INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Today's most frequently used F_0 estimation methods, and their accuracy in estimating male and female pitch in clean speech

Sofia Strömbergsson

Division of Speech and Language Pathology, Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet (KI), Stockholm, Sweden
sofia.strombergsson@ki.se

Abstract

Variation in fundamental frequency (F_0) constitutes a valuable source of information for researches across many disciplines, with a shared interest in speech. Different methods for estimating F_0 vary in estimation accuracy and accessibility, and there is yet no gold standard. Through a bibliometric survey, this study examines what methods were the most frequently used in the speech scientific community during the years 2010-2016. Secondly, the most used methods are evaluated against a ground truth reference, with a specific focus on their accuracy in estimating F_0 in male and female speakers, respectively.

The results show that Praat is the dominant method by far, followed by STRAIGHT, RAPT and YIN. This pattern holds across a range of different research areas, although within Acoustics and Engineering, Praat's dominance is less pronounced. In the evaluation including Praat, RAPT and YIN – with their default and gender-adapted settings – Praat also proved to be the most accurate. The finding that adapting Praat's pitch range settings by gender leads to further improvements should encourage researchers to do this routinely.

Index Terms: fundamental frequency, pitch tracking, pitch estimation, speech prosody, intonation.

1. Introduction

Estimation of the fundamental frequency (F_0), or pitch tracking, underlies all data-driven analyses of intonation. Although large datasets require automatic procedures, there are known uncertainties involved in using existing methods, raising reliability as an obvious concern. However, although different F_0 estimation methods are in use, there is no gold standard. This investigation sets out to survey which F_0 estimation methods are currently the most well-spread, and to bench-mark these against a ground truth reference.

Information regarding F_0 in speech is relevant to a range of different research areas and applications. In basic linguistic research, F_0 characteristics have been linked to various linguistic and pragmatic functions, e.g. conversational contrast [1] or interrogative signaling [2]. In dialogue systems, real-time analysis of the F_0 contour may guide the interpretation of a speaker's intentions [3]. In a clinical context, F_0 is an important feature in the description of atypical prosodic and vocal features [4]. Within the field of acoustics, new pitch tracking algorithms are continuously suggested, for clean recordings of single speakers as well as for more challenging conditions. However, it is unclear to what extent these algorithms are accessible to researchers in other disciplines.

Many approaches exist for estimating F_0. In phonetic and linguistic research, Praat [5] has been referred to as "the de facto standard speech analysis program" [6], and is indeed a well-spread software for the acoustic analysis of speech. For F_0 estimation, the default method provided in Praat is the autocorrelation function [5]. Another flavor of the autocorrelation method is available in YIN [7], which has been referred to as "one of the most popular and most efficient methods of pitch estimation" [8]. The RAPT [9] algorithm (also referred to as ESPS or get_f0) is another well-known method, which instead uses a cross-correlation function. In most, if not all, speech analysis frameworks, users can adjust parameters like framerate and pitch range, to tailor their analysis to their purpose. F_0 estimation accuracy varies across speakers, with accuracy generally being lower for female speakers than for male speakers [10]. As shown by Vogel and colleagues [11], adjusting the pitch-range settings by gender may improve F_0 estimation accuracy, even to levels comparable to applying individualized speaker settings. However, little is known of whether such adjustments are generally made, and of the degree of improvement across different F_0 estimation approaches.

