ARCHIVES OF ACOUSTICS
31, 4 (Supplement), 219–226 (2006)

COMPARATIVE STUDY OF THE SELECTED METHODS OF LARYNGEAL TONE DETERMINATION

W. WSZOŁEK, M. KŁACZYŃSKI

AGH University of Science and Technology
Department of Mechanics and Vibroacoustics
Al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: wwszolek@agh.edu.pl

(*received June 15, 2006; accepted September 30, 2006*)

Laryngeal tone (pitch), understood as fundamental frequency of the voice (F_0), is one of the most important features in the field of acoustic studies of the speech signal. Time dependence of the laryngeal tone contains very essential information, providing a basis for speech recognition and extraction of distinctive features in the voice signal. The parameters of the laryngeal tone also contain some features that are useful in the medical diagnosis and therapy. Detailed knowledge of the laryngeal tone enables also both qualitative and quantitative analysis of speech deformations, related to pathological changes in the larynx area. Due to the progress in professional methods of registration and processing of acoustic signals, several algorithms for determination of laryngeal tone have been developed. Yet an accurate and detailed determination of the laryngeal tone characteristics for a given speaker still presents a problem in the field of speech signal analysis. In the present paper selected methods of F_0 determination have been described and compared.

Key words: speech analysis, pathological speech, speech recognition, pitch determination.

1. Introduction

The process of human speech generation is a complex phenomenon, comprising many topics in psychology, biology, medicine, as well as aerodynamics and acoustics. In a simplified description one can distinguish two basic layers of features that are specific for a given speaking person: the physical layer – originating from the anatomical structure of the vocal tract (the source and filters) and the psychological layer, related to the individual manner of controlling the phonation and articulation organs. In the physical layer it is necessary to distinguish the two stages of speech generation and recognize the possibility of separate definition and measurement of the parameters of source and filter. The speech signal, treated as a time-dependence of acoustic pressure (Fig. 1), exhibits a complicated time-course, reflecting the complex nature of the process of its articulation. The signal parameters are affected by the source (vibrating vocal cords, or the noise of turbulent flow of the air-stream through the straits in speech organs) as well as the dynamical properties of the vocal tract, forming the final structure of the signal.

In the time domain the signal can be mathematically described using a convolution of the original time dependence of the source signal $g(t)$ and the impulse response of the vocal tract $h(t)$:

$$p(t) = \int_{0}^{t} h(t-\tau)\, g(\tau)\, \mathrm{d}\tau. \tag{1}$$

Interpretation of the above-mentioned formula indicates that in the time-dependent acoustic speech signal the properties of the source and the properties of the sound forming voice channel are closely related [8].

Therefore the fundamental problem in the extraction of the source parameters is the unavailability of direct access to the acoustic signal $g(t)$ generated by the source. It should be mentioned that there are non-acoustic methods of estimation of the vocal folds vibrations, e.g. electroglottography – a method based on the measurement of the electrical impedance of the glottis. A serious disadvantage of the method is the necessity to apply a relatively expensive, specialized measuring equipment, and on the other hand its possibly adverse effects, when the method is applied to children or persons after surgical treatment in the larynx area. Considerable progress in professional methods of digital registration and processing of acoustic signal now enables the possibility of extraction of the source-related features from the speech signal. It is assumed that the acoustic parameters of the source contain the individual features of the speaker, originating from his/her individual anatomical features of the phonation organ and the specific manner of controlling the voice pitch modulation (the voice fundamental frequency F_0). The studies of laryngeal tone, carried out by many researchers also confirm, that it is possible to accomplish both qualitative and quantitative analysis of speech deformations, related to larynx impairment. The voice organ pathology related to irregularities in larynx functioning is most frequently visible in changes of the fundamental tone F_0 parameters [6].

The laryngeal tone (excitation for voiced sounds), called the F_0 formant and bearing F_0 parameter as its frequency, is a signal of variable frequency resulting from individual features of the speaker, and a rich spectrum, in which the higher harmonics are attenuated with a slope close to 12 dB/octave. Figure 1 presents the time-dependence of the acoustic speech signal and the plot of vocal fold vibrations during pronunciation of the vowel “a”.

