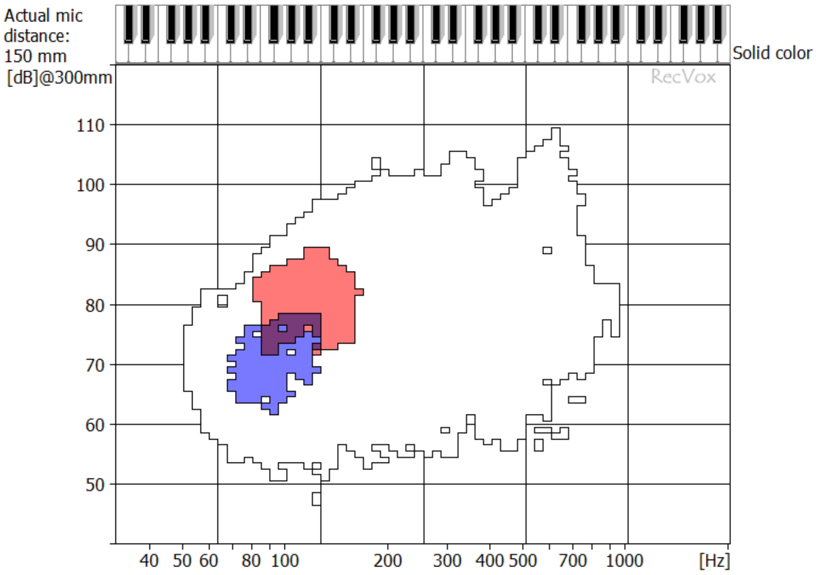
Description and validation of RecVox – a free software for real-time phonetography

Svante Granqvist
-Tolvan Data
-Division of Speech and Language Pathology, Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet, Stockholm, Sweden
-KTH Royal Institute of Technology, School of Engineering Sciences in Chemistry, Biotechnology and Health (CBH), Dept. of Biomedical engineering and Health systems, Stockholm, Sweden and

Introduction

The phonetogram is a 2-dimensional graph that illustrates fundamental frequency (f_o) on the horizontal axis and the sound pressure level (SPL) on the vertical axis. The fundamental frequency is typically expressed in semitones, and the sound pressure level in decibels, that is, both the horizontal and vertical axes are logarithmic. A third dimension can be displayed in terms of color density. If this third dimension represents occurrence, the phonetogram can be seen as a 2-dimensional histogram of the f_o and SPL. However, the third dimension can also be used to represent voice quality measures, such as the crest factor or the level difference between the first two harmonics, L_{H1}-L_{H2}.



Three phonetograms generated in RecVox. The blue phonetogram is a voice range profile recorded in a quiet environment. The red phonetogram is recorded wearing headphones playing pink noise, provoking a louder and more high-pitched voice. Finally, the white phonetogram is a voice range profile, demonstrating the full range of the voice. The microphone is placed at 15 cm distance from the mouth, RecVox presents SPL as if the microphone had been placed at 30 cm distance.

Early phonetograms were recorded manually by asking the participant to phonate at several pitches as soft and loud as possible, thus registering the maximum range of the voice, in terms of f_o and SPL as functions of each other. The phonetogram provides more information than measurement of the f_o and SPL ranges separately as the full SPL range is not possible to reach for all fundamental frequencies. Computer based real time phonetography adds feedback to the process which further encourages the participant to extend the range of the phonetogram, and thus providing improving the measurement of the actual extremes that can be reached with the voice. In the clinic, coaching is however still crucial to register the full range of the voice, and a well-working procedure for the coaching has been developed by for example Hallin et. al. (2012)
Over the years some, but not many, computer programs for real-time phonetography have been made commercially available. Among these are Dr Speech, Lingwaves, Kay Pentax CSL, Phog and the Voice profiler program. The market is however small, and manufacturers struggle with economic viability. The small market

also leads to a relatively high cost for the software licenses, which in turn may prevent clinics from using real-time phonetography.