Many evaluations of F_0 estimation methods have been presented during the years, for conditions involving different challenges (e.g. clean speech [10] [12], noisy conditions [8], multi-speaker conditions [13], and for singing voice [14]). According to Pirker et al. [15], YIN and RAPT are the best performing algorithms for F_0 estimation in single speakers. Camacho [10] substantiated this statement, in finding YIN and RAPT performing at comparable levels, whereas Praat's autocorrelation method performed slightly worse. According to a more recent evaluation by Ghahremani and colleagues [16], many off-the-shelf pitch trackers (YIN and RAPT included) are outperformed by Kaldi Pitch [16], an algorithm specifically tuned for automatic speech recognition. Such differences in evaluation outcomes may be contributed to different methodological choices. For example, whereas some evaluations base their statistics on instants where all candidate trackers agree on voicing (e.g. [10] [16]), the voiced/unvoiced decision is treated like a feature of the candidate trackers in other evaluations (e.g. [8] [17]). Hence, comparing outcomes across evaluations is not always straightforward. Additional concern may be raised regarding the fact that not all evaluations include the same trackers; this indicates that the selection of trackers is somewhat arbitrary.

The accuracy of estimated glottal activity (as reflected in F_0 frequency) is best evaluated with reference to recorded actual glottal activity, e.g. as described in [18]. The CSTR [18], the Mocha-TIMIT [19] and the Keele [20] corpora are well-known resources containing parallel recordings of laryngograph and microphone signals. A more recent database is the PTDB-TUG

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-240

corpus [15], which includes more speakers and more recorded sentences than previous resources. For a general purpose evaluation of F_0 estimation approaches, PTDB-TUG therefore appears as the best available reference today.

Without quantitative evidence, impressions that some F_0 estimation approaches are more widely used than others remain subjective. Although evaluations of existing methods have been made, the selection of included approaches has not necessarily reflected their usage in the scientific community. Moreover, validations against ground truth references of smaller size may need to be updated when better reference materials become available. And finally, although evidence holds that adapting pitch range settings by speaker gender may improve F_0 estimation accuracy, it is yet unknown to what extent such adjustments are actually made, and to what gain. The present paper addresses these concerns by investigating the following research questions:

1. What F_0 estimation methods have been used the most within the wide speech scientific community, during the last 5 years?
2. How do these methods perform, when used "as-is", and evaluated against the best available ground truth reference?
3. For each of the methods evaluated, how is estimation accuracy affected by adapting settings by speaker gender?

A bibliometric survey addressing the first research question is described in Section 2. Based on this survey, the most frequently used approaches are then evaluated against a ground truth reference; this evaluation is described in Section 3. Lastly, the findings are summarized and discussed in Section 4.

2. Bibliometric survey

2.1. Data retrieval

The first research question, relating to what F_0 estimation methods are currently in use, was explored through a bibliometric survey using the Web of Science (WoS), within its Core Collection. The search terms used to extract publications (articles or proceedings papers) were a) any of the terms *pitch*, *intonation*, *fundamental frequency* or *f0*, together with any of the terms b) *track**, *estimat**, *curve*, *contour*, *slope*, *rising* or *falling*. Any one of the two terms *speech* or *voice* was also included as a required term. (For the *speech* alternative, the string *prosod** was specified as a required term.) Research areas included were Acoustics, Computer Science, Communication and Linguistics[1]. The search was performed in the Web of Science on March 20, 2016, and included publications between 2010 and 2016. Based on these criteria, the search resulted in 360 hits. After the manual exclusion of publications describing non-human sounds (e.g. animals or musical instruments), 351 items remained.

2.2. Data analysis

The 351 publications were manually examined with regards to whether they included any F_0 estimation, and – if so – what specific method was used. If F_0 values were referred to in the publication without any explicit specification regarding how these were derived, this was also noted. F_0 estimation settings were noted if such were explicitly stated.

Information regarding WoS research area classification was registered for all publications. In this analysis, many publications were counted multiple times, reflecting the fact that publications may represent more than one research area.

2.3. Bibliometric results

74 of the retrieved 351 publications were not accessible online, neither through the author's university library, nor through a Google Scholar search. Of the remaining 277 publications, 128 contained no information regarding F_0 estimation. Of these, a substantial proportion constituted reports of F_0 modelling or re-synthesis (n = 53, e.g. [21], [22]), thus implicitly involving analysis of F_0. In these, Hz values were often reported but with no explicit information regarding how these were derived.[2] A minor proportion (n = 25) were publications that neither directly nor indirectly involved any F_0 analysis, e.g. in reviews like [23], or in studies based on existing ToBI-annotated data, e.g. [24].

