euronoise

A system for automatic detection and correction of detuned singing

Michał Lech and Bozena Kostek

Gdansk University of Technology, Multimedia Systems Department, 11/12 Gabriela Narutowicza Street, 80-952 Gdansk, Poland
mlech@sound.eti.pg.gda.pl

The aim of the paper is to show a system engineered for automatic detection and correction of detuned singing. For this purpose, existing methods of fundamental frequency detection and pitch correction are reviewed. In addition, main characteristics of some existing detuning systems are presented. As algorithms for fundamental frequencies detection and pitch correction, the fast autocorrelation and HPS (Harmonic Product Spectrum), and the modified phase vocoder and PSOLA (Pitch-Synchronous Overlap-Add) are chosen and examined. Four possible combinations of the algorithms are reviewed not only in the context of fundamental frequency detection and pitch shifting correctness but also with regard to the quality of the resulting singing signal. Experiments are performed on both male and female singing samples consisting of a variety of tones and various articulations. Basing on the obtained results, it is concluded that the HPS and PSOLA algorithms are the optimum choice as means to correct detuned singing. In addition, listening tests are performed in order to confirm objective measurements of pitch detection and correction. The system is implemented in JAVA. Conclusions are drawn and proposals of improvements are provided.

1 Introduction

Within the past ten years in the musical market, especially the one connected with the popular music, there has developed the fashion for creating records putting great emphasis on quality and tone with simultaneously attaching less importance to feeling. Singers were demanded to sing ideally in tune even if it resulted in lack of emotions and if their efforts did not meet producers' expectations as to their voices, thanks to rapid development of modern technology, were corrected using computer systems. Today this fashion is gradually changing allowing for barely audible out of tune notes if they are sung or played with extraordinary feeling but till now a lot of applications able to improve this feature have been developed.

As early as the beginning of the nineties the systems were indeed able to correct false notes but simultaneously caused audible changes to sound. The real breakthrough was dated 1996 when Auto-Tune system of Antares Audio Technologies, which was able to shift a pitch without significant interference in original sound, was presented. Today, the pitch correction systems provide not only pitch shifting but also possibility of changing voice timbre or adding 'artistic hoarseness' to voice.

2 Fundamental frequency detection

In the common approach to pitch correction at the first step the fundamental frequency detection is performed. There are many methods of fundamental frequency detection, operating in time domain, frequency domain or, thanks to time-frequency transformations, in the field of both domains [1, 7]. Using time-domain methods one can retrieve fundamental frequency directly from the time form of a signal, without the need for complex transformations. The typical characteristics of time methods are: good resolution, occurrence of octave errors and low resistance to noise. Despite the fact that there is no necessity of performing complex transformations, without some optimization modifications these methods can be time expensive. Among time methods of fundamental frequency detection one can mention: threshold methods, ACF (*Autocorrelation Function*), AMDF (*Average Magnitude Difference Function*), envelope analysis [7, 10, 12, 13].

Frequency methods of fundamental frequency detection are based on a signal spectrum analysis. In case of sound having a definable pitch, its spectrum composes of series of peaks corresponding to fundamental frequency and harmonic frequencies being its multiplicity. Analyzing a distribution of these peaks it is possible to define fundamental frequency of a sound [1]. As an example, following frequency methods of fundamental frequency detection can be mentioned: HPS (*Harmonic Product Spectrum*), double Fourier transformation, cepstral method [1, 3, 7, 9, 10].

Another worth-mentioning type of means for fundamental frequency detection are perceptual methods. They are based on the way of perceiving sound by the human hearing system. As an example of perceptual methods, fundamental frequency detector based on Licklider dualism theory of pitch perception can be mentioned [5, 11]. The algorithm of the detector, developed by Slaney and Lyon [11], is based on the utilization of cochlea model in connection with a set of values of the autocorrelation function. A so-called correlogram, which is a result of performing autocorrelation, is filtered, non-linearly amplified and summed up among each channel values. Basing on the analysis of resulting peaks fundamental frequency can be determined. The algorithm is resistant both to noise and phase changes [11].

3 Pitch correction methods

Pitch correction, like fundamental frequency detection, can be performed in time domain, frequency domain or in the field of both domains. The bases for construction of pitch shifting algorithms are: phase vocoder in frequency domain and time scaling in time domain. Both algorithms in their original form result in audible unwanted changes to sound. However, today's computer computational power is sufficient enough to introduce some improvements, such as phase adjusting among adjacent frames. More advanced methods based on human perception or on the usage of wavelets are also in use [2, 4].

