THE VOICE SOURCE. MODELS AND PERFORMANCE

Gunnar Fant
Dept of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden

ABSTRACT

This is a summary of research into functional models of the human voice source with considerations to production theory, experimental techniques and individual and contextual variations in connected speech. The emphasis is on work carried out in our department, including the development of a transformed LF-model, and studies of source-tract interaction. The voice source as a prosodic parameter is discussed. Of special interest is the covariation of source parameters, F_0 and inferred contours of lung pressure variations found in focal accentuation.

INTRODUCTION

A major tool for the study of the human voice source is inverse filtering. Over the years a substantial amount of work in this area has been carried out at KTH, see the review in [1].

Inverse filtering is a processing of undressing the vocal tract filter function of the speech wave thus regenerating a replica of the underlying source. This process provides us with some insight in the production mechanism and also a physical substance to be quantified and described within a suitable parameter system.

Early parameter systems concentrated on main shape aspects of glottal flow pulses such as rise time, decay time and open quotient. The importance of the flow discontinuity at closure as an excitation function was early discovered in connection with inverse filtering and was included in a Laplace transform production modeling in 1979 [2]. Five years later the importance of the return phase in the flow derivative was fully acknowledged [3] and became a major constituent of the LF-model [4]. The effective duration of the return phase, T_a, was proved to be inversely proportional to a frequency $F_a=1/2\pi T_a$ where the source spectrum attains an extra -6dB/oct slope. Increasing T_a thus implies a low pass filter effect, a relative attenuation of formants located above Fa. This parameter is usually of greater significance than the main pulse shape parameters.

The ability to capture wave shape essentials has promoted a wide use of the LF model. However, human data from inverse filtering may deviate substantially from model data, and mainly in terms of a superimposed fine structure which displays both typical recurrent patterns and a seemingly randomness. The underlying mechanisms for this structure has been extensively studied in several publications from KTH [1, 5-8].

There exist systematic covariations in the LF parameters which have been exploited in a transformed version [9] of the model. It operates with a fewer number of parameters retaining wave shape essentials, combined with a more detailed specification in terms of deviations of the original LF-parameters from default values. This new system also has advantages from an experimental point of view and as a basis for rule oriented speech analysis and synthesis.

The covariation of source and filter functions, in more general terms phonatory and articulatory processes, is of particular interest. It is the combined gesture rather than the source function alone which has a communicative function. Supraglottal constrictions impede the voice source [10] and glottal abduction introduces additional bandwidths and F_1 increase, subglottal coupling and aspiration noise adding to the source features [7-8, 11-12]]

A specific topic of interest in prosody is the coordination of glottal adjustments, adduction/abduction gestures and F_0-control, and lung pressure. There are apparent differences between singing and speech that need to be studied in greater detail, e.g. vowel consonant contrast, relative emphasis and accentuation.

BASIC SOURCE-FILTER MODEL

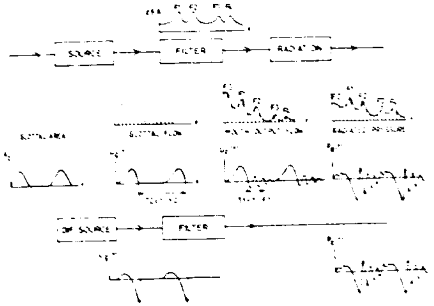

Figure 1. Frequency- and timedomain view of the production of voiced sounds.

The basic concept of source-filter decomposition of voiced sounds in the frequency domain and in the time domain is illustrated in Figure 1. It conveys the traditional view of the source as a raw material of spectral harmonics which is shaped by a filter function. The latter, imposing the formant structure is made up of two parts, the vocal tract transfer function relating the volume velocity flow at the lips to the glottal flow, and a radiation transfer from flow at the lips to the radiated sound pressure wave at some distance from the lips. The radiation transfer is usually approximated by a simple differentiation, in the frequency domain a +6dB octave spectral rise.

In the time domain representation a glottal flow pulse is a skewed version of the glottal areafunction. Glottal parameters are often defined with respect to the time derivative of glottal flow. One advantage of the differentiated source, see the bottom part of the figure, is that it accounts for the radiation transfer component.