Six publications included more than one F_0 estimation algorithm [6] [14] [25]-[28]. For ease of interpretation, only the 143 publications where one F_0 estimation method was specified are included in the presentation in Table 1. As seen, in the majority of the publications where one F_0 estimation method was specified, this method is Praat [5]. Of the 11 cases where information regarding the specific estimation algorithm was unavailable, but where software environment was specified, 3 had employed Snack [29], 2 had used CSL[3], whereas the remaining 6 had all been performed in different environments. In 17 of the 143 publications, the output from the automatic F_0 estimation was checked (and potentially modified) manually. In 9 of the 80 publications based on F_0 estimation in Praat, the default pitch range setting (75-500 Hz) was modified across all speakers (n = 5), adapted by gender (n = 2), or by speaker (n = 2).

Table 1. *Methods used in the 143 publications where one F_0 extraction algorithm was specified.*

Extraction method	# of publications
Praat	80
RAPT	12
N/A*	11
STRAIGHT [30]	8
YIN	5
SWIPE [31]	3
Hu & Wang [32]	2
Others**	22
Total	143

* Information regarding the F_0 analysis environment was specified, whereas the specific method was not.
** Each appearing in only one publication.

The 143 publications where one F_0 estimation method was specified were also analyzed with regards to what research area

[1] In WoS, publications on phonetics (e.g. Journal of Phonetics) are listed under the research area Linguistics.

[2] In some of these publications, the authors referred to another publication for details regarding how F_0 data had been derived. Such second-hand references were not included in this analysis.
[3] Computerized Speech Lab, PENTAX.

they represented. In this analysis, 7 areas represented by only one publication each were collapsed into a single category: “Other areas”. As indicated in Figure 1, Praat is the dominant method (> 60% of the publications) in most research areas. Only within Engineering and Acoustics, Praat’s dominance is less pronounced, falling below 50%. (NB: Figure 1 illustrates that, for example, YIN is represented in 7 research areas, which may seem contradictory to the data in Table 1, where YIN is reported in only 5 publications. However, this reflects the fact that a publication may represent more than one research area.)

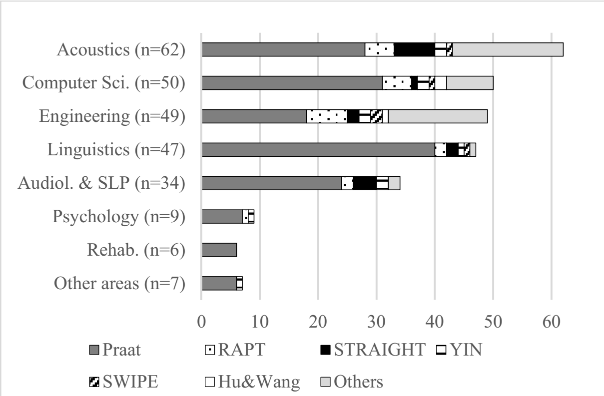

Figure 1. *Number of publications using the different F_0 estimation methods across different research areas.*

3. Evaluation

3.1. Speech data

The PTDB-TUG corpus [15] was used as a ground truth reference. This corpus contains 4720 recorded sentences from 20 speakers (10 male and 10 female). The sentences are recorded both through microphone and laryngograph at a 48 kHz sampling rate, with a 16 bit resolution. From this material, the reference F_0 trace – as extracted by means of RAPT [9] from the high-pass filtered laryngograph signal as described in [15] – was used as the ground truth. Microphone signals were used as input to the F_0 extraction as described below.

