Methods in
Empirical Prosody Research

Edited by
Stefan Sudhoff, Denisa Lenertová, Roland Meyer,
Sandra Pappert, Petra Augurzky, Ina Mleinek,
Nicole Richter, Johannes Schließer

Offprint

Walter de Gruyter · Berlin · New York

Christophe d'Alessandro (Orsay)

Voice Source Parameters and Prosodic Analysis

1 Voice source and prosody

The most striking acoustic features of prosody are melody and rhythm. But there is much evidence that other acoustic features related to voice quality are also relevant for prosody. This chapter presents a review of prosodic dimensions related to voice quality. Clearly, fundamental frequency (F0) and duration variations in speech are by-products of linguistic and expressive tasks rather than explicitly controlled variations: contrary to music, speech melodies and rhythms are not based on explicit and pre-defined scales and the speaker has no idea of the precise pitches and durations produced. Voice quality variations are also acoustic consequences of intended linguistic and expressive tasks. When considering speech in expressive communication situations (rather than in constrained situations like text reading or other laboratory speech), it is clear that voice quality and articulation variations also play a significant prosodic role, together with melody. One can hypothesize that the analysis of prosody and musical theory (in the Western cultural area) followed the same path of thinking in putting more emphasis on melody and rhythm at the expense of other sound attributes. Although, until recently, prosodic research has focused mainly on F0 and temporal aspects of speech, there is much evidence that voice quality must also be regarded as an inherent dimension of several aspects of prosody (quoting Ní Chasaide and Gobl, 2004, see also Fant, 1995).

The need for considering voice quality emerged in various fields of prosodic research, due to several factors. One factor is the importance of voice quality in the linguistic description of some languages that have been only recently studied with the help of laboratory equipment. Another factor is speech synthesis, a domain where F0 and duration modelling reached a limit in terms of quality and naturalness, as something more seems needed. A third factor is the emergence of research on the processing of expressive speech, for which voice quality variations are much more relevant than for laboratory speech read from written texts.

1.1 Prosody and articulation

In the general sense of suprasegmental speech variation and inflection, prosody involves both the vocal tract, or articulation, and the voice source, or phonation. Following phonetic descriptions of voice quality (see e.g. Laver, 1980; Catford, 1977), both articulatory settings and voice source settings should be considered in prosody. Accentuation, expression and other aspects of prosody may possibly affect the speed, rate, force, and place of articulation. The vocal tract may be dynamically lengthened or shortened, the place of articulation may be more to the back or the front; the lips may be rounded or spread. Then, phenomena usually associated with prosody, for instance accentuation may be marked in speech articulation as well as in speech phonation (De Jong, 1995, Fougeron and Keating, 1997). Phenomena associated with expression, e.g. smiling, may also involve vocal tract articulation rather than the voice source. An account of expressive use of articulation can be found in the pioneering work of Fónagy (1983). Not all of these aspects will be reviewed here, but it is important to keep in mind that, as far as emotion and expression of affects are concerned, the whole body, and particularly the whole vocal apparatus is likely to be involved.

1.2 Prosody and phonation

Prosody is usually associated with phonation. In addition to F0 and segmental durations, a number of additional parameters can characterize the voice source. First of all, the voice may or may not vibrate. This "voicing" parameter is not a binary parameter. Varying degrees of voicing must also be taken into consideration, depending on the degree of periodicity of the glottal vibration. Aperiodicities are important features of the individual voice quality. But they also evolve dynamically during speech for a same speaker. They usually play a role in prosodic variation.

The shape of the glottal flow is another aspect of prosodic variation. This shape can be parameterized according to several glottal flow models. Glottal flow parameters are discussed in detail below. Although there is no general agreement in that matter, I will argue that the glottal flow shape is responsible for the two main perceptual attributes of the voice, loudness and tenseness.