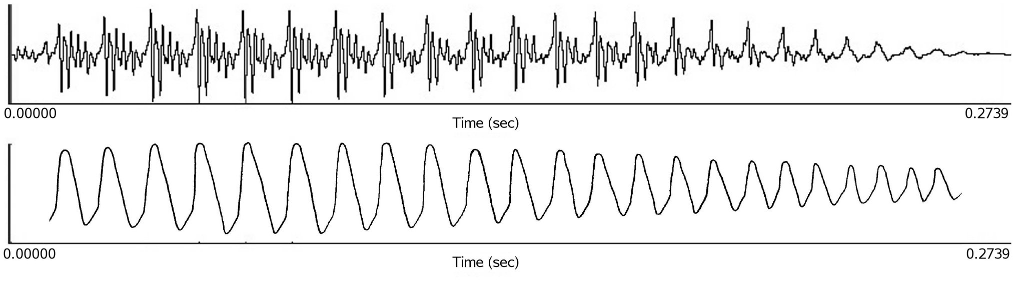

Fig. 1. Acoustic speech signal, vibrations of vocal folds.

Frequencies of the laryngeal tone for population of Poland, depending on the individual speaker's voice type, can be located in the following ranges (Table 1).

Table 1. The ranges of fundamental (laryngeal) tone.

bass	80–320 Hz
baritone	100–400 Hz
tenor	120–480 Hz
alto	160–640 Hz
mezzo-soprano	200–800 Hz
soprano	240–960 Hz

2. Methods of laryngeal tone determination

There are several algorithms dedicated for determination of the laryngeal tone F_0. Most of them is based on the signal analysis in the time or frequency domain. In the present paper the attention has been focused on three methods based on: the zero crossing analysis, the cepstral analysis and the determination of the harmonic to subharmonic frequency ratio.

Zero Crossing Measure is realized in the time domain and is based on determination of the points, where the time-dependent values of the acoustic speech signal $p(t)$ cross the time (ordinance) axis, namely the points t_i, for which $p(t_i) = 0$. In practice, or effectively in the digital signal processing procedures, the ZCM is based on the calculation of the signum function, for n consecutive samples of the examined signal $p(n)$, according to Eq. (2) [3].

$$\rho_0(x,n) = \frac{|\operatorname{sgn}(p(n)) - \operatorname{sgn}(p(n-1))|}{2}, \tag{2}$$

where

$$\operatorname{sgn}(n) = \begin{cases} +1 & \text{when } p(n) \geq 0, \\ +1 & \text{when } p(n) < 0. \end{cases} \tag{3}$$

The main directions of the ZCM analysis are the evaluation of the crossing points density ρ_0 and the analysis of the distribution of time intervals between the consecutive zero crossing points. The knowledge of statistical properties for these distributions of time intervals allows the determination of average frequency of the laryngeal tone and its local values in consecutive periods. However it should be noticed that the registered acoustic speech signal $p(t)$ must be subject to the so-called pre-emphasis operation – in order to separate the low frequency and high frequency components.

A method for F_0 estimation, which is also based on the operations in time and frequency domains, is the cepstral analysis method. It allows the evaluation of mutual relations between the spectral component frequencies, contained in the examined signal. A particular feature of the cepstrum is the possibility of separation in the $p(t)$ signal the components related to the functioning of the sound source from the effects related to the transmission of the vocal tract. After denoting by $E(\omega, m)$ the spectra of sound emission in consecutive m-th time window, the cepstrum of the signal can be represented by Eq. (4) [3].

$$c_e(n, m) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log |E(\omega, m)| e^{j\omega n} \, \mathrm{d}\omega. \tag{4}$$

After denoting by $f_z(\omega, m)$ the power spectrum of the sound source and by $f_t(\omega, m)$ the frequency characteristics of the vocal tract during emission of the respective sound, the emission spectra in consecutive m-th time windows can be presented as in Eq. (5)

$$\log |E(\omega, m)| = \log |f_z(\omega, m)| + \log |f_t(\omega, m)|. \tag{5}$$

Figure 2 presents the result of cepstral analysis for the "a" vowel (see Fig. 1). One can notice clear separation between the effects of sound source functioning and the effects of vocal tract transmittance. Estimation of the laryngeal tone F_0 from the cepstrum calculated for the m-th time window consists of a search for the maximal value in the respective plot in the 70–500 Hz frequency band for males and 160–960 Hz frequency band for females, according to the formula given by Eq. (6). The frequency bands are respectively recalculated onto the discrete cepstrum-frequency domain.

