INTERACTIVE PHENOMENA IN SPEECH PRODUCTION

GUNNAR FANT

Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH), Box 70014
S-100 44 Stockholm, Sweden

ABSTRACT

A brief overview of interactive phenomena on several levels of speech production modeling has been attempted. Special attention has been devoted to the dependent covariation of phonation and articulation and the implications for a source-filter decomposition of speech. The growing insight in the voicing mechanism and voice source dynamics provides a broader basis for description of segmental as well as prosodic features.

INTERACTION AND THE SPEECH CODE

Speech production processes are inherently interactive in the sense that component parameters and descriptors seldom function independently. Interaction has thereby become a key word in theoretical issues concerning the speech code of almost the same dignity as variability and invariance and is related to these topics.

There exists a large literature on invariance and variability, e.g., the volume edited by Perkell and Klatt /1/. The various standings on these issues seem to reflect consequences of varying definitions and interpretations of terminology rather than true divergencies. I must admit that there exists a similar vagueness in the interpretation of the term interaction which I will use in a rather general sense. One obvious comment on the invariance issue is that we must first accept that phonological transformations and deletions frequently interact with the planning of an utterance, thus accounting for a deviant set of phonological representations in less precise speech. The expected phonemes may simply not be there. Apart from this extreme, there is a continuity of the extent of information-bearing properties in the speech signal ranging from weakly induced traces to the presence of well-defined phonetic segments and cues.

However, it is not fair to refute the invariance issue by reference to either missing or weakly manifested features or to the listeners' complimentary "top down" expectancy. The situation has been neatly summarized by Lindblom /2/ who points to comprehension as the ultimate level of invariance. Personally, I do not favor any specific definition of invariance but I feel it has an important role in the discussion of distinctive features /3/ in their literary sense as constituents of the full speech code. It is also important to point out the relational basis of features. To search for distinctive elements in the speech wave is not a matter of hunting for a very specific golden grain of information that should always be there. It is rather a matter of finding context-biased manifestations of relational contrasts. A feature is just as much a matter of what could have been present in the speech wave as what is actually found.

In practice, we thus have to make up for the frequent lack of direct invariance by resorting to a rule-oriented analysis of variabilities which in the far end preserves an output more or less appropriate for the specific situation with its constraints and demands. Production has to recruit a substantial amount of coordinated interaction within the system to accomplish its complex task.

How do we now define interaction? In a general sense, interaction is an interdependency of constituents of a descriptive system applicable to complex transformations and departures from linear orthogonal relations. A variation of one parameter usually implies a nonlinear influence on the values and variational limits of other co-varying parameters and the extent to which each of the parameters influences the final output.

The many-to-one and one-to-many relations between linguistic and acoustic entities, e.g., relating a sequence of phonemes or a bundle of phonological features to acoustic segments and events and vice versa /4/ has its parallel in transforming from one level of speech production to a previous or to a following one, e.g., from neuromuscular activity to articulatory movements and further on to vocal tract area functions, aerodynamical events and speech wave patterns.

The movement of a single articulator is generally an interplay of several muscle functions displaying synergism or antagonism, with a large allowance for individual variations, combining with other articulatory activity to preserve an adequacy of the final output. Sensory feedback adds to the complexity of interactions, see the contribution of V. Sorokin to this symposium. Let's hope that the now popular "action theory" /5-7/ will find a sufficient close tie to neurophysiology so that we at some stage may transform our present hypothetical generalizations into a more complete insight in actual speech motor behavior.

I have often complained about our lack of speech analysis data. For speech production modeling, the need is even greater. Research has dealt more with tracking of the movements of specific reference points of articulators than a mapping of complete time-varying area functions and aerodynamic states. There remains a great deal of work to map cavity dimensions and speaker specific topologies. We need more insight in the general relations of speech production and speech wave patterns, e.g., with respect to consonants. Cavity-formant relations are complex but these can be handled with appropriate models /8/. The lack of descriptive physiological data remains the bottleneck.