RecVox is a free software written by the author of this article. Its purpose is to provide real-time phonetography for research and education, and to inspire other developers to create clinically useful phonetography software. RecVox has no integration with patient journals and is not approved as a medical device, but it has some features regarding calibration and installation that is particularly aimed at a setting where different people have different responsibilities for installing and calibrating the equipment and running the recordings.

The purpose of this article is to describe some of the design considerations made in RecVox that may be of general interest for users and developers of real-time phonetography. The purpose of the article is also to validate and test the accuracy of the measurements made in RecVox.

Note on terminology: In this article the term “phonetogram” is used for the graph itself. Consequently, “phonetography” is the process of creating a phonetogram. In principle, a phonetogram can represent any signal, also non-vocal signals. The term “Voice range profile” is used for the phonetogram that is produced when the subject is asked to cover as large area as possible in the phonetogram, possibly only the outline. The term “Speech range profile” is used for the phonetogram produced when habitual speech is recorded, for example when reading a standard text.

Method

Brief description of RecVox

RecVox records phonetograms, while providing immediate visual feedback to the coach and the participant. Both the participant and coach can be recorded simultaneously with separate microphones into separate phonetograms. The coach phonetogram is rarely interesting, but the possibility to separate the two signals into separate phonetograms enables the coach to sit in the same acoustic space as the participant. Phonetograms can be saved, and several phonetograms can be opened and overlayed to enable comparisons, for example to visualize a change of location of the SRP before and after treatment. A wide range of statistics can be extracted from the phonetogram. Audio is recorded during the registration of the phonetogram, making RecVox potentially useful as an all-in-one recording program providing audio, phonetogram and statistics. SPL is calibrated prior to the recording and SPL can be recalculated for 30 cm distance even if the microphone is positioned at another distance. The audio files saved from RecVox can include calibration. The fact that the audio is calibrated immediately when saving has been shown to be a much-appreciated feature in the clinic in the case of the Phog program, as it eliminated the need for keeping track of separate calibration tone files.

This paper describes version 0.0.22 of RecVox but is valid back to version 0.0.19. In this paper, it is assumed that the settings of RecVox are left at their default values unless otherwise is stated.

Description of RecVox algorithms

Audio is registered at a sampling rate of 44.1 kHz using 24-bit resolution. Fundamental frequency and sound pressure level is detected for 40 ms frames using an overlap of 20 ms. In addition to f_o and SPL a third parameter can be displayed as color density in the phonetogram. All parameters are registered at 20 ms intervals, or 50 registrations per second. Wav and smp audio files can be saved at full sampling rate for the audio and the parameters and scaled to 16-bit resolution. Sopran audio files are saved using different sampling rates (44100 and 50 Hz) for the audio and the parameters, as calibrated 32-bit floating-point numbers. Phonetograms can also be saved in the standard phonetogram format, stdpg, that is also used by Phog and Voice Profiler.

Spectrum

The spectrum of each frame is determined by applying a fast Fourier transform (FFT) to the audio frames. A Hanning window and zero-padding the sample number to the next power of two is applied to the audio samples. The spectrum is used consecutively for f_o detection, blocking criteria and third parameter calculation.

Sound pressure level

The SPL of each frame is measured from the power of the samples in the frame. Prior to the measurement, the signal is filtered by a fourth order Butterworth high-pass filter at 40 Hz. This filter suppresses noise with frequencies below the lowest measurable fundamental of 50 Hz. It affects the level at f_o by 0.5 dB at 50 Hz. The corresponding level change for C-weighting is -1.3 dB according to IEC 61672-1. Thus, the voice levels in RecVox are neither A- or C-weighted, but rather Z-weighted and with less effect on the voice signal, and with more suppression of subsonic frequencies than the C-weighting offers. This extra suppression has proven important in some cases where the ventilation system generates subsonic sound. The subsonic filter is only applied to the analysis, the soundfiles generated by RecVox are not affected.