$$\exists (i \in n): \quad \left[i \in \left\langle \frac{f_s}{LF}; \frac{f_s}{HF} \right\rangle \wedge c_e(i, m) = \max(c_e(n, m)) \right], \tag{6}$$

where f_s – sampling frequency used for the speech signal [Hz], HF – 70 Hz for males, 160 Hz for females, LF – 500 Hz for males, 960 Hz for females, $c_e(n, m)$ – cepstrum function for the speech signal emission for the m-th window.

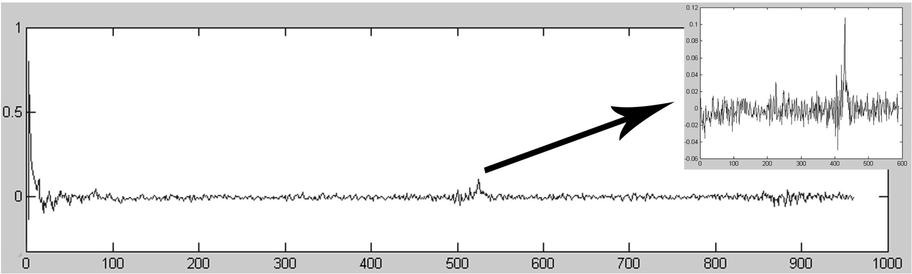

Fig. 2. Cepstrum pitch detection.

It happens sometimes that the algorithms for automatic determination of the fundamental tone makes a mistake, e.g. by doubling the pitch or dividing it by half. One of the reasons for such a mistake is the observed alternation of the amplitude cycles and/or periods in the speech signal. For a regular speech the alternating cycles usually occur in connection with a creaky voice or hoarseness and a low value of F_0, and then the sound is often characterized as a "harsh" voice. For a pathologically deformed speech signal, the extra cycles occur rather frequently. All these phenomena have a direct relation to short-term instability of the vocal chords vibrations. Figure 3 presents the examples discussed above.

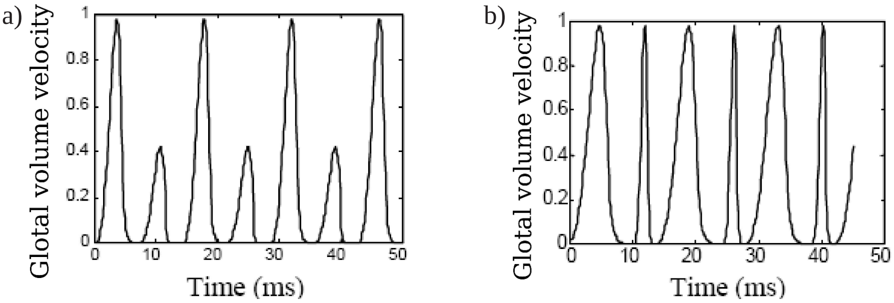

Fig. 3. A schematic representation of glottal pulses with alternating pulse cycle: a) amplitude alternation, b) period alternation.

A method proposed in [9], based on the algorithm for determining the relations between the harmonic and sub-harmonic frequencies, turned out to be very helpful. The low-frequency component is often called a sub-harmonic, which can be in some sense integral with the fundamental frequency F_0 $\left(\text{e.g. } \frac{1}{2}F_0,\ \frac{1}{4}F_0,\ ...,\ \frac{1}{n}F_0\right)$. The relative amplitude change is given by Eq. (7), while the respective frequency change is given by Eq. (8):

$$M_{AM} = \frac{A_i - A_{i+1}}{A_i + A_{i+1}}, \tag{7}$$

$$M_{FM} = \frac{T_i - T_{i+1}}{T_i + T_{i+1}}. \tag{8}$$

Previous research has shown that the determination of the fundamental tone F_0 is closely related to the ratio between the sub-harmonic frequencies and the harmonic ones (SHR – Subharmonic-to-Harmonic Ratio). When the SHR value is high enough (above 0.4), the sub-harmonic frequencies become visible in the speech signal spectrum, and the reception of F_0 leads to values by one octave lower than the original frequency of the vocal folds vibration. This regularity often leads to a statement that the SHR estimation is a good determinant of the laryngeal tone.