Production theory [2,7] states a proportionality between the amplitude of glottal flow derivative at its negative discontinuity, which usually is identical to the negative peak, and formant amplitudes. With an abrupt return to the zeroline and assuming a single formant filter function there is a continuity between the negative peak amplitude E_e and the initial amplitude of the corresponding damped oscillation in the radiated wave This is indicated in the figure. However, the mouth output volume velocity flow, which is the integral of the radiated wave, shows a relative reduction of oscillatory energy but retains the pulse shape of the initial (non-differentiated) glottal flow.

As a matter of fact, integrating the speech wave provides an approximation to the maximum amplitude of glottal flow U_0, constant leakage omitted, while the E_e amplitude information is approximately retained in the envelope contour of the negative side of the radiated speech wave [1, 9, 13] Since U_0 and E_e are the main constituents of glottal waveshape as proposed in the transformed LF model [9], important information about the temporal variation of voice source parameters can be derived without proper inverse filtering.

SELECTIVE INVERSE FILTRING

Inverse filtering experiments confirm these general statements. Figure 2 illustrates regenerated glottal flow and so called selective inverse filtering [1] with cancellation of all formants but one, in this case F1, which appears as a damped oscillation following each glottal flow derivative pulse. The pattern for the [ae] is typical of a sonorous male voice.

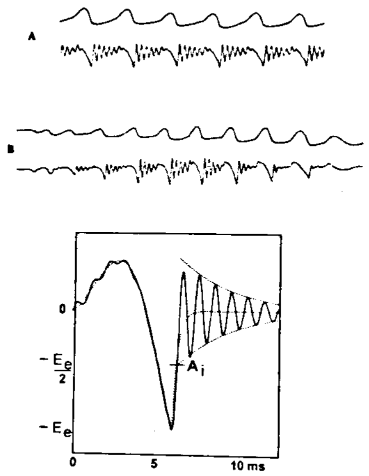

Figure 2. Selective inverse filtering retaining the F1 oscillation. A; vowel [ae]. B; vowel [a] preceding an unvoiced stop. The lower graph is model generated with $F_1/F_a=2$.

The lower pair of curves pertain to a vowel [a] preceding the occlusion of an unvoiced aspirated stop. Within the series of the three final pulses the F1 oscillation decays more rapidly than the supporting flow derivative E_e residue. This is a matter both of increasing damping, i.e. of a formant bandwidth increase following glottal abduction and of the F1 initial amplitude becoming progressively smaller than the negative E_e peak. The latter is the time domain equivalent to an increasing low pass filtering associated with the decreasing F_a which is also a consequence of the abduction gesture. Figure 3 illustrates the principal relation of the F1 initial amplitude being reduced by the same amount as implied by the spectral tilt [1]. Similar effects also appear when a formant is under influence of a neighboring zero, e.g. in a nasalized vowel.

VOCAL TRACT-SOURCE INTERACTION.

More marked instances of pre-occlusive aspiration is treated in [8, 25] In addition to increased spectral tilt and first formant bandwidth, glottal pulse modulated noise appears in the final part of the vowel, in extreme cases combined with pole-zero spectral modifications and extra formants from the subglottal system Noise components are also consistently found in breathy voices [11-12, 16].

A number of other interaction effects complicate the source-filter interpretation of inverse filtering data. One obvious aspect is that a constant leak during the maximally closed glottal interval will pose a problem of how to tune F1 bandwidth and frequency. If these are set for maximal cancellation the inverse filtering will not provide a picture of the true glottal flow. Instead, an ideal regeneration of the true glottal flow would require a setting of the inverse filter to cancel the supraglottal transfer function alone which differs from that of the complete system and can not be derived from the speechwave. The true flow, which has the theoretical burden of conveying the difference between the coupled and the uncoupled system, may have a more complex fine structure than what is seen in ordinary inverse filtering. An example is the appearance of formant oscillations in the maximally closed phase.

A prominent interaction effects is the nonlinearity of the glottal impedance, i.e. the second power dependency of pressure drop on flow, in combination with the presence in the transglottal pressure drop of oscillations evoked from previous excitations [6-7, 15). A typical feature is the double peak appearance of the positive part of the glottal flow derivative and a corresponding spectral dip around $2F_1$ in the source spectrum. [5, 8]

Other aspects of nonlinearities is that a constant glottal chink may counteract the F_a induced spectral fall in the mid and high frequency range of the source spectrum [7, 17, 18] . It is also found [18] that the T_a of the glottal flow derivative becomes larger than an equivalent T_a of the underlying glottal areafunction.