Background noise

RecVox estimates the spectral content of the background noise by first assuming that sound pressure levels between 25 and 40 dB @ 30 cm consists solely of background noise. For these frames, the spectrum is analyzed, and for each frequency a histogram of the level distribution is generated. The median level for each frequency is assumed to be the background level for that frequency. By using this background noise floor, any static background noise can be prohibited from being detected as phonation. The accumulation of background noise data starts as soon as the recording button is pressed, but the recording of phonation starts only after the pause button is released, thus giving a short time for background noise detection prior to the actual recording. A reasonably accurate SPL calibration (within a few decibels) is required for the background noise detection to work well.

When analyzing a pre-recorded soundfile, the same algorithm is applied. Obviously, the unrecorded sound registered between pressing the record and pause button cannot be used, so soundfiles should have a second or more of silence in the beginning, otherwise detection will not start until there is a silent part within the recording. This has implications for generating synthetic signals for testing and validation. Test files must contain “silent” part with only “background noise” at between 25 and 40 dB, otherwise detection will not start. In the future a more intelligent algorithm may be implemented to look for silent parts throughout the entire soundfile before starting the analysis.

Fundamental frequency

The f_o is detected by finding the highest-level peak in the spectrum. The frequency of the highest spectral peak is determined by quadratic interpolation based on the three FFT samples around the peak. This peak is assumed to represent a partial of the signal, and the frequency is divided by 1, 2, 3… until 50 Hz is reached. Each of these new frequencies are considered candidates for fundamental frequency, and the f_o candidate producing the highest total harmonic power is assumed to be the fundamental. Partials with lower level than 30 dB under the highest peak, and less than 10 dB above the background noise floor are disregarded.

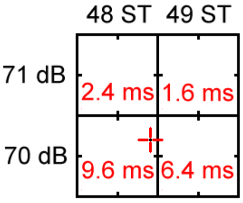
Blocking criteria

As with all methods for f_o measurement, it is difficult to extract stable and reliable data at onset and offset of voicing and during sections where instability or perturbation is high. In phonetography, such difficulties become particularly problematic as these faulty and/or uncertain data points often appear outside the VRP or SRP, quite visible to the eye. Thus, RecVox uses a battery of criteria to block unreliable phonation detections and these criteria are briefly listed here.

- If the energy below 50 Hz is 10 dB higher than the energy between 50 and 2000 Hz, the detection is blocked. The purpose of this is to block frames with spurious low frequency noise that easily travels through ventilation systems, such as sounds from distant doors closing.
- If the energy above 2 kHz is above -15 dB relative to the energy between 50 and 2000 Hz and the SPL is below 70 dB @ 30 cm, the detection is blocked. The purpose is to suppress fricatives and breathing sounds, but still accept loud vowels.
- If the rate of f_o change is larger than 7 semitones per frame, the detection is blocked. Rapid changes in f_o is typically a sign that the detection is erroneous.
- During voice onset and offset the SPL of voice at sudden vocalization stops there is still reverberant energy in the room that might cause a dropping “tail” of detections at the low end of the VRP. Thus, during rapid decays faster than 100 dB/s, detection is blocked. The level drop rate should be greater than the reverberation drop rate of the room at all frequencies, and 100 dB/s is equivalent to reverberation time T_{60}=0.6s, as 60dB/0.6s=100 dB/s .
- The detection is blocked if harmonic-to-noise ratio (HNR) is less than 5 dB.
- Detection is also blocked if the piano keyboard is pressed to play a reference tone.
- If dual microphones are used, the coach microphone can block the participant microphone, and vice versa, thus allowing the coach to sit in the same room as the participant without influencing the phonetogram of the participant.

Binning

The values for f_o (in semitones) and SPL (in dB) are typically not perfect integer numbers, so the duration of a frame (20 ms) is distributed as accumulated time over four bins according to the decimals of the f_o expressed in semitones, and the SPL expressed in dB. Distributing the frame time over the bins like this results in accurate averages of SPL and f_o even for pure, static tones.