Other striking features of the voice source are voice registers. The vibration mechanisms of the source have several distinct patterns, often referred to as vocal fry, modal voice, falsetto voice, and whistle. The changing of vocal register usually involves an abrupt change of vocal parameter settings, and then of timbre. Register changes are perceptually very salient and can thus be used as expressive means.

1.3 Two functions of prosody

The functions of prosody are manifold. This includes linguistic functions, i.e. functions associated with linguistic aspects such as phonology, lexicon, morphology, syntax, and semantics. Prosody is for instance involved in matters such as sentence and word segmentation, syntactic phrasing, stress, accentuation, phonological distinctions in tone languages. Prosody also features pragmatic and expressive functions. A given sentence in a given context generally expresses much more than its linguistic content (the same sentence, with the same linguistic content may have plenty of different expressive contents or pragmatic meanings). Examples of expressive content are: the identity of the speaker, her/his attitude, mood, ages, sex, sociolinguistic group, and other extralinguistic features. Examples of pragmatic meaning encompass the speaker/listener attitudes (aggressive, submissive, neutral, etc.), the relationships of the speaker and her/his discourse (belief, confidence, assertiveness, etc.), and various other aspects of the specific speech act performed.

Depending on the situation, one aspect of speech or another may dominate a communication act. Sometimes, the "semantic" or "linguistic" content of speech dominates and is almost the sole conveyor of meaning; speech can then resemble its written form, including well formed and predictable prosodic patterns. In other situations, expressive variation dominates, the words per se are less important because it is the "way of pronouncing" an utterance that takes prevalence. Conventions for well-formed prosodic utterances corresponding to written texts have probably little in common with the much more varied situations encountered in expressive prosodic utterances.

1.4 Linguistic functions and the voice source

Intonation patterns, i.e. the melodic and rhythmic aspects of "linguistic prosody", have been described with the help of several phonological and morphological models. The aim is generally to decide what a well formed intonation pattern is and possibly to interpret it. General rules for intonation pattern analysis and synthesis have been proposed, including semantic and pragmatic interpretations. However, there is some evidence that intonation is not the only prosodic domain, even for Indo-European languages. Three aspects of the voice source are more or less explicitly associated with linguistic functions in the prosodic literature: spectral richness, voice tenseness and voice registers.

Variations of spectral richness, related to voice source effort, seem to be important cues for marking accentuation in speech (Sluijter et al., 1996, 1997).

Vocal tenseness seems also to be used as a marker for prominence and phrase boundaries (Epstein, 2003). Analysis of vocal tenseness may help analyse linguistic aspects of prosody such as accentuation and phrasing.

Vocal registers are implicitly incorporated into some intonation models, e.g. the modal/creak distinction. The modal/creak distinction is a change in voice source settings rather than purely a change in F0.

1.5 Expressive functions and the voice source

In phylogenesis, expression has been the first and the main function of vocalization, well before the emergence of speech and language. The expressive functions of prosody have been studied as a part of linguistics in the field of phonostylistics (Léon, 1993). A large research program on the prosodic aspects of emotion in speech has also been conducted (e.g. Scherer, 1986, 2003) and currently seems to attract a lot of attention in speech research with a growing number of studies in "expressive prosody" being conducted.

Recently, automatic speech processing, and particularly speech synthesis renewed interest in this research, because as the sound quality of speech synthesizers increased, their lack of expression became more evident and less acceptable (Eide et al., 2004, d'Alessandro and Doval, 2003).

Several factors have prevented wide-scale investigations of voice quality parameters in prosodic research: while they are generally of secondary importance compared to F0 in read speech and other laboratory speech conditions, they also involve specific and sometimes intricate signal analysis techniques.

The remainder of this paper is organized as follows: In the following section, a typology of phonation settings is proposed, depicting the richness of the basic sound material potentially used for prosodic variation. However, the voice quality dimensions involved in prosody can be reduced to four main axes that are proposed in Section 2. Unfortunately there is no one-to-one correspondence between the four basic voice quality dimensions and basic acoustic parameters. Then in Section 3, voice source models and voice source parameters are reviewed. Voice source parameters are then linked to voice quality dimensions. Section 4 deals with signal processing techniques for voice source parameter analysis as well as for voice quality parameter analysis. The final section presents a conclusion to the paper.