3.2. F_0 estimation

Three of the four most frequently used F_0 estimation methods, as reported above, were used: Praat [5], RAPT [9] and YIN [7]. These were selected on the basis of their frequency of use, and of their potential to run “as-is”, e.g. without relying on voicing decisions from other sources. (For this reason, STRAIGHT was not included in the analysis.)

For all three methods, F_0 values were computed in steps of 0.01 seconds, thus matching the frame rate in the reference material. All methods were implemented with two different configuration settings; one with their respective default settings (described below), and one with pitch range values adapted by gender, according to the recommendations in the online Praat manual [33]. Hence, for male speakers, the pitch range was set to 75-300 Hz, and for female speakers to 100-500 Hz.

For F_0 estimation in Praat, the standard autocorrelation method was used. Default values were used in the standard configuration setting, with a pitch search range of 75-500 Hz. For the gender-adapted settings, the increased pitch floor from 75 to 100 Hz in the female settings results in a reduced frame rate (93 frames/sec instead of 100 frames/sec in the default/male settings). To compensate for this reduction, the number of frames in the generated female pitch files was stretched by interpolation by a factor of 1.07.

RAPT was run through an implementation in Snack [29], with default settings (method: ESPS, maxpitch: 400 Hz, minpitch: 60 Hz, window length: 0.0075). In the gender-adapted version, pitch range was adapted as described above.

YIN was run through a Matlab implementation available at [34]. In the standard configuration, default settings as described in [7] were used (minf0: 30 Hz and maxf0: SR/(4*dsratio), threshold: 0.1). In the gender-adapted configuration, pitch range settings were adapted as described above.

3.3. Analysis

The F_0 traces yielded by each of the three extraction methods were each evaluated against the reference F_0 traces in the PTDB-TUG corpus. In accordance with [8], the following evaluation metrics were used:

- **Gross Pitch Error (GPE)**: the proportion of frames – considered voiced by both pitch tracker and ground truth – where the relative pitch error is higher than 20%.
- **Fine Pitch Error (FPE)**: the standard deviation of the distribution of relative error values (in cents) from the frames that do not have gross pitch errors.
- **Voicing Decision Error (VDE)**: the proportion of frames for which an incorrect voiced/unvoiced decision is made.
- **F0 Frame Error (FFE)**: the proportion of frames for which an error (either according to the GPE or the VDE criterion) is made. FFE can be considered a single measure of overall performance [17].

3.4. Evaluation results

Table 2 displays the results of the evaluation of the different F_0 estimation methods, for female and male speakers, and for the group as a whole. It is clear from the figures in Table 2 that Praat’s overall performance, as estimated by the FFE, is better than that of both RAPT and YIN. However, it is also clear that this pattern is largely driven by Praat’s superior accuracy in detecting voicing, particularly when compared to YIN; in this respect, YIN does not at all meet the performance of the other trackers. A closer look at the YIN’s inaccurate voicing decisions reveals that a majority of these errors (93% in the default setting, and 79% in the gender-adapted setting) are cases of over-identification of voicing. It should be observed, however, that on frames where the three candidate trackers agree with the reference data on voicing (66-75% for YIN, as compared to 95% for Praat), YIN is more accurate (as measured both by GPE and FPE) than the other two candidates, although the advantage over Praat is quite marginal.

Adaption of pitch range settings by gender is most beneficial for Praat, whereas the positive effects of a similar adaption for RAPT and YIN are less obvious. For YIN, in fact, the gender-adapted settings generally lead to deteriorated accuracy. For Praat, however, the adapted settings benefits the F_0 estimation accuracy for both female and male speakers. However, not even in the gender-adapted version of Praat does the FFE for female speakers reach the performance on male speakers in the non-adapted/default version.