Time-domain methods are based on the assumption that in the sufficiently small frame (e.g. 1024 samples) the signal is periodic [8]. Pitch shifting within these methods is basically modification of fundamental period within each frame. The commonly used time-domain method is PSOLA (*Pitch-Synchronous Overlap-Add*), which performs pitch correction basing on series of marks positioned in the signal in determined distance from each other. The ideal distribution is such that points are positioned in the signal peaks and simultaneously in equal distance from each other. Due to the fact that fundamental period slightly changes

within the chosen frame such distribution is not possible. Therefore, one aims at such distribution that a distance between neighbouring points is close to the first, detected fundamental period and points are positioned near the signal peaks [8]. Goncharoff-Gries algorithm is used in this case. In the next stage a new vector of points, spaced in identical distance equal fundamental period corresponding to the desired pitch, is generated. For each new point the nearest mark in the original vector of points is being found and a part of signal within two original fundamental periods separated by this point is copied into a new place determined by a new point. Summed up, overlapping parts compose the pitch-corrected signal [2, 8].

Pitch correction in frequency methods lies in modification of spectral bins composing peaks with retaining existing relationship among them. In a phase vocoder, each peak of a spectrum is shifted by a determined value multiplied by a number of harmonic frequency corresponding to the peak being processed at the given moment. To determine the shift value properly the frequency detection (in frequency domain) should be done using parabolic interpolation of peak maximum and neighbouring maxima [4, 6].

4 Existing pitch correction systems

After the success of previously mentioned Auto-Tune application there have appeared many various, continuously being improved solutions on the market. Among most popular systems one can mention Antares Auto-Tune [14], Celemony Melodyne [15], Serato Pitch'n'Time [16], TC-Helicon VoiceOne [17]. The systems are available as plug-ins of various types for popular music editors such as Steinberg Cubase and Pro Tools or as the autonomous rack units being able to correct pitch in real time. Below, there are some characteristics of the mentioned systems.

Work with Antares Auto-Tune can be started with choosing gender of voice or instrument. This enables the system to choose correction algorithm appropriate for the input characteristic. Pitch correction can be performed in one of two available modes: automatic and graphic. In the automatic mode a correction is performed basing on a key automatically retrieved from the MIDI pattern or, in case there is no particular key in the system database, manually entered using virtual or external MIDI controller. In the graphic mode, detected frequencies are presented as a contour which can be freely modified using various graphic tools. The application enables to control the level of correction to avoid excessive adjustment of sung or played phrase to the pattern [14].

The Celemony Melodyne application was for the first time presented in 2001 during winter NAMM exhibition. Its constructors have used innovative approach to sound representing which is presenting each note as the object of shape, length and height determining its characteristic. Height of the object represents velocity, length – duration, and vertical position – pitch. Within each object and between adjacent objects there is a frequency contour which represents frequency modulations and pitch drift. One can modify each note by manipulating the corresponding object and contours [15].

Another pitch correction application, Serato Pitch'n'Time, is based on human sound perception and is available in three versions, which differ in possibilities and number of available functions. Most advanced one, version Pro, allows to change pitch by ±36 semitones and simultaneously change tempo (independently) in a range of 12.5% up to 800% of the original value. One can modify pitch using simple function of increasing or decreasing it by a chosen number of semitones, operating on graphical representation of a signal or determining pitch by tempo settings. Application provides a processing of stereo tracks without phasing and processing of matrix encoded tracks without losing surround information. Serato Pitch'n'Time is intended for use with Pro Tools [16].

TC-Helicon VoiceOne as opposed to the previous systems is an autonomous unit equipped with DSP processor able to correct pitch in real time. The equipment is utilizing both the classical pitch correction algorithms basing on formants and algorithms specially intended for human voice. Pitch correction is performed basing on one of 48 predefined keys or on a key entered by user with MIDI controller [17].

5 Research on algorithms

To develop own correction system of detuned singing the research on chosen algorithms of fundamental frequency detection and pitch correction was performed. Examined algorithms were: fast autocorrelation, HPS, PSOLA and modified phase vocoder. The Matlab codes of the algorithms come from Connexions website [18].