2 Voice quality dimensions and phonation types

2.1 Voice quality dimensions

Phoneticians' first contact with speech stimuli and the first analysis that they perform is usually auditory analysis. The relationships between impressionistic, auditory analyses, signal analyses, physiological analysis, and phonation types are indeed of the utmost importance for experimental phonetics. In the present state of research, it is only possible to form a rough sketch of what these correspondences are. For the sake of prosody research, the rather large amount of more or less independent voice source acoustic parameters (reviewed in Section 3) can be organized using four main prosodic dimensions:

1. the voice register dimension
2. the noise dimension
3. the pressed/lax dimension
4. the effort dimension

Let us review the significance of these four dimensions for speech prosody.

Voice registers depend on the underlying voice mechanisms (as explained latter in this section). Changes of mechanisms may often be used as stylistic features for expressive tasks. The lowest tones of intonation models often correspond to creaky voice, which is a register change.

The noise dimension represents the relative amount of noise in the speech signal. The amount of noise reveals relevant indices for several prosodic phenomena related to breathiness or hoarseness. It is an indication of physical or simulated proximity (for instance in the soft breathy voice used for welcoming customers in vocal server voices), hoarseness is an indication of the emotional state; breathiness is a marker for the end of sentences or phrases (it appears at the end of the so-called “breath group”).

The pressed/tense dimension has also several prosodic functions; in some tonal languages, this dimension is used in tonal distinctions (voice quality tones, or “strangled” tones). Tenseness of the voice is a cue for emotional speech analysis. Tenseness can serve also of stylistic feature (e.g. “popular voices” in French or other languages).

The vocal effort dimension is important for signalling accentuation. Pitch accents are generally considered to be the main correlates of accentuation, but the spectral balance (an acoustic consequence of vocal effort) is another important acoustic accentuation cue. Obviously vocal effort is an important cue for emotions, affect and attitude analyses.

2.2 Sound sources in the voice

The voice is generated by the larynx. Glottal folds, more or less periodically, modulate the airflow exhaled by the lungs (or inhaled, for ingressive voice). According to the position and tension of the larynx musculature, tissues and cartilages, the resulting modulated airflow (the glottal flow) exhibits various acoustic properties and results in various phonation types.

Several phonation types have been described in the phonetic literature. Unfortunately a general consensus on the number and description of these types can hardly be found (Laver, 1980, Catford, 1977). The three main sources of sound in the larynx are:

1. vocal fold vibration (voiced speech)
2. turbulent noise produced through open vocal folds (unvoiced speech)
3. ventricular band vibrations (ventricular speech)

In languages like German or French, ventricular voice is usually an indication of some sort of vocal pathology, or of an aged voice. It can be used in some expressive vocalization like shouting or crying. Creaky voice can be produced by ventricular phonation. These three basic sound production mechanisms can combine and form mixed phonation types.

2.3 Voice registers

As for voiced speech, variations in tension and mass of the vibrating part of the folds give birth to different voice mechanisms (or voice source registers). This is well known and well documented in the singing voice literature, albeit almost never mentioned in the phonetic and speech literature. Four voice mechanisms have been identified, according to physiological measurement of vocal muscle activity (Roubeau et al., 1987, 1997). Creaky voice corresponds to mechanism 0: thick and heavy vocal folds vibrate at a rate of about 10-20 Hz, due to a very relaxed vocal musculature. Modal or chest voice corresponds to mechanism 1: thick and heavy vocal folds vibrate along their whole length. Head or falsetto voice corresponds to mechanism 2: lighter and thin vocal folds vibrate only along their anterior length, due to the tension of the vocal musculature. The whistle register corresponds to mechanism 3, a very high voice that is not used in speech. Changes in voice mechanisms result usually in F0 jumps, from modal to creaky voice, and from modal to falsetto voice. Changes of mechanisms may have linguistic and expressive functions.