A consequence of glottal impedance nonlinearity is that the superposition imposed by an integer relation between formant frequency and F_0, i.e. when a harmonic hits the formant peak, also effects the driving source function as well as the vocal folds vibratory pattern.. It has indeed been found that the amplitude of F2 and F3 seem to follow the F_1/F_0 ratio rather than the F_2/F_0 and F_3/F_0 ratios [19] An extreme aspect of the nonlinear superposition is that the air consumption is minimized when F_1 hits F_0 but is maximal when F_1 is in the region of 1.5 F_0 which has consequences for soprano singers [6].

A major aspect of vocal tract-source interaction is that a supraglottal narrowing anywhere in the vocal tract or at the lips will be associated with a pressure drop which reduces the transglottal pressure [9-10] and thereby the excitation amplitude E_e and changes the waveshape of glottal flow, increasing the open quotient and the return time T_a. This effect is maximal in voiced plosives and in voiced fricatives but is also noticeable in narrow vowels and in nasals specially in Swedish [1, 13, 16]

VOICE SOURCE MODELING

We shall now return to the more pragmatic aspects of quantifying voice production and source characteristics. In general, irrespective of the particular parameterisation, we may note the close correspondence between the peak value U_o of glottal flow and the amplitude H_1 of the voice fundamental in a harmonic representation of the source component of the speech wave at a distance a cm from the speaker [8].

$$H_1 = U_o k \pi F_0 (\rho / 4 \pi a) \qquad (1)$$

where k is close to 1 for opening qoutients of the order of 0.5-0.7. Adopting the notation $F_a=1/(2\pi T_a)$ where T_a is the effective duration of the return phase we may write the following expression for the amplitude H_m of any harmonic of frequency f_m well above F_0 in the glottal flow derivative spectrum submitted to an extra +6 dB/octave rise with respect to F_0.

$$H_m = (E_e/\pi)(\rho/4\pi a)(1+f_m^2/F_a^2)^{-0.5} \qquad (2)$$

The relative levels of the fundamental and the next two harmonics have to be treated separately by an analysis of the specific glottal pulse shape as in (1). The result is an additional reinforcement, a "glottal formant" located at a mean frequency of $F_g=1/2T_p$ and providing a few dB larger gain than implied by (2)

A consistent mapping of time domain features into the frequency domain allows us to perform an inversion and predict glottal flow shape and magnitudes from absolute calibrated spectral data [7].

An alternative to the Fourier analysis is to decompose the glottal pulse into a sequence of discrete excitation functions [2]. This is necessary for the understanding of the details of observed waveforms and of interaction phenomena. Assuming a single bell shaped glottal pulse with a rising branch of $(U_o/2)(1-\cos 2\pi f_g t)$ and a symmetrical falling branch the flow derivative becomes $U_o \pi f_g \sin 2\pi f_g t$ which is similar to that of the LF-model. The derivative discontinuity at the onset of the rising branch thus contributes with a -12 dB/oct spectrum slope, i.e. -18 dB/oct in the flow domain. Providing the falling branch does not include an additional discontinuity prior to its end it will provide the same excitation function as the rising branch but with opposite phase, and if $T_0=1/F_0=2T_p$, i.e. OQ=1, the net effect in the source domain is a sinewave. This is one extreme condition to be preserved in a parametric scheme. In general, however, formants exited at the onset will be damped out quicker than those at the offset. The major excitation will thus be at the offset even if it does not contain an additional discontinuity. The limiting value of the source spectral tilt is accordingly -12 dB/oct (in the flow derivative) as with an extremely low F_a. On the other hand an abrupt and instantaneous return of the flow derivative at the excitation point T_e provides a spectrum slope of -6dB/oct.

An additional high frequency gain in the source spectrum can be attained only if the duration of the falling branch of the flow is very short, i.e. with an extreme asymmetry and a very small opening quotient, in which case the E_e spike becomes very narrow. This extreme is generally not encountered but can be approached within reasonable limits in simulations.

Occasionally there is to be seen an abrupt step in the flow derivative at the opening phase which adds an excitation of the same type as at closure. This has been taken into account in a modification of the LF model proposed in [23].

THE EXTENDED LF-MODEL

The LF-model [4] is illustrated in Figure 3. We have already discussed the significance of the return phase which accounts for the degree of spectral tilt through the $F_a=1/(2\pi T_a)$ parameter which is frequently used as an alternative to $R_a=T_a/T_0$.