5.1 Fundamental frequency detection algorithms

At the first step of examining fast autocorrelation algorithm, impact of correlation threshold on fundamental frequency detection effectiveness was checked. Analysis was performed for values equal 0.005, 0.01, 0.015, 0.02, 0.025, 0.03 with frame length equal 8192 samples and hop size equal 2048 samples. The input signal was male voice singing notes from A3 to E4. It was assumed that the proper detection was such that the relative error should be less than 3%. The error threshold of such level let treat fundamental frequency as correctly detected when it was in range described by the Eq. (1). In this equation P denotes detected pitch whereas P_{ref1} and P_{ref2} are, respectively, reference pitch of the nearest tone from the twelve-semitones scale lower than reference tone for P and reference pitch of the nearest tone higher than reference tone corresponding to P.

$$P - \frac{P - P_{ref1}}{2} < P < P + \frac{P_{ref2} - P}{2} \tag{1}$$

Duration of a tone considered is within the particular number of frames among which each has 8192 samples. Effectiveness of particular fundamental frequency detection is designated as a ratio of number of correct detections (number of frames among which detection was correct) and all detections for given tone (number of all frames containing examined tone). The results of research described above are given in Figs. 1 and 2.

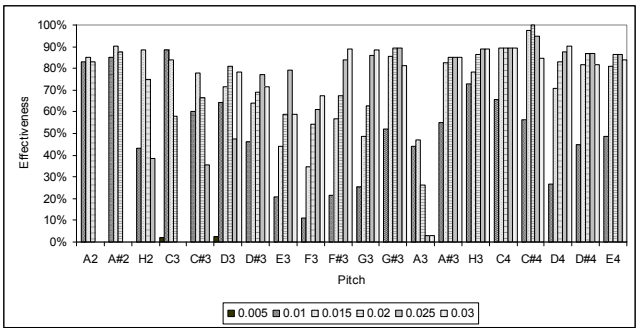

Fig. 1 Fundamental frequency detection effectiveness using fast autocorrelation algorithm for particular tones depending on correlation threshold

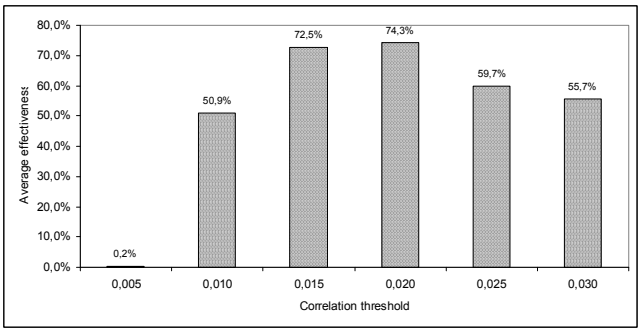

Fig. 2 Average fundamental frequency detection effectiveness for fast autocorrelation algorithm depending on correlation threshold

One can notice that the optimal threshold value is contained in range [0.015, 0.025]. The best results were obtained for threshold equal 0.020 and such value was used for further experiments.

The next stage of analyzing fast autocorrelation algorithm was to check fundamental frequency detection correctness in relationship with frame length and hop length. Tests were performed basing on the sample of male voice singing A3 – E3 notes and female voice singing H4 – E4 notes with glissando articulation in both cases. The following frame lengths were used: 512, 1024, 2048, 4096, 8192 and 16384 samples. For each frame length w experiment was performed thrice, for hop sizes equal $\frac{1}{4}w$, $\frac{1}{2}w$, $\frac{3}{4}w$. The results of the experiment are presented in Figs. 3 and 4.

Analyzing the results obtained for female voice one can notice that detection effectiveness is higher for lower frame lengths, beside the fact that for lengths 512 up to 2048 differences are negligible. However, using male singing sample for frame length equal 512 samples and hop size $\frac{1}{4}w$ the obtained results are distinctly worse than for three next frame lengths. Also, comparing results with these obtained for female voice one can notice that for lengths equal 8192 and 16384 samples results are worse. These differences might be caused by individual characteristics of both sung samples such as velocity, attack, voice strength. For both male and female samples at the same time the best results were obtained for frame lengths equal 2048 and 4096 samples and hop size equal $\frac{1}{2}w$.