Speech production is a key to the understanding of the speech code. Speech production research is now enjoying a renaissance as a support of speech perception theory and also offers intriguing potentialities for a more natural articulatory-based synthesis. Although a complex of coordinated activities of several articulators may be involved in securing a specific auditory-perceptual effect, the opposite can also be true. What appears to be a complicated set of context-dependent, perceptually interacting segments and cues in the speech wave can often be related to a single production parameter.

An example is the role of a vocal fold adbuction-adduction gesture determining a sequence of associated events in the speech wave of an unvoiced stop including possible aspiration and preaspiration which we may contrast with a voiced stop. Presence or absence of a voice bar, the initial F0 and F1 at relase, F1 cut back and a shorter duration of a preceding vowel are all functions of one and the same underlying glottal gesture. Preaspiration usually terminates the preceding vowel prior to supraglottal closure inducing breathy termination of the vowel which may end with a consonantal occlusion noise.

Other factors than abduction-adduction may contribute. Thus, the phonemic contrast above may contain covarying elements along the tense-lax dimension. Anyhow, this example illustrates that what may seem quite complex in a pure perception-oriented analysis may have a simple correspondence on the production level. Such relations support motor theories of speech perception /9/.

Apart from top-down effects, I would prefer to conceive of speech perception not as a process of looking for equivalent production patterns but rather as involving direct responses to complex auditory patterns which we have learned to associate with linguistic entities. These may not entirely conform to speech-motor patterns, the full equivalence being reached at a higher level of message representation only. An association of auditory patterns to one's own motor capacity could be of importance in the learning stage /10/.

Compensatory modes of articulation have not been studied extensively. Compensation is never complete if we look at fine acoustic details but has to satify perceptual criteria.

The output-oriented function of speech production is often illustrated by the classical bite-block experiment of Lindblom, Lubker and Gay /11/. A speaker aiming at the vowel [i] compensates for an unnatural fixed high jaw opening by raising the tongue to an appropriate position. It is an open question whether the execution relies more on an invariant command for anchoring the tongue blade in a certain contact position than a recalculation of what the tongue has to do with respect to the jaw.

Coarticulation is generally a matter of complex interactions which might obscure the interpretation of spectrographic patterns. Thus, tracking the transition in the release phase of a labial stop, it might be hard to catch the initial delabialization cues and keep them apart from the more slowly progressing main tongue body movements. An insight in production mechanism is apparently at an advantage. Still it is questionable whether this reference also operates in normal speech perception.

Speech output norms vary with the situational demands. Vowel space shrinks in casual style and is expanded in "hyper-speech" modes (ref. /2/). This is analogous to the relation between unstressed and stressed vowels in Swedish /12/. A related observation is that of Zhang Jialu who in his paper for this session reports shifts of formant frequencies and F0 as a function of voice output level.

An adequate theory of prosody must take into account systematic interactions of stress and emphasis with most speech production parameters.

Words within sentence focus display an increased articulatory or "dynamic" contrast whereas unstressed words will be produced with less contrast between successive segments. With emphasis, vowels and unvoiced consonants increase in intensity, whereas voiced consonants display decreased intensity due to more effective constrictions. With emphasis, an otherwise voiced [ɦ] tends to loose its voicing due to a more extreme abduction of the vocal folds, and noise generation takes over. In a destressed position, a voice bar of a stop tends to turn into a semi-vowel with but little contrast to adjacent vowels /13/. Fig. 1 illustrates the degree of contrasts within a Swedish word "behålla", [behɔl:a], uttered in sentence focal position and prefocus. The oscillogram and the voice source excitation parameter display similar contours which bring out the difference in dynamic contrast. We are now engaged in more general studies of how voice source parameters enter prosody.

PHONATORY AND ARTICULTORY INTERACTION. THE HUMAN VOICE SOURCE

A decomposition of the acoustic stage of speech production into a source function and a filter function has a counterpart in the terms phonation and articulation but the correspondence is not perfect. The lack of coherence is in part a matter of terminology, in part a matter of physical interaction.