Table 2. *Evaluation results for the three F_0 estimation methods, with default (def.) and gender-adapted (m/f) settings.*

Method	Female speakers				Male speakers				All speakers			
	GPE (%)	FPE (cents)	VDE (%)	FFE (%)	GPE (%)	FPE (cents)	VDE (%)	FFE (%)	GPE (%)	FPE (cents)	VDE (%)	FFE (%)
Praat (def.)	2.33	30.47	5.06	7.39	1.83	37.10	4.87	6.70	2.09	33.83	4.96	7.05
Praat (m/f)	2.10	28.85	**4.82**	**6.92**	1.28	34.82	**4.18**	**5.47**	1.69	31.91	**4.51**	**6.20**
RAPT (def.)	3.89	41.87	8.88	12.77	5.47	48.17	6.81	12.29	4.67	44.95	7.86	12.53
RAPT (m/f)	3.86	41.90	8.39	12.25	4.90	48.13	7.51	12.41	4.37	44.86	7.96	12.33
YIN (def.)	**1.66**	31.18	24.17	25.83	**1.12**	32.61	26.78	27.90	**1.39**	31.85	25.47	26.86
YIN (m/f)	1.80	**25.52**	34.28	36.08	1.32	**29.40**	33.34	34.65	1.56	**27.49**	33.81	35.38

4. Discussion

By surveying the last five years' literature for speech-related studies involving F_0 estimation, the present investigation has identified Praat, RAPT, STRAIGHT and YIN as the most frequently used. By evaluating three of these methods against a state-of-the-art ground truth reference, it was further shown that the most frequently used method – Praat – was also the most accurate. Moreover, the comparison of gender-adapted pitch range settings against default settings revealed that – at least for Praat – accuracy could be further improved by applying different settings for male and female speakers.

A secondary finding in the bibliometric survey is the distribution of different F_0 estimation methods across research areas. Praat was found to hold a position as the dominant method across a wide range of research areas, such as Linguistics, Computer Science and Audiology & Speech-Language Pathology. On the other hand, many methods or algorithms are reported in single or few studies, particularly within Engineering and Acoustics. Considering that these are the areas where much of the development of new methods presumably takes place, this should not come as a surprise. Time will tell if any of these methods will be more frequently used in the future. A rather discomforting discovery was that many studies report F_0 values but with no specification regarding how these were derived. From the perspective of replicability and reliability, this is – of course – unfortunate.

In any bibliometric survey, there is a chance that the search criteria may not cover all relevant publications. This risk is present also here. Moreover, with only one person performing the manual lookup of the publications, reliability may be questioned. However, the task of identifying whether or not F_0 estimation method is explicitly reported, and – in case it is – what method this is, is quite straightforward, and does not rely on subjective interpretation. Therefore, data noise of this type should not be a major concern.

Regarding the evaluation, there are some aspects that deserve to be discussed. First, the high error rates observed for YIN on voicing decisions are not surprising, considering that YIN is designed only to provide F_0 estimates, treating the voiced/unvoiced decision as a separate issue [8] [7]. From this perspective, the suboptimal evaluation results for YIN may be considered unfair. However, for the purpose of this investigation, the different estimation methods were intended to be used "as is", with default settings. In other contexts, coupling YIN with voicing decisions from another source is a way of improving the F_0 estimates [12]. One may argue that YIN is disprivileged also in the default configuration settings, as the pitch range is wider than in the default settings of the other two candidates. YIN's default settings (with 40 Hz as the lower pitch threshold) are presumably less tailored to speech than the other two candidates. However, the fact that YIN with its default configuration actually performs better than when pitch range settings are adapted by gender, serves to indicate that this potential disfavor does not have a major effect.

The fact that RAPT is used both in the processing of laryngograph data (available in the PTDB-TUG corpus) and as a candidate F_0 estimation method, one may argue that this method is given an advantage in the evaluation. Hence, the evaluation results for RAPT may overestimate its actual performance.