The $R_k=(T_e-T_p)/T_p$ parameter specifies the relative duration of the falling branch from the peak at time T_p to the discontinuity point T_e.

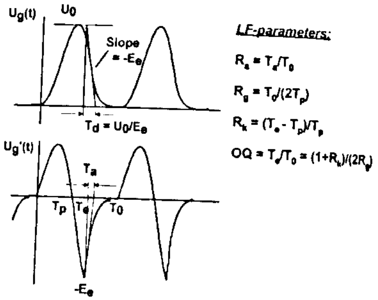

Figure 3. The LF-model. Glottal flow and flow derivative.

The $R_g=T_0/2T_p$ parameter increases with a shortening of the rise time T_p. A large R_g and a small R_k thus produce a small opening quotient, $OQ=(1+R_k)/2R_g$. An alternative common definition is $OQ=(1+R_k)/(2R_g)+R_a$.

Typical values for male vowels are F_a=700 Hz, R_k=0.35, R_g=1.20 and for female vowels F_a=500 Hz, R_k=0.45 and R_g=1. An increase of R_k or a decrease of F_a as in breathy phonation will produce an increase of U_o and thus of the voice fundamental H_1 at constant E_e. An increase of R_g at constant E_e will increase the relative level of the second harmonic of the source spectrum at the expense of a lowered U_o and produces a decrease of OQ which is typical of pressed phonation. A sonorous voice has a relative high F_a of the order of 2000 Hz.

The R_d-parameter

A statistical and functional analysis of covariation of LF-parameters ranging from an extreme tight adducted phonation with low OQ and high F_a to a very breathy abducted phonation with high OQ and low F_a brings out characteristic trends. These can be quantified along a single shape parameter R_d which is closely related to the effective pulse decay time $T_d=U_0/E_e$ (in ms) of the falling branch, see Figure 3.

$$R_d=(U_0/E_e)(F_0/110) \qquad (3)$$

T_d is of the order of 0.5-1 ms for both male and female vowels.

Within a population of vowels and voiced consonants we find a statistical relation:[9]

$$R_a= (-1+4.8R_d)/100 \qquad (4)$$

and

$$R_k= (22.4+11.8R_d)/100 \qquad (5)$$

An important additional finding is that R_d can be estimated from the geometrical constraints of the LF model given the set of R_a,R_k,R_g

$$R_d =(1/0.11)(0.5+1.2R_k)(R_k/4R_g+R_a) \qquad (6)$$

R_g can be derived statistically in the same way as R_a and R_k [9], but a better approach is to calculate R_g from R_d given R_a and R_k. This ensures a conformity with the LF model.

An interesting finding [9] based on female vowel data supplied by Karlsson [20] is that given her full specification of the R_a, R_k and R_g values of a set of nine Swedish vowels these can be predicted with considerable accuracy from R_d alone. This involves the process of first condensing R_a, R_k and R_g values into a single R_d parameter (6) and then applying (4-6).

A conclusion is thus that essentials of the glottal source wave shape may be contained into a single default parameter, $R_d =(U_o/E_e)(F_0/110)$ which is relatively easily accessible from a primitive inverse filtering which has special merits for tracking temporal variations in connected speech. However, for more detailed analysis we need the full set of LF parameter from a proper inverse filtering. Deviations of these from default predicted values can be specified in terms of coefficients $k_a=R_a/R_{ap}$, $k_g=R_g/R_{gp}$ and $k_k=R_k/R_{kp}$ for extra aspiration, press, or flow respectively.

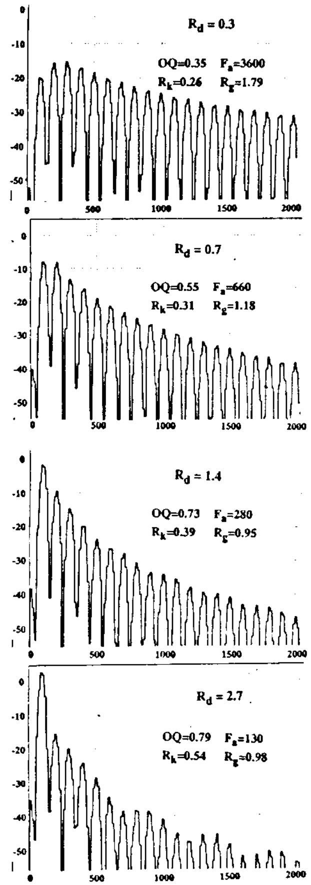

Figure 4. Glottal flow derivative spectra in the frame of R_d values with default LF-parameters included