We may thus speak of laryngeal or glottal articulations as determinants of the voice source as well as of quality changes related to accompanying changes in vocal tract configurations (e.g., a "throaty voice"), or we could imply glottal stops. In connection with Fig. 1, we have already noted that a highly constricted supraglottal articulation impedes the glottal flow which causes apparent changes in glottal pulse shape and intensity /14/. Furthermore, a glottal abduction induced by an [h]-sound or appearing at the boundary towards an unvoiced segment causes changes in formant frequencies and bandwidths in addition to changes in glottal pulse shape all of which contribute to the breathiness or the local aspiration. Thus, both articulatory and phonatory processes may influence the voice source whilst the filter function is determined by articulatory as well as by phonatory adjustments including lung pressure variations. The validity of the last statement, however, may depend on the particular definitions adopted for source and filter. These are not self-evident.

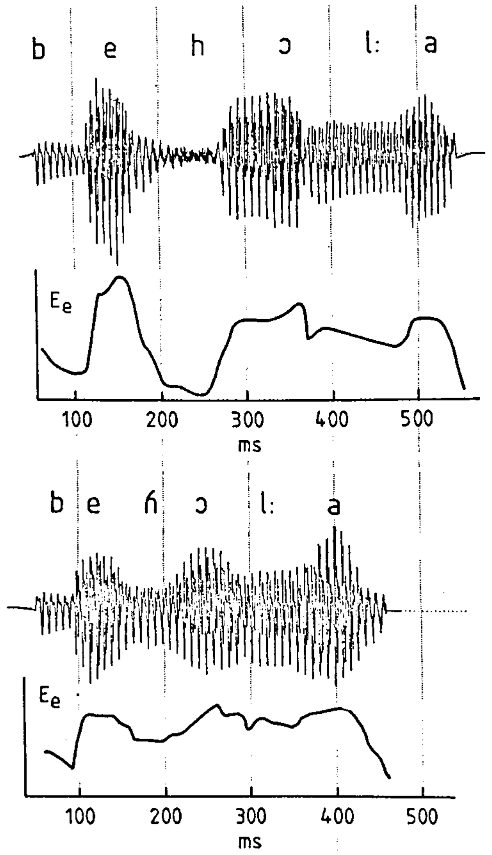

Fig. 1. Oscillogram and equivalent source amplitude E_e of the word "behålla" [beh'ɔl:a]. Above the word in sentence focus, below in prefocus position.

One approach is to refrain from decomposing the speech into a source and a filter function. This is the basis for the Flanagan vocal tract analog which includes a two-mass self-oscillating representation of the vocal folds /15/. It has been a most important and influential tool for simulations.

Over the last eight years, a substantial amount of voice source studies and modeling has been carried out in our department /16-24/. Our recent modeling has been based on a definition of the source as the actual air flow passing through the glottis. The filter function is, accordingly, defined as the supraglottal volume velocity transfer function relating the output flow at the lips, or with radiation included, the sound pressure in front of the speaker to the glottal flow. In inverse filtering, this transfer function is canceled which ensures an output of true glottal flow. It should be observed that the source becomes a property of the entire system just as any flow or pressure within the vocal tract whilst the filter function excludes the glottal and subglottal impedance. Its sole function is to translate from glottal flow to output flow.

A consequence is that the instantaneous resonance frequencies of the whole system may differ somewhat from the corresponding resonance frequencies of the supraglottal system. Also, the rate of formant damping is enhanced during the glottal open period. These circumstances as well as the nonlinearity of glottal impedance and the presence of distributed excitations within the glottal cycle account for a modulation of the instantaneous frequency, phase and damping of formants during the open period. This interaction is usually a second-order effect. However, it puts the burden on the voice source to introduce these modulations in combination with the constant noninteractive filter function. The result is a ripple superimposed on an otherwise smooth glottal pulse shape and the presence of a pattern of distributed zeros in the source spectrum.

These irregularities are especially enhanced by the superposition of formant oscillations from previous voice periods which may occur at a high F0. They enter as components of the instantaneous pressure above and below the vocal folds and thus, to the transglottal pressure drop which has a square-law relation to the resulting flow. This nonlinearity accordingly accounts for an interaction between the existing flow-pressure state and a following excitation.