It should be acknowledged that the findings presented here are generalizable only to similar conditions. Hence, it is reasonable to assume that the evaluation results generalizes to other situations involving clean recordings of single adult speakers; however, they may not extend to more challenging conditions like speech-in-noise or pathological speech samples. Moreover, researchers may want to analyze the accuracy of different F_0 estimation methods in more detail; in such contexts, tools like WinPitch [6] may provide better assistance than a general evaluation of the kind presented here.

As scientific progress is continuous, the current investigation will need to be extended with the inclusion of new F_0 estimation methods when such methods become available. By relying on publically available data (the PDTB-TUG corpus), and standard evaluation outcome measures, the experimental setup can easily be replicated by others.

In the absence of a golden standard for the analysis of F_0 in speech science, deciding what method to use is an ad hoc choice. The present study has addressed this unsatisfying state of affairs by identifying which methods are currently the most frequently used, and by evaluating their performance against the currently best available ground truth reference. The findings should provide comfort to researchers who rely on (one of) the most accessible tools available for F_0 estimation – Praat – that their estimations are among the most accurate (if not the most accurate) that can be achieved. Also, the evaluation gives a reasonable estimation of the error rates that can be expected. Moreover, the finding that adapting pitch range settings by gender leads to further improvements in accuracy – at least for Praat – should encourage researchers to do this routinely.

5. Acknowledgements

The work presented here is funded by the Swedish Research Council projects (VR 2009-1764) *Intonational variation in questions in Swedish* and (VR 2015-01525) *Functional consequences of children's misarticulation in continuous speech*.

6. References

[1] M. Zellers, ”Prosodic variation for topic shift and other functions in local contrasts in conversation,” *Phonetica,* vol. 69, nr 4, pp. 231-253, 2012.

[2] S. V. Crespo, C. Kaland, M. Swerts and P. Prieto, ”Perceiving incredulity: The role of intonation and facial gestures,” *Journal of Pragmatics,* vol. 47, nr 1, pp. 1-13, 2013

[3] I. Siegert, K. Hartmann, D. Philippou-Huebner and A. Wendemuth, ”Human Behaviour in HCI: Complex Emotion Detection through Sparse Speech Features,” in *International Workshop on Human Behavior Understanding (HBU)*, Barcelona, Spain, 2013.

[4] D. P. Snow and D. J. Ertmer, ”Children's development of intonation during the first year of candlear implant experience,” *Clinical Linguistics & Phonetics,* vol. 26, nr 1, pp. 51-70, 2012.

[5] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noice ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, Amsterdam, The Netherlands, 1993.

[6] P. Martin, ”Multi methods pitch tracking,” in *Proceedings of Speech Prosody*, Shanghai, China, 2012.

[7] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA,* vol. 111, no. 4, pp. 1917-1930, 2002.

[8] T. Drugman and A. Alwan, ”Joint Robust Voicing Detection and Pitch Estimation Based on Residual,” in *Proceedings of Interspeech 2011*, Florence, Italy, 2011.

[9] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., Amsterdam, Elsevier Science, 1995, pp. 495-518.

[10] A. Camacho, *A sawtooth waveform inspired pitch estimator for speech and music,* University of Florida, 2007.

[11] A. Vogel, P. Maruff, P. J. Snyder and J. C. Mundt, ”Standardization of pitch-range settings in voice acoustic analysis,” *Behavior Research Methods,* vol. 41, nr 2, pp. 318-324, 2009.

[12] A. de Cheveigné and H. Kawahara, ”Comparative evaluation of F0 estimation algorithms,” in *Proceedings of Eurospeech*, Aalborg, Denmark, 2001.

[13] M. Wohlmayr and P. F., ”Multipitch Tracking Using A Factorial Hidden Markov Model,” in *Proceedings of Interspeech*, 2008.

[14] O. Babacan, T. Drugman, N. D'Alessandro, N. Henrich and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Proceedings of ICASSP*, Vancouver, Canada, 2013.