Example of binning. For a level 70.2 dB and frequency 48.4 ST (C3 + 40 cent) the 20 ms frame duration would be distributed over four bins; 9.6 ms being added to the (48, 70) bin, 2.4 ms to the (48, 71) bin, 6.4 ms to the (49, 70) bin and 1.6 ms to the (49, 71) bin.

Threshold

A bin of the phonetogram is displayed when the accumulated time at that particular f_o-SPL combination has reached a threshold. However, as the f_o-extraction algorithm sometimes results in faulty values for f_o, the accumulated time must exceed a threshold of 40 ms for that bin to be displayed. The accumulated time includes 25% of the adjacent bins (above, below, to the left and to the right) in order to facilitate clusters of bins to be shown, but to suppress isolated bins.

Third parameter

A third parameter can be displayed as color intensity in the phonetogram. RecVox calculates four such parameters for each frame.

- L_{H1}-L_{H2} is calculated for each frame by measuring the spectral levels at f_o and $2 \cdot f_o$. The peak levels are determined using quadratic interpolation as with the f_o extraction.
- The harmonics-to-noise ratio (HNR) is by dividing the power from all harmonics by the power from the noise between harmonics. The algorithm presently used is not optimal and may be replaced in later versions of RecVox.
- The spectral power centroid is measured as the weighted mean of the powers in the spectrum.
- The maximum partial number is the number of the partial found first during f_o detection, as described above.
- The crest factor is calculated as the level difference between the highest absolute peak value in the frame and the frame power.

If there are multiple hits at the same bin, the third parameter values are averaged and weighed by the time allocated to the bin in the binning process.

Statistics

Statistics for the selected part of the phonetogram is calculated from the contents of the phonetogram bins. For the SPL the equivalent sound pressure level, the maximum and minimum SPL, percentiles and range can be calculated. For the f_o the average, minimum, maximum, median, mode and percentile values can be calculated. The f_o can be expressed in Hz, a semitone number relative to C1 or as a note name (e.g. "C3"). Because of binning, the minimum, maximum and mode f_o values are quantized to whole semitones. Also, the recorded time, phonated time, and the selected area (in STdB) can be calculated. Also, an average of the third parameter can be extracted.

It should be noted that not only the f_o measures, but also the SPL measures refer to the phonated frames, and that silent parts and fricatives are excluded. Thus, while possibly being more representative of the phonation, the SPL values will in most cases differ from measures that include the complete sound signal. In particular, the Leq detected by RecVox is typically 3-6 dB higher than Leq detected for the speech if the voiceless sounds and pauses are included.

Calibration

As RecVox does not have any dedicated hardware, but runs on most standard soundcards, special routines for SPL calibration are necessary. RecVox has two different strategies for calibration. The recommended procedure involves a 94 dB calibrator that is attached to the microphone, and the operator adjusts the volume control of the soundcard so that 94 dB is registered by RecVox. By using this procedure, a maximum peak value of the signal of 126 dB is also set. There is also another calibration procedure using a sound level meter mounted close to the microphone, and a sinusoidal tone played by a loudspeaker. This procedure uses whatever position is set on the soundcard and calibrates RecVox accordingly. The maximum peak level is displayed post calibration.

File formats and normalization

The most common file formats for audio do not include saving the calibration with the signal. Wav files are commonly assumed to have full scale represent the number “1”. This could be used to represent 1 Pa, which in turn would allow for a maximum SPL of 91 dB for a sinusoidal signal. This is not enough to represent the full range of human voice. Therefore, wav files saved from RecVox are considered for listening only and are saved as 16-bit files, scaled so that the maximum peak corresponds to full scale.

To save the calibrated sound the sopran or smp file formats can be used. The smp format is the most common of the two, but only implements 16-bit data. For this format, audio data is scaled to full scale, calibration is saved in the header and is readable for the softwares that handle the smp format. Parameter data is padded to 44100 Hz. The Sopran format saves the audio in 32-bit floating point samples calibrated to pascals and keeps the 44100 and 50 Hz sampling rates for audio and parameters, respectively. The sopran format is presently only used by the Sopran software, also from Tolvan Data, but is the format that preserves the data and analysis best. Sopran can further export the data to several other formats, including calibrated 32-bit floating point wav, importable to for example Matlab.