A more basic instance of vocal tract - source interaction is the tendency of a delay of glottal flow towards the end of the glottal open phase. The main pulse shape is "skewed" to the right in comparison with the profile of the time-varying glottis opening. A consequence is a greater steepness of the flow pulse at closure /25/. This steepness quantified by the maximum flow derivative at closure becomes a scale factor of formant excitations. The larger the negative flow derivative, the larger values the formant amplitudes will be. The maximum glottal flow amplitude (or more precisely, the total volume of the pulse) is a main determinant of low-frequency energy, e.g., the amplitude of the voice fundamental. Increased flow derivative at closure under the condition of constant pulse amplitude thus increases the level of formant amplitudes versus the fundamental. The pulse skewing increases with overall vocal tract inductance, i.e., with the length and inversely with the cross-sectional area of the main vocal tract constriction. Therefore, there is a small difference in inherent voice source strength of vowels. The [i] and [a] and [u] will thus gain about a decibel compared to less constricted vowels (see ref. /19/). These relations can be upset at high F0 values.

All these acoustic interaction phenomena display a seemingly random pattern of perturbations of the voice pulse shape which presumable adds to naturalness /26-27/. They are illustrated in Fig. 2 which shows glottal pulse shapes and spectra under two conditions, the source without any load and with the full load of sub- and supraglottal systems, glottal inductance and viscous resistance included.

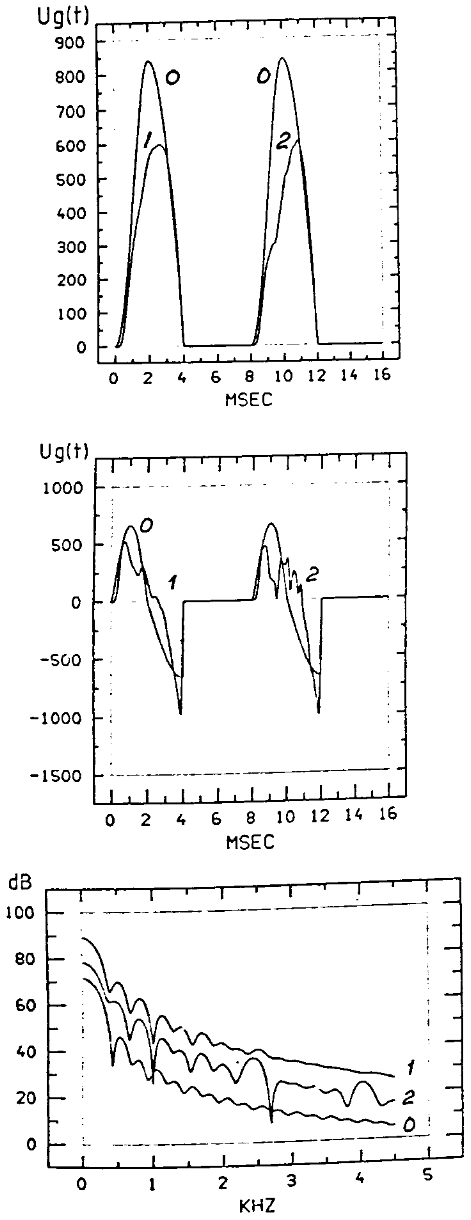

Fig. 2. From top: glottal flow, flow derivative, and flow derivative spectra from modeling of the vowel [i]. 0 stands for source without load, 1 for the first pulse and 2 for the second pulse with load.

The interaction ripple is larger in the second pulse than in the first pulse because of the nonlinear superposition effects. We may also observe spectral cancellation and reinforcement effects in the vincity of F2 and F4 of the vowel [i] which reflects a redistribution of spectral energy to fit the specific source-filter model.

In Fig. 3, pertaining to the vowel [a], we observe a zero between F1 and F2 in the vowel spectrum and an extra peak between F2 and F3 also associated with the nonlinear superposition /28/. The tendency becomes enlarged at large glottal openings and small losses within the vocal tract, and when a formant is much dependent on cavity structures close to the larynx. Figs. 2 and 3 originate from systematic simulations with our model. In true speech, we occasionally observe similar effects of extra peaks between formants which are not related to nasalization. The origin is the nonlinearity element in the source-filter system, see further the contribution of René Carré to this symposium.