Glottal flow derivative spectra assuming a constant E_e and F_0=100Hz are shown in Figure 4. The four samples of R_d =0.3, R_d=0.7, R_d=1.4 and R_d=2.7 illustrate the variation from a medially compressed phonation with a small open quotient and a high F_a to a highly abducted, phonation with a large open quotient and a low F_a. Observe the large variation in the ratio H_1/H_2 of voice fundamental to second harmonic amplitude. Females show higher R_d and

H_1/H_2 values than males, and voiced consonants, and aspirated vowels show higher R_d and H_1/H_2 values than regular vowels which is in agreement with earlier findings [11-12].

VOICE SOURCE DYNAMICS

Studies of voice source dynamics, i.e. of temporal variations in connected speech is a developing area which has not yet received the same attention as stationary voice qualities. There remains much to learn about the coordination of glottal adjustments with intonation and lung pressure within a phonetic-linguistic frame.

From our recent work we find systematic covariations of E_e and U_o with F_0. These occur both in glissando sustained phonations of a vowel [15] and in connected speech [9, 13] Both U_o and E_e increase with F_0 up to a maximum or a plateau which is located in the speaker's mid frequency F_0 range, somewhat higher for E_e than for U_o, and E_e increasing more steeply than U_o. Statistical data sampled from prose reading have shown a location of the E_e maximum at around F_0= 100-130 Hz for two males and at F_0=215 Hz for a female voice [13]. In a neutral intonation without focal accents we see clear tendencies of the E_e contours following the general pattern of the F_0 contour. An exception is when in focal accentuation F_0 overshoots a critical value of maximum E_e in which case the temporal contour of E_e and intensity may show a minimum at the F_0 peak with local maxima on both sides. The minimum is not always present. It will be flattened out under the influence of a subglottal pressure rise. Lung pressure is known to increase with F_0 in singing but in speech F_0 operates largely independent of pressure.

INTENSITY VARIATIONS

An important physiological parameter in voice production is the time varying glottal area. $A_g(t)$. At one and the same lung pressure the E_e and the intensity (SPL) increases with A_{gmax} [7, 15] which is capitalized by trained singers [27]. For a more complete understanding of prosodic phenomena we need more data on how A_{gmax} [28] and subglotttal pressure [29] covary with supraglottal and glottal articulations, F_0, SPL, E_e and source spectral shape parameters. Increased lung pressure, and thus subglottal pressure, is found in contrastive and higher degrees of stress but is probably not a necessary component of focal accentuation.

It has long been known that increasing voice effort is associated with a relative emphasis at higher frequencies In an early study [21-22] it was found that a 10 dB increase in the F_1 region was accompanied by about 4 dB increase in the voice fundamental and 14-18 dB in the F_2-F_3 region. This spectral nonlinearity can be interpreted as R_d and R_a decreasing (F_a increasing) with voice effort. Local increments of this magnitude are seldom encountered in speech [26]. The average intensity difference between stressed and unstressed syllables is about 2 dB only and 3 dB with high frequency preemphasis. Twice these values are normally encountered in contrastive stress marking. The intensity parameter has a greater importance as a boundary marker and shows temporal variations similar to those of an accompanying F_0 declination within a phrase.

ACKNOWLEDGEMENTS

This research has been supported by a grant from The Bank of Sweden Tercentenary Foundation and by Telia Promotor Infovox AB.

REFERENCES

[1] Fant, G., (1993), "Some problems in voice source analysis," *Speech Communication* 13, pp. 7-22, 1993.

[2] Fant, G. (1979), "Glottal source and excitation analysis". *STL-QPSR* 1/1979: 85-107

[3] Ananthapadmanabha, T.V. (1984), "Acoustic ananlysis of voice source dynamics", *STL-QPSR* 2-3/1984, pp.1-24.

[4] Fant, G., Liljencrants, J. & Lin, Q. (1985), "A four-parameter model of glottal flow", *STL-QPSR* 4/1985, pp. 1-13.

[5]. Fant, G., (1982), "Preliminaries to the analysis of the human voice source," *STL-QPSR* 4/1982, pp. 1-27.