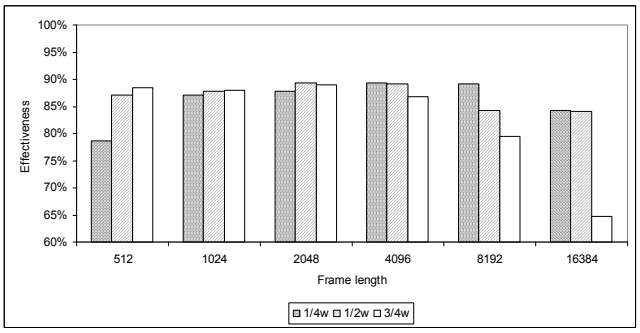

Fig. 3 Fundamental frequency detection effectiveness using fast autocorrelation algorithm for male voice depending on frame length and hop size

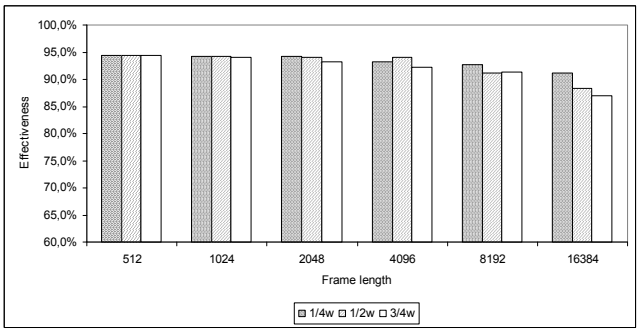

Fig. 4 Fundamental frequency detection effectiveness using fast autocorrelation algorithm for female voice depending on frame length and hop size

The research on relationship between frame length and the fundamental frequency detection correctness was also performed for HPS algorithm. Utilized input samples as well as frame lengths and hop sizes were the same as in the previous case. The obtained results have been presented in Figs. 5 and 6.

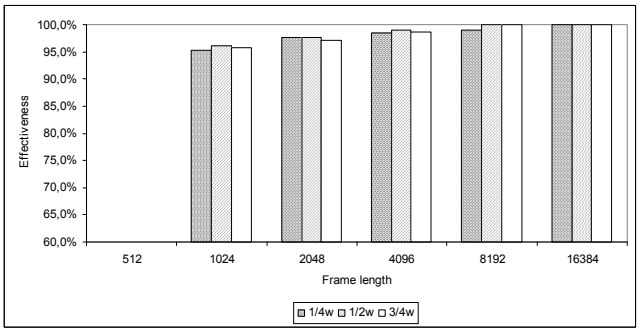

Fig. 5 Fundamental frequency detection effectiveness using HPS algorithm for male singing sample depending on frame length and hop size

Using HPS algorithm, for longer frames better results were obtained, although for frame lengths equal 1024 up to 16384 samples differences were slight (within 5% change). For frame length equal 16384 samples fundamental frequency detection effectiveness was near the level of 100%. For length of 512 samples and male singing sample the effectiveness was less than 7%. For female voice such drawback of the algorithm was not observed (the effectiveness was over 95%). Again, like in the fast autocorrelation algorithm such difference might have been caused by specific articulation used with the male singing.

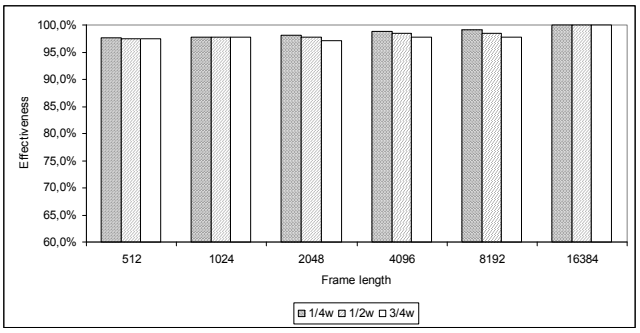

Fig. 6 Fundamental frequency detection effectiveness using HPS algorithm for female singing sample depending on frame length and hop size

5.2 Pitch correction algorithms

The next stage of the research was examining pitch correction algorithms with regard to correctness and quality of resulting signal. Four possible configurations of fundamental frequency detection algorithms and pitch correction algorithms were reviewed. Tests were performed using male and female with glissando articulation singing sample. The correction based on increasing the first tone of the glissando and preserving it for the whole duration of the sample. It was assumed that the proper correction was such that the resulted pitch equaled the reference pitch and quality was subjectively rated as level of general similarity in sound with the original signal.