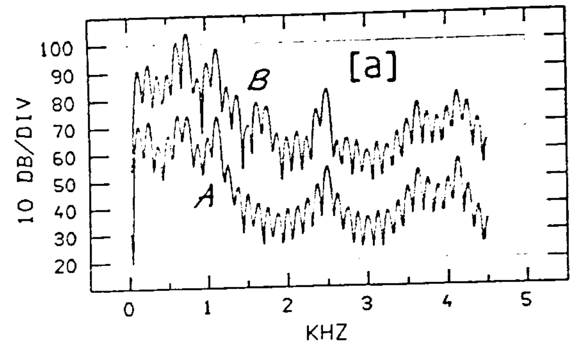

Fig. 3. Modeling of the spectrum of a vowel [a]. A=without interaction, B with acoustic interaction.

Another approach to defining source and filter which relies on approximations suitable for synthesis work is to start out by assuming a specific pulse skewing effect, i.e., the main shape of a glottal pulse which is incorporated in a constant volume velocity source feeding into the vocal tract terminal which is loaded by the glottal impedance /29/. Various alternatives exist such as incorporating the subglottal impedance also or to resort to a time average only of the load.

In formant synthesis one may take into account the variable loading by a modulation of formant bandwidths and frequencies within the glottal open period. This has been successfully exploited by Cheng and Guèrin /30/. However, available experimental data to assess the subjective gain of various solutions is still meager.

Summarizing interaction phenomena in voice production, we have described an acoustic interaction related to the dependency of the excitation mechanism on the instantaneous value of transglottal pressure drop in which components of formant oscillations gain prominence when the impedance of the supraglottal system is comparable to or larger than the glottal impedance. The main objection to selecting the equivalent constant current transformation is the nonlinearity of the glottal impedance. To this acoustic interaction adds the mechanical interaction, related to the change in the aerodynamic forces, affecting the vocal folds as a consequence of a supraglottal constriction which may impede the flow as earlier described and in general causes perturbations of both voice fundamental frequency and flow pulse shapes, see further the contribution of K.N. Stevens who also treats interaction phenomena in the generation of unvoiced sounds.

Acoustic interaction alone can explain an interesting phenomenon in soprany singing. An articulation maintaining F1 close to F0 will not only maximize acoustic output but will also minimize the air consumption (see refs. /22-23/).

At increasing F0 and constant vocal tract filter function, formant amplitudes display periodic amplitude variations, the range of which is lowered by the extra damping associated with interaction. At the same time, the fluctuations of F2 amplitudes are no longer determined by the F2/F0 ratio only and appears to be influenced by the F1 component of transglottal pressure. This is demonstrated in Fig. 4. The full effects observed experimentally by Fant et al. /31/, probably include the vocal fold sound pressure mechanical interaction (see also ref. /20/).

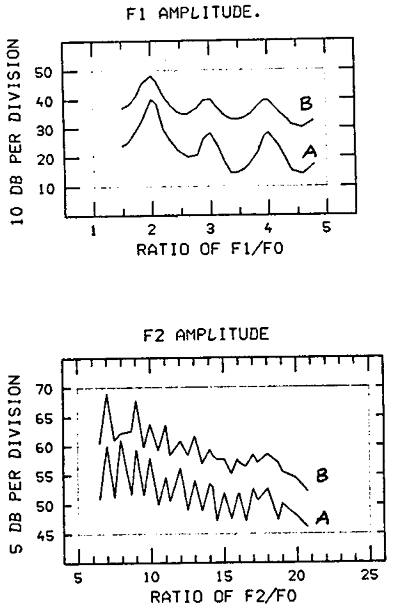

Fig. 4. F1 and F2 amplitude variations as a function of F0 in A: noninteractive, B: with acoustic interaction model. Vowel [ɛ].

More about interactive voice source effects will be reported by René Carré. The Grenoble group has also contributed to other aspects of source filter or rather source vocal tract interactions. One is related to the problem of the origin of inherent F0 of vowels. Guérin and Boe have shown that the aerodynamic forces on the vocal folds tend to lower F0 when F1-F0 is small and positive as could be the case for narrow front vowels /32/.