[6] Fant, G. (1986), "Glottal flow: models and interaction," *J of Phonetics*, 14 Nos (3/4), pp.393-399.

[7] Lin, Q. (1990), "*Speech Production T[heory and Articulatory Speech Synthesis*", Ph.D. Thesis, Dept. Speech Com. and Music Acoust., KTH, Stockholm.

8] Fant, G., & Lin,Q. (1988), "Frequency domain interpretation and derivation of glottal flow parameters", *STL-QPSR* 2-3/1988, pp. 1-21.

[9] Fant, G., Kruckenberg, A., Liljencrants J. & Båvegård, M. (1994), "Voice source parameters in continuous speech. Transformation of LF-parameters", *ICSLP-94*, Yokohama.

[10] Bickley, C.C. and Stevens, K.N. (1986), "Effects of a vocal tract constriction on the glottal source: Experimental and modelling studies", *Journal of Phonetics* 14, pp. 373-382.

[11] Stevens, K.N. & Hanson, M. (1994), "Classification of Glottal Vibration from Acoustic Measurements", in Eds. Osamu Fujimura and Minoru Hirano, *Vocal Fold Physiology 1994,* Singular Publ. Group. pp.147-170.

[12] Klatt, D., & Klatt, L. (1990), "Analysis, synthesis and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am.* 87, pp. 820-857.

[13] Fant, G. and Kruckenberg, A. (1994), "Notes on stress and word accent in Swedish", *Proc. Int. Symp. on Prosody*, Sept. 18 1994, Yokohama. Also published in *STL-QPSR* 2-3 1994, pp.125-144.

[14] Fant, G., & Kruckenberg, A. (1995): "The Voice Source in Prosody", ICPhS 95

[15] Fant, G. (1982) "The voice source. Acoustical modelling",STL-QPSR 4/1982, pp.28-48.

[16] Karlsson, I. & Neovius, L., (1993), "Speech synthesis experimens wit the GLOVE synthesiser, *Eurospeech 93*, pp.925-928.

[17] Cranen, B. and Schroeter, J. (1993), “Modelling a leaky glottis”. *Proc. Dept. of Language and Speech* 16/17, University of Nijmegen. pp 56-64.

[18] Båvegård M., Fant G. (1995), “Interactive voice source modelling". ICPhS-95.

[19] Fant, G., Fintoft, K., Liljencrants, J., Lindblom, B. & Martony, J. (1963), "Formant amplitude measuremets", *J. Acoust. Soc. Am.* 35, pp.1753-1761.

[20] Karlsson, I. (1990) "Voice source dynamics for female speakers,"*Proc. I.C.S.L.P. Kobe*, pp. 69-72, 1990.

[21] Fant, G. (1959), "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", Ericsson Technics No. 1 /1959)

[22] Fant, G. (1980), "Voice source dynamics", *STL-QPSR* 2-3/1980, pp.17-37.[

[23] Schoentgen, J.(1995), "Dynamic models of the glottal pulse", *Levels of Speech Communication: Relations and Interactions,* C. Sorin et al. (Eds) Elsevier Science, B.V. pp.249-266.

[24] Gobl. C. (1988), "Voice source dynamics in connected speech," *STL-QPSR* 1/1988, pp. 123-159.

[25] Gobl, C. "& Ní Chasaide, A. (1988), "The effects of adjacent voiced/voiceless consonants on the vowel voice source: a cross language study", *STL-QPSR* 2-3 1988, pp 23-59.

[26]. Stevens, K.N (1994), "Prosodic influences on glottal waveform: Preliminary data", *Int. Symp. on Prosody,* Sept. 18 1994, Yokohama, pp. 53-63.

[27] Sundberg, J., Titze, I., & Sherer, R. (1993), "Phonatary Control in Male Singing: A Study of the Effects of Subglottal Pressure, Fundamental Frequency, and Mode of Phonation on the Voice Source", *Journal of Voice*, Vol. 7, No. 1, pp. 15-29.

[28] Sundberg, J. (1994) "Vocal fold vibration patterns and phonatory modes", *STL-QPSR* 2-3/1994, pp.69-80.

[29] Sundberg, J., Elliot, N., Gramming,, P., & Nord, L. (1993): "Short-Term Variation of Subglottal Pressure for Expressive Purposes in Singing and Stage Speech: A Preliminary Investigation", *Journal of Voice*, Vol.7, No. 3, pp. 227-234.