Finally, the phonetogram bins can be saved in the stdpg file format, which is compatible with Phog and Voice profiler. The audio and the variation over time of parameters is not saved in this format.

Description of validation signals

Validation of RecVox performance was made by generating 80 synthetic signals by adding 10 partials at f_o=110, 220, 440 and 880 Hz, and SPL 40, 60, 80, 100 and 120 dB, and spectral slopes of -6, -9, and -12 dB/octave. Finally, to simulate the background noise of a room, white noise lowpass filtered at 20 Hz with a slope of -6 dB/octave and SPL of 27 dB was added. The signals were generated in Matlab and stored in a calibrated 32-bit floating point wav file. The soundfile was run through RecVox and the result was saved as a sopran file for testing the accuracy of SPL, f_o and L_{H1}-L_{H2} measurements.

Results

The analysis of the synthetic validation signal reveals that RecVox detects SPL with an error no more than 0.04 dB, f_o with an error of no more than 0,13 % and L_{H1}-L_{H2} with an error of no more than 0,17 dB, see table. Clearly, the SPL errors produced for these synthetic, stationary signals are ridiculously small compared to those typically caused by calibration error, microphone distance etc. The maximum error for f_o corresponds to 2.2 cents or 0.022 semitones, which is barely audible in a straight tone. The L_{H1}-L_{H2} deviations are also small, at the most 0,17 dB from the true value.

	SPL deviation dB	**f_o deviation %**	**L_{H1}-L_{H2} deviation dB**
avg	0,00	0,06	-0,08
stdev	0,02	0,05	0,05
max	0,04	0,13	-0,03
min	-0,03	-0,02	-0,17

Not all the signals were however detected as phonation by RecVox. The complete table of f_o detections is shown in the table. The signals marked with a single asterisk * the fundamental is too near the background noise and the signals marked with a double asterisk ** was blocked due to a large high-frequency content.

RecVox takes advantage of the fact that soft voice at high f_o is dominated by the fundamental in real voices and blocks these signals as non-voice.

-6 dB/oct	**110**	**220**	**440**	**880**	**-9 dB/oct**	**110**	**220**	**440**	**880**	**-12 dB/oct**	**110**	**220**	**440**	**880**
40	*	*	**	**	40	110,05	*	440,13	**	40	110,14	220,22	440,13	879,85
60	110,13	220,22	**	**	60	110,13	220,21	440,13	**	60	110,13	220,22	440,13	879,85
80	110,13	220,21	440,13	879,85	80	110,13	220,22	440,13	879,85	80	110,13	220,21	440,13	879,85
100	110,13	220,22	440,13	879,85	100	110,13	220,22	440,13	879,85	100	110,13	220,21	440,13	879,85
120	110,13	220,22	440,13	879,85	120	110,13	220,22	440,13	879,85	120	110,13	220,22	440,13	879,85

Discussion

Analyzing the voice signal in phonetography is more difficult than in most voice analysis applications. In the VRP the full range of the voice is recorded, both in terms of SPL and f_o. The phonetogram software must be able to handle a f_o range from 50 to 2000 Hz (5.3 octaves or 64 semitones) and at SPLs from 40 dB to 120 dB (80 dB range). Many f_o extraction algorithms restrict the f_o and SPL range to improve reliability for f_o extraction of speech signals. The phonetogram algorithms cannot benefit from such limitations.

The great f_o range also calls for compromises with respect to speed of f_o change and lower f_o limit. When using a windowed approach, such as in Recvox, the window length should be long enough to include a reasonable number of glottal cycles in the window, but also short enough to register glissandi without too much f_o variation between the start and end of the window. Practical experiments with RecVox show that there is a rather narrow interval around 40 ms that is appropriate for the window length. If f_o=50 Hz, this allows for only two cycles of the signal to fit in the window but still this results in a surprisingly high precision of f_o extraction due to interpolation of the spectral data. On the other hand, f_o fluctuations such as in tremor or vibrato are in the range of 4-8 Hz in the human voice, which would allow for 3-6 analysis frames within such an f_o fluctuation cycle.