The subject of inherent pitch has also been treated by Zhang Jialu in his presentation to the congress. He finds that inherent pitch operates in the Chinese language and fairly independent of both distinctive tones and speaker sex. Inherent pitch differences are greatest at high F0. Zhang concludes, as is now generally accepted, that the mechanism of inherent pitch is effected by vocal fold tension passively induced by tongue elevation.

Experiments in France with the Flanagan type vocal tract analog have verified the raise of F0 with subglottal pressure of the order of 2-4 Hz per cm H_2O pressure increase /33-35/. The role of the subglottal system appears to be rather small. It adds slightly to the source excitation parameter and has a quite small effect on F0. Our experiments in Stockholm point to rather small influences of the subglottal system on formant frequencies and bandwidths except when the abduction is relative large and the subglottal pressure low.

There is evidence that an increase of subglottal pressure alone is followed by an approximately square-root dependent increase of maximum vibratory glottal area, see Fig. 8 of Flanagan, Ishizaka and Shipley /36/. Since particle velocity is proportional to the square root of pressure, and the volume velocity is the product of glottal area and particle velocity, it follows that glottal peak flow should increase in direct proportion to subglottal pressure. The accompanying shortening of the pulse base length and the increase of F0 accounts for an additional 3 dB increase in formant amplitudes, i.e., a doubling of subglottal pressure is associated with 9 dB overall spectral level gain (ref. /20/).

In the shift towards a pressed voice, there is an increase of maximum flow derivative at closure and thus of formant amplitudes at constant or even reduced glottal peak flow and a decrease of the open quotient.

We are now engaged in a project of parameterizing the voice source and tracking source parameters in connected speech /ref. 17/. One important parameter is the projection of the initial slope of the return phase on the time axis. This is a measure of the effective duration of the interval from maximum flow discontinuity in the closing branch to complete closure (see refs. /22-23/). This parameter is expecially apparent in breathy phonation. It is associated with reduced excitation and extra formant damping whilst the maximum flow may increase. These studies are also directed to the mapping of individual and of age- and sex-related specifics.

It appears to be fruitful to incorporate voice source parameters as correlates to prosodic categories. Rule-oriented studies are now under way to sort out segmentally induced interactions from underlying prosodic patterns. An example was given in Fig. 1. It is apparent that both prosodic-suprasegmental and inherent-segmental structures are related to all factors of speech production, articulation as well as phonation, and thus source as well as filter functions (ref. /13/).

ACKNOWLEDGMENTS

Christer Gobl has contributed with Fig. 1 and Lin Qiguang with Figs. 2-4.

REFERENCES

/1/ J. Perkell, D. Klatt, Eds., "Invariance and variability of speech processes", Lawrence Erlbaum, New York 1986.

/2/ B. Lindblom, "Phonetic invariance and the adaptive nature of speech", lecture at 30th Ann. of the IPO, Eindhoven, 1987.

/3/ R. Jakobson, G. Fant, M. Halle, "Preliminaries to speech analysis. The distinctive features and their correlates", MIT Press, Cambridge, MA, 7th ed. 1967.

/4/ G. Fant, "Descriptive analysis of the acoustic aspects of speech, Logos 5, 3-17, 1962.

/5/ F.J. Nolan, "The role of action theory in the description of speech production", Linguistics 20, 287-308, 1982.

/6/ B. Lindblom, P. MacNeilage, "Action theory, problems and alternative approaches", J. of Phonetics 14, 29-60, 1986.

/7/ J.A.S. Kelzo, E.L. Saltzman, B. Tuller, "The dynamical perspective on speech production: data and theory", J. of Phonetics 14, 29-60, 1986.

/8/ G. Fant, "The relations between area functions and the acoustic signal", Phonetica 37, 55-86, 1980.

/9/ A.L. Liberman, I.M. Mattingly, "The motor theory of speech perception revised", Haskins Lab., SR-82/83, 63-93, 1985.

/10/ G. Fant, Auditory patterns of speech", in W. Wathen-Dunn, ed., Symp. on models for the perception of speech and visual form 1964, M.I.T. Press, Cambridge, Ma, 1967.

/11/ B. Lindblom, J. Lubker, T. Gay, "Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulation", J. of Phonetics 7, 147-162, 1979.