Analyzing the obtained results one can notice that irrespective of the utilized detection or correction algorithm high impact on the final effect has the hop size corresponding to the chosen frame length. Performing correction with hop size equal $\frac{3}{4}w$ result in chopped signal. The specific tremolo effect of a speed depending on used frame length is audible. Utilizing small hop size (e.g. $\frac{1}{4}w$) let minimize this effect by multiple summing of overlapping frames multiplied by Hanning window. Hereby, the signal being the average of parts of the signal among adjacent frames in terms of shape and amplitude is obtained.

Analysis of the results respectively to chosen frame length has showed that the correction effectiveness depends on the particular fundamental frequency detection algorithm. Using autocorrelation algorithm with a long frame resulted in skipping tones of a short duration (shorter than frame length). This effect could be very clearly observed for the correction of glissando articulation with frame length equal 16384 samples. Such problem did not exist using HPS algorithm as it does not operate on time-domain form of a signal.

The research on a quality of corrected signals depending on length of used frame showed that for PSOLA algorithm the shorter frame used the more audible distortion or flutter to the sound. For modified phase vocoder there was no relationship between frame length and sound quality observed but negative effect on formants resulting in unnatural metallic sound was noticed.

6 System design and validation

6.1 System design

Basing on the results of the research described in the previous section it was concluded that the optimum choice for the correction of detuned singing are HPS and PSOLA algorithms. For the chosen configuration the best results were obtained using frame of 8192 samples and hop size equal 2048 samples. These are default values in the developed system. The research has also showed that in some cases, e.g. in glissando articulation, shorter frames are necessary. Therefore, in designed system one is able to chose different frame length and hop size from the predefined set of values.

The system was implemented in JAVA, as it provides many, free sound libraries. The development environment used was Netbeans IDE 5.5 with JDK 1.6 and the runtime environment was JRE 1.6. The user graphical interface was developed using Swing library. In Fig. 8 there is the main window of the application showing correction of the input signal.

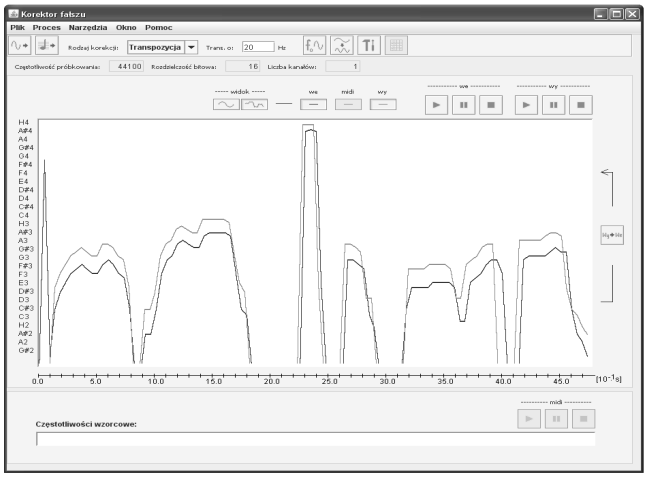

Fig. 8 The main window of the application with a view of pitch of the original signal and corrected one changing in time

It was assumed that the signal to be corrected is always stored in the WAVE PCM file of frequency sampling equal 44100Hz and bit resolution equal 16bps. The signal is single mono track. The system provides two ways of pitch correction. The first one is based on the MIDI pattern loaded from SMF file and the second one lies in decreasing or increasing pitch basing on a given fixed value in Hz entered by user. The additional requirement was to provide possibility of performing detection without proceeding with correction. Before performing detection or correction user has a possibility to chose frame length from the set of following values: 1024, 2048, 4096, 8192, 16384 samples. For the chosen frame length one can set hop size to $\frac{1}{4}w$, $\frac{1}{2}w$ or $\frac{3}{4}w$. Default hop size is $\frac{1}{4}w$. Additionally, user can set downsampling factor of the HPS algorithm and path slope for PSOLA algorithm. Default downsampling factor is 5 and default path slope equals 4.

6.2 System validation

The pitch correction provided by the system was validated using male singing sample consisted of notes H3 to G4 sung in sequence, female and male glissando articulations used previously for testing Matlab algorithms and the part of vocal track of the own composition. For the sequence of tones four MIDI patterns were used. The first two patterns contained sequences increased and decreased by whole tone. The third pattern consisted of sung notes, therefore its aim was to level each out of tune note. Tones of the last pattern were determined by random number generator giving numbers from 59 (note H3 MIDI code) to 67 (note G4 MIDI code). For male and female glissando articulation three patterns were prepared. The first pattern consisted of the note beginning the glissando increased by a whole tone, the second one – the note with which the glissando begun and the third one – the note beginning the glissando decreased by a whole tone. For the part of vocal line of the own composition MIDI pattern containing phrase increased by fourth was prepared.