The SHR algorithm is based on analysis in the frequency domain with logarithmic scale. For the linear scale the sum of harmonic amplitudes can be presented as Eq. (9), while the sum of sub-harmonic amplitudes as Eq. (10).

$$\text{SH} = \sum_{n=1}^{N} A(nf_0), \tag{9}$$

$$\text{SS} = \sum_{n=1}^{N} A\left(\left(n - \frac{1}{2}\right) f_0\right). \tag{10}$$

For the logarithmic scale Eqs. (9) and (10) take the forms of Eqs. (11) and (12), respectively.

$$\begin{aligned}\mathrm{SH}_{\mathrm{LOG}} &= \sum_{n=1}^{N} A(\log n + \log f_0), && (11)\\ \mathrm{SS}_{\mathrm{LOG}} &= \sum_{n=1}^{N} A\left(\log\left(n-\frac{1}{2}\right)+\log f_0\right). && (12)\end{aligned}$$

For the logarithmic scale the abscissa values, for which the f_0 related values are marked, are shifted, in contrast to the linear scale spectrum, by constant distances, like $\log(2)$, $\log(4)$, $\ldots$ $\log(2N)$. For the whole spectrum, these shifted values are added together and form a sum denoted as Eq. (13):

$$\mathrm{SUMA}(\log f)_{\mathrm{even}} = \sum_{n=1}^{N} A(\log f + \log(2n)). \tag{13}$$

According to Eq. (13) the previous Eq. (11) now takes the form of Eq. (14):

$$\begin{aligned} \mathrm{SH} &= \mathrm{SUMA}(\log(0.5f_0))_{\mathrm{even}} && (14)\\ \text{and} \qquad \mathrm{SH}+\mathrm{SS} &= \mathrm{SUMA}(\log(0.25f_0))_{\mathrm{even}}. && (15)\end{aligned}$$

By doing the same for the sub-harmonic components Eq. (16) is obtained:

$$\begin{aligned} \mathrm{SS} &= \mathrm{SUMA}(\log(0.5f_0))_{\mathrm{odd}}\,, && (16)\\ \text{and} \qquad \Delta &= \mathrm{SUMA}(\log(0.25f_0))_{\mathrm{odd}}\,. && (17)\end{aligned}$$

The difference between $\mathrm{SUMA}_{\mathrm{even}}$ and $\mathrm{SUMA}_{\mathrm{odd}}$ can be presented as:

$$\mathrm{DA}(\log f) = \mathrm{SUMA}(\log f)_{\mathrm{even}} - \mathrm{SUMA}(\log f)_{\mathrm{odd}} \tag{18}$$

and after using the Eqs. (14)–(18) one can determine:

$$\begin{aligned} \mathrm{DA}(\log(0.5f_0)) &= \mathrm{SH}-\mathrm{SS}, && (19)\\ \mathrm{DA}(\log(0.25f_0)) &= \mathrm{SH}+\mathrm{SS}-\Delta. && (20)\end{aligned}$$

For a "regular" speech, when $\mathrm{SS} = 0$, the maximum value of $\mathrm{DA}(*)$ is found for $\log(0.5f_0)$. However for an unstable signal, when sub-harmonics are observed, the maximum value $\mathrm{DA}(*)$ is found for $\log(0.25f_0)$ or for $\Delta \approx 0$, and the next in sequence maximum values is found for $\log(0.5f_0)$. Finding the location of the maximum value $\mathrm{DA}(*)$ is related to the determination of the fundamental tone F_0. The Subharmonic-to-Harmonic ratio (SHR) can be expressed as follows:

$$\mathrm{SHR} \approx 0.5\frac{\mathrm{DA}(\log(0.25f_0) - \mathrm{DA}(\log(0.5f_0))}{\mathrm{DA}(\log(0.25f_0) + \mathrm{DA}(\log(0.5f_0))} = \frac{\mathrm{SS}-0.5\Delta}{\mathrm{SH}-0.5\Delta}\,. \tag{21}$$

The above relation is true when $\mathrm{SS} < \mathrm{SH}$.