/12/ G. Fant, U. Stålhammar, I. Karlsson, "Swedish vowels in speech material of various complexity", in G. Fant, ed., Speech communication, Vol. 2, Almqvist & Wiksell Int., Stockholm.

/13/ G. Fant, L. Nord, A. Kruckenberg, "Segmental and prosodic variabilities in connected speech. an applied data-base study", paper XI ICPhS, Tallinn, 1987.

/14/ C.A. Bickley, K.N. Stevens, "Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies", J. of Phonetics 14, 385-392, 1986.

/15/ K. Ishizaka, J.L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell System Techn. J. 51, 1233-1268, 1972.

/16/ G. Fant, "Vocal source analysis - a progress report", STL-QPSR 3-4/1979, 31-54 (KTH, Stockholm).

/17/ G. Fant, "Voice source dynamics", STL-QPSR 2-3/1980, 17-37 (KTH, Stockholm).

/18/ G. Fant, "Preliminaries to the analysis of the human voice source", STL-QPSR 4/1982, 1-27 (KTH, Stockholm).

/19/ T.V. Ananthapadmanabha, G. Fant, "Calculation of true glottal flow and its components", Speech Communication 1, 167-184, 1982.

/20/ G. Fant, T.V. Ananthapadmanabha, "Truncation and superposition", STL-QPSR 2-3/1982, 1-17 (KTH, Stockholm).

/21/ L. Nord, T.V. Ananthapadmanabha, G. Fant, "Signal analysis and perceptual tests of vowel responses with an interactive source filter model", STL-QPSR 2-3/1984, 25-52 (KTH, Stockholm).

/22/ G. Fant, Q. Lin, C. Gobl, "Notes on glottal flow interaction", STL-QPSR 2-3/1985, 21-45 (KTH, Stockholm).

/23/ G. Fant, J. Liljencrants, Q. Lin, "A four-parameter model of glottal flow", STL-QPSR 4/1985, 1-13 (KTH, Stockholm).

/24/ T.V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics", STL-QPSR 2-3/1984, 1-24 (KTH, Stockholm).

/25/ M. Rothenberg, "An interaction model for the voice source", STL-QPSR 1/1981 (KTH, Stockholm), 1-17.

/26/ G. Fant, "Glottal flow: models and interaction", J. of Phonetics 14, 393-400, 1986.

/27/ L. Nord, T.V. Ananthapadmanabha, G. Fant, "Perceptual tests using an interactive source filter model and considerations for synthesis strategies", J. of Phonetics 14, 401-404, 1986.

/28/ Q. Lin, G. Fant, "Complete simulation of voice source - vocal tract interaction", paper, Int.Conf. on Information Processing, China, 1987.

/29/ G. Fant, "The voice source-filter concept in speech production", STL-QPSR 1/1981, 21-37 (KTH, Stockholm).

/30/ Y.M. Cheng, B. Guérin, "Dynamically controlled excitation source for a time-varying formant synthesizer, 2003-2006, ICASSP 86, Tokyo, 1986.

/31/ G. Fant, K. Fintoft, J. Liljencrants, B. Lindblom, J. Mártony, "Formant amplitude measurements", J.Acoust.Soc.Am. 35, 1753-1761, 1963.

/32/ B. Guérin, L.J. Boe, "Etude de l'influence du couplage acoustique source-conduit vocal sur F0 des voyelles orales", Phonetica 37, 169-192, 1980.

/33/ B. Guérin, L.J. Boe, "A two-mass model of the vocal cords: determination of control parameters and their respective consequences", 583-586, IEEE-ICASSP, 1977.

/34/ B. Guérin, D. Degryse, L.J. Boe, "Acoustical consequences of parameters controlling of a vocal cord model coupled with the vocal tract", Report from Symp. on articulatory modelling, Grenoble, 1977.

/35/ B. Guérin, Effects of the source-tract interaction using vocal fold models", in J.R. Titze, R.C. Scherer, eds., Vocal Fold Physiology, The Denver Center for the Performing Arts, Denver, 1985, 482-499.

/36/ J.L. Flanagan, K., Ishizaka, K.L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell System Techn. J. 54, 485-506, 1974.