The accuracy of extracted f_o and SPL by RecVox is high over great f_o and SPL ranges. It is likely that this precision is good enough for almost any application. However, the most important, and at the same time least well-described factor regarding fundamental frequency extraction are the criteria for accepting a frame as phonated. There is a fine balance between accepting frames that contain phonation and rejecting frames that do not. In phonetography, this balance should be shifted towards rejecting non-phonated frames, as these often appear outside the voice range profile. A single incorrectly accepted frame may thus result in an isolated dot outside the voice range profile. Such a dot disturbs the appearance of the phonetogram much more than a missing frame inside the VRP. Thus, it is quite important to detect and block such false detections of phonation. On the other hand, the solutions for blocking false detections can involve many choices of the designer or the advanced user that are not easy to describe and have a great impact on the appearance and statistics of the phonetogram. It was a deliberate choice made in RecVox to prioritize function over simplicity, but to provide default settings that mostly do not have to be changed.

Due to the many design choices that exist in phonetography, comparisons between phonetograms generated by different computer programs or even with the same program using different settings is difficult. Measures that rely on extremes are particularly sensitive to settings and errors, such as the minimum and maximum f_o and SPL. This is unfortunate given that they have great intuitive and diagnostic importance, so these difficulties must be handled with good choices of thresholds but also an educated interpretation of the phonetogram.

The dynamic range of the equipment is quite important when recording a voice range profile. The SPL of a voice can range from 40 dB to 120 dB, or a dynamic range of about 80 dB or more is required to be recorded. In addition, the softest signals still need a decent signal-to-noise ratio (SNR) to allow accurate measurements. A 16-bit recording system has a theoretical maximum SNR of about 96 dB, and this may not be sufficient for the purpose of phonetography. Therefore, a 24-bit recording system is preferable, having a theoretical maximum SNR of 144 dB, even though the SNR of available soundcards rarely exceed 120 dB.

Background noise plays an important role for the lower profile of the phonetogram. The softest human voice is about 48 dBC @ 30 cm (5th percentile) and given a background noise of 38 dBC as suggested by Šrámková (2015) this results in only a 10 dB signal to noise ratio. For microphone distances closer to the mouth than 30 cm the SNR is improved. It is easily imagined that different f_o and phonation detection algorithms can give different results under these conditions. Therefore, if the lower profile is of interest, users should be made aware of the importance of a quiet environment, and the possibility of placing the microphone closer to the mouth.

Conclusion

The purpose of this paper is to validate the performance of the real-time phonetography computer program RecVox and to describe the design considerations and document its properties in order to help other software designers to find appropriate design choices.

It was found that RecVox produces very accurate measurements of f_o, SPL and L_{H1}-L_{H2} for synthetic signals, but it was also pointed out that the detection of phonation and thresholding of accumulated time may produce much larger measurement differences in e.g. average f_o in running speech.

Acknowledgements

The author of this article is also the developer of RecVox and Sopran freely available from www.tolvan.com . He was also involved in the early development of the Phog phonetograph. The author has no economic interests to declare. RecVox and Sopran are genuine freeware, without advertisements, data harvesting or similar limitations. The only time RecVox contacts Tolvan Data is when looking for updates.

References

Anna Eva Hallin, Karin Fröst, Eva B. Holmberg & Maria Södersten (2012) Voice and speech range profiles and Voice Handicap Index for males — methodological issues and data, Logopedics Phoniatrics Vocology, 37:2, 47-61, DOI: 10.3109/14015439.2011.607469

Hana Šrámková, Svante Granqvist, Christian T. Herbst, Jan G. Švec; The softest sound levels of the human voice in normal subjects. J. Acoust. Soc. Am. 1 January 2015; 137 (1): 407–418. https://doi.org/10.1121/1.4904538