When employing the presented algorithm in the first stage one determines the SHR value and then the fundamental tone frequency F_0 – by finding the location of the global maximum denoted as $\log(f_1)$. In the second stage, starting from the previous point the location of the local maximum, denoted as $\log(f_2)$ is determined, knowing that the location must be contained in $[\log(1.75f_1), \log(2.25f_1)]$ interval. Finally Eq. (21) may be presented as:

$$\mathrm{SHR} = 0.5\frac{\mathrm{DA}(\log f_1) - \mathrm{DA}(\log f_2)}{\mathrm{DA}(\log f_1) + \mathrm{DA}(\log f_2)}\,. \tag{22}$$

If the SHR value is less than 0.2, the f_2 frequency is used as the laryngeal tone frequency value F_0, in the opposite case the f_1 frequency is used instead.

3. Summary and conclusions

The algorithms for determination of the laryngeal tone F_0, described in the present paper, have been implemented in the MatLab environment (MathWorks Inc.). The research material used for the present study has been obtained from a group of 45 persons (men and women), exhibiting a correct, but not specially "trained" pronunciation. Each person uttered in the anechoic chamber environment the same phonetic test, which has been also used in the previous studies [2, 6, 10]. The recordings, stored in the computer in the form of time samples, have been used as input data for further processing. From the collected research material the laryngeal (fundamental) tone values have been determined for every word and sound uttered by the research group members, using the algorithms described in the present work. In the final effect comparison has been made between the acoustic methods of F_0 determination and the observed values of vocal chords vibration frequencies (measure by EGG method). The Electroglottograp (EGG) is an instrument for investigating the vibratory characteristics of the vocal folds., which is a precisely and a non-invasive device provides a waveform representation of vocal fold dynamics and relative contact patterns during phonation applied in medicine. The EGG transducers, positioned on the surface of the user's neck at the level of the thyroid cartilage, detect changes in impedance across vocal folds during the vibratory cycle. The output of the Electroglottograph is an EGG waveform (also called the Lx waveform) which shows duration, coordination, and relative contact patterns within a glottal cycle [5] which is F_0 cycle. For each case the comparison has been carried out for the same sample recording. Table 2 presents a summary of comparison results for the maximum errors of average fundamental tone evaluation F_0 for all 45 persons group. The re-

Table 2. Comparison for acoustics F_0 pitch determination to EGG determination methods.

Methods	error [%]
Pitch determination by Zero Crossing Analysis	5.5
Cepstrum pitch determination	4
Pitch determination algorithm based on SHR	2.5

search concludes that a general error of the fundamental frequency determination does not exceed 6% in relation to the F_0 determined by EGG method. The Subharmonic-to-Harmonic Ratio (SHR) method, proposed in [9], definitely provides the best available estimation accuracy.

References

[1] BASZTURA CZ., *Źródła, sygnały i obrazy akustyczne*, WKiŁ, Warszawa 1988.

[2] ENGEL Z., MODRZEJEWSKI M., WSZOŁEK W., *Akustyczna ocena operacji krtani z wykorzystaniem parametrów tonu podstawowego*, Zeszyty Naukowe AGH, Mechanika, **16**, 1 (1997).

[3] DELLER J. R., PROAKIS J. G., HASEN J. H., *Disrete-time processing of speech signals*, Macmillan Publising Company, New York 1993.

[4] HESS W., *Pitch determination of speech signals*, Springer-Verlag Berlin, Heidelberg, New York, Tokyo 1983.

[5] MARASEK K., *Electroglottography description of voice quality*, Phonetic AIMS, Universität Stuttgard 1997.

[6] MODRZEJEWSKI M., *Skuteczność chirurgicznego leczenia chorych na raka krtani piętra głośni*, Rozprawa habilitacyjna, CM UJ, Kraków 1996.

[7] OZIMEK E., *Podstawy teoretyczne analizy widmowej sygnałów*, PWN, Warszawa-Poznań 1985.

[8] TADEUSIEWICZ R., *Sygnał mowy*, WKiŁ, Warszawa 1988.

[9] TITZE I. R., *Principles of voice production*, Prentice-Hall Inc., Englewood Cliffs, New Jersey 1994.

[10] WSZOŁEK W., KŁACZYŃSKI M., *Acoustics methods of voice estimation after surgical treatment of the ocla tract*, Archives of Acoustics, **30**, 4 (Supplement), 193–197 (2005).