[15] G. Pirker, M. Wohlmayr, S. Petrik and F. Pernkopf, "A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario," in *Proceedings of Interspeech 2011*, Florence, Italy, 2011.

[16] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer and J. &. K. S. Trmal, ”A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition,” in *Proceedings of ICASSP*, Florence, Italy, 2014.

[17] W. Chu and A. Alwan, ”Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend,” in *Proceedings of ICASSP*, Taipei, Taiwan, 2009.

[18] P. C. Bagshaw, S. M. Hiller and M. A. Jack, ”Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching,” in *Proceedings of Eurospeech*, Berlin, Germany, 1993.

[19] A. Wrench, ”A multichannel/multispeaker articulatory database for continuous speech recognition research,” in *Phonus*, 2000.

[20] F. Plante and G. &. A. A. Meyer, ”A pitch extraction reference database,” in *Proceedings of Eurospeech*, Madrid, Spain, 1995.

[21] K. Hirose, K. Andi, R. Mihara, H. Hashimoto and D. &. M. N. Saito, ”Adaptation of Prosody in Speech Synthesis by Changing Command Values of the Generation Process Model of Fundamental Frequency,” in *Proceedings of Interspeech 2011*, Florence, Italy, 2011.

[22] K. J. Kohler, ”Communicative Functions Integrate Segments in Prosodies and Prosodies in Segments,” *Phonetica,* vol. 68, nr 1-2, pp. 26-56, 2011.

[23] I. Mennen, J. M. Scobbie, E. de Leeuw and S. &. S. F. Schaeffler, ”Measuring language-specific phonetic settings,” *Second Language Research,* vol. 26, nr 1, pp. 13-41, 2010.

[24] J. Tepperman and E. Nava, ”Where should pitch accents and phrase breaks go? A syntax tree transducer solution,” in *Proceedings of Interspeech 2011*, Florence, Italy, 2011.

[25] F. Kurth, A. Cornaggia-Urrigshardt and S. Urrigshardt, ”Robust F0 estimation in noisy speech signals using shift autocorrelation,” in *ICASSP*, Florence, Italy, 2014.

[26] T. Ewender and B. Pfister, ”Accurate Pitch Marking for Prosodic Modification of Speech Segments,” in *Proceedings of Interspeech 2010*, Makuhari, Japan, 2010.

[27] T.-C. Yeh, M.-J. Wu, J.-S. R. Jang, W.-L. Chang and I.-B. Liao, ”A hybrid approach to singing pitch extraction based on trend estimation and Hidden Markov Models,” in *Proceedings of ICASSP*, Kyoto, Japan, 2012.

[28] G. Seshadri and B. Yegnanarayana, ”Performance of an Event-Based Instantaneous Fundamental Frequency Estimator for Distant Speech Signals,” in *Proceedings of ICASSP*, Prague, Czech Republic, 2011.

[29] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *International Conference on Speech and Language Processing*, Beijing, China, 2000.

[30] H. Kawahara, J. Estill and O. Fujimura, ”Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” in *Proceedings of MAVEBA*, Florence, Italy, 2001.

[31] A. Camacho and J. Harris, ”A sawtooth waveform inspired pitch estimator for speech and music,” *JASA,* vol. 124, nr 3, pp. 1638-1652, 2008.

[32] G. N. Hu and D. L. Wang, ”Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. Neural Networks,* vol. 15, nr 5, pp. 1135-1150, 2004.

[33] P. Boersma and D. Weenink, ”Intro 4.2. Configuring the pitch contour,” Institute of Phonetic Sciences, The University of Amsterdam, 30 August 2005. [Online]. Available: http://www.fon.hum.uva.nl/praat/manual/Intro_4_2__Configuring_the_pitch_contour.html. [Accessed March 15, 2016].

[34] A. de Cheveigné, ”Alain de Cheveigné,” Equipe Audition, [Online]. Available: http://audition.ens.fr/adc/. [Accessed March 10, 2016].