The processes of fundamental frequency detection and pitch correction were performed for default values. After correction listening tests were performed as well as checking obtained pitch values by treating the corrected signal as an input of previously examined Matlab HPS algorithm. Analyzing obtained results it was stated that the three last tones were not shifted correctly (fundamental frequency detection effectiveness equal respectively 5.3%, 0.0% and 1.7% for sample containing notes decreased by a whole tone and 71.4%, 70.6%, 5.2% for sample consisting of notes increased by a whole tone). For other notes the average fundamental frequency detection effectiveness equaled 81%. When using randomly generated MIDI pattern, although pitch was shifted correctly, quality of resulting sound was very low. Analysis of the singing sample let conclude that the problems were caused by voice articulation. Three last notes were sung with much greater attack than the others.

7 Conclusions

The listening tests of the developed application have shown that using classical, common algorithms of fundamental frequency detection and pitch correction it is hard to develop the system providing faultless correction and preserving the original sound quality. To obtain satisfying results, creating such system one should consider perceptual methods and wavelet transformations. The research on the algorithms implemented in Matlab has shown that due to the non-deterministic aspects of human voice simple mathematical models are not sufficient means to describe it.

Considering the developed system better results could be achieved by implementing also the other two algorithms reviewed in the research and connecting them with existing ones. Then, depending on a type of correction to perform, time-domain or frequency-domain algorithm could be used or both algorithms could run simultaneously and basing on the results from the population of adjacent frames more reliable results could be chosen. Another interesting feature would be variable frame length depending on timing value defined in MIDI pattern.

References

[1] Dziubiński M., Kostek B., "Octave Error Immune and Instantaneous Pitch Detection Algorithm", *J. New Music Research*, 34, No. 3, 273 – 292, 2005.

[2] Holzapfel M., Hoffmann R., Höge H., "A Wavelet-Domain PSOLA Approach", *Third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998

[3] Hu J., Xu S., Chen J., "A modified pitch detection algorithm", *IEEE Communications Letters*, 5, (2), 2001.

[4] Laroche J., Dolson M., "New Phase-Vocoder Techniques for Pitch-Shifting, Harmonizing, and other Exotic Effects", *Proc. 1999 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 91-94, 1999.

[5] Licklider J., "A duplex theory of pitch perception", *Psychological Acoustics*, Stroudsburg, PA, 1979.

[6] Middleton G., "Frequency Domain Pitch Correction", *Connexions Project*, mod. m11715, 2003.

[7] Middleton G., "Pitch Detection Algorithms", *Connexions Project*, mod. m11714, 2003.

[8] Middleton G., "Time Domain Pitch Correction", *Connexions Project*, mod. m11711, 2003.

[9] Noll A. M., "Cepstrum Pitch Determination", *J. Acoust. Soc. of America*, 14, 293-309, 1967.

[10] Rabiner L. R., Cheng M. J., Rosenberg A. E., McGogenal C. A., "A comparative performance study of several pitch detection algorithms", *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-24, (5), 1976.

[11] Slaney M., Lyon R., "A Perceptual Pitch Detector", *International Conference on Acoustics Speech and Signal Processing*, vol. 1, 357-360, 1990.

[12] Tan L., Karnjanadecha M., "Pitch Detection Algorithm: Autocorrelation Method and AMDF", *Proceedings of the 3rd International Symposium on Communications and Information Technology*, 2:551-556, 2003

[13] Ying G. S., Jamieson L. H., Mitchell C. D., "A Probabilistic Approach To AMDF Pitch Detection", *Proc. 4th Int. Conf. on Spoken Language Processing*, Philadelphia, PA, October, 1201-1204, 1996.

[14] Antares Audio Technologies Auto-Tune official website, http://www.antarestech.com/

[15] Celemony Melodyne official website, http://www.celemony.com/cms/index.php?id=products_plugin

[16] Serato Pitch'n'Time Pro official website, http://www.serato.com/products/pnt/

[17] TC-Helicon VoiceOne official website, http://www.tc-helicon.com/VoiceOne

[18] Connexions website, http://cnx.org/