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 994–1006, 2012.

[17] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 34–43, 2007.

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

[19] S. Mahadeva Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[20] T. Drugman, "Advances in glottal analysis and its applications," Ph.D. dissertation, Ph. D. thesis, University of Mons, 2011.

[21] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 82–91, 2012.

[22] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1602–1613, 2008.

[23] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," *Signal Processing Letters, IEEE*, vol. 20, no. 4, pp. 387–390, 2013.

[24] N. Henrich, "Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception [study of the glottal source in speech and singing: Modeling and estimation, acoustic and electroglottographic measurements, perception]," Ph.D. dissertation, Université Paris 6, 2001.

[25] N. Henrich, C. d'Alessandro, M. Castellengo, and B. Doval, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1417–1430, 2005.

[26] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, pp. 36–43, 1992.

[27] N. Henrich, "Mirroring the voice from garcia to the present day: Some insights into singing voice registers," *Logopedics Phoniatrics Vocology*, vol. 31, pp. 3–14, 2006.

[28] B. Roubeau, N. Henrich, and M. Castellengo, "Laryngeal vibratory mechanisms: the notion of vocal register revisited," *J. Voice*, vol. 23, no. 4, pp. 425–438, 2009.

[29] M. R. Thomas and P. A. Naylor, "The sigma algorithm for estimation of reference-quality glottal closure instants from electroglottograph signals," in *Proc. European Signal Processing Conf.(EUSIPCO), Lausanne, Switzerland*, 2008.

[30] E. Candes and J. Romberg, "l1-magic: Recovery of sparse signals via convex programming," 2005.

[31] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[32] "VOICEBOX Matlab Toolbox," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, [Online; accessed 24-March-2013].