Register changes are important indications for expressive speech. For adult male voices, the chest register (mechanism 1) is usually associated with normal voice. The usage of falsetto "lighter" voice quality (mechanism 2) may be interpreted as an indication of emotional state change, or an indication of a specific (e.g. submissive) attitude. The usage of creaky voice (mechanism 0) is sometimes integrated into prosodic models (as an infra-low pitch register). It may also be a marker of expression. For adult female voices, both "chest" or "head" registers can be used, depending on individual habits, but register changes are likely to serve also as an indication of some nuance in speech.

In professional voice training, the art of the singer or the speaker is to control his voice, and particularly control and smooth register changes. Voice register changes can be masked by the appropriate articulatory compensation of specific vocal gestures.

2.4 Aperiodicities

Different forms of aperiodicities appear in different phonation types. In ventricular phonation or hoarse voice, the random perturbation of the voice source is mainly due to irregular vibratory patterns. This type of perturbation is often called "structural noise" because the vibration of the voice apparatus is globally affected. A second class of perturbation is the noise produced by a turbulent flow at a constriction in the glottis. The two main physical situations encountered in speech production are thus:

1. Additive noises. This source of aperiodicity represents continuous noise added to the periodic component. This noise is produced by a turbulent flow at the glottal constriction. This situation is encountered in whispered speech (narrow glottis constriction), or in breathy phonation (larger glottis opening). The vocal cord vibration in breathy vowels modulates the turbulent flow generated at the glottal constriction.
2. Structural noises. The noises produced by the random modulation of the amplitude, period and shape of the glottal waveform from period to period are known as structural noises (Klingholz, 1987). This type of noise is linked to structural perturbations of the vocal fold vibration, and is not due to an additional sound source:
   a. Jitter: This is a random fluctuation of the duration of fundamental periods;
   b. Shimmer: This is a random fluctuation of amplitude for successive periods.

In non-pathological voices, additive noise obviously represents the main contribution to the aperiodic component since the structural noises are usually low.

2.5 Lax-tense dimension

The vocal folds can be pressed together more or less strongly at their posterior extremities (arytenoids cartilages). This posterior movement of the folds is used to narrow the glottis, which is the open space between the folds. When the folds are close together, or pressed together by this posterior constriction, the voice quality is pressed (sometimes called "tense" or "sharp"). Note that this pressed quality may be relatively independent of the vocal effort. On the opposite, if the arytenoids are separated, a chink is created at the posterior part of the glottis. The resulting voice quality is lax. For unvoiced speech, the degree of posterior constriction differs in breathy voice (which is lax) and whispered voice (which is more pressed).

2.6 Vocal effort dimension

An important phonetic correlative of stress and accentuation is loudness. In terms of voice quality, loudness results from vocal effort or vocal force. The exact mechanisms for increasing vocal effort are still partly unknown. Tension and stiffness of the vocal folds can be increased with the action of the extrinsic vocal musculature on the cryco-thyroid cartilages and by contraction of the intrinsic vocal musculature (vocalis muscle). Sub-glottal pressure must be increased. Some adjustment of the voice source and the vocal tract may also play an important role in vocal effort changes. Local variation of vocal effort is an important feature of stress and accentuation. Of course, vocal effort is an extensively used expressive feature.

2.7 Summary of phonation types and voice quality dimensions

A summary of phonation types is given in Table 1. All these phonation types are potentially useful for prosodic analysis, because every difference in sound is significant at the suprasegmental level.

Phonation type	**Description**	**Production**
	Ventricular phonation	
Ventricular	A harsh quality, with a lot of aperiodicities, low F0	Produced between the ventricular bands, or "false vocal folds"
Ventricular creak	Very low frequency, periodic air pulses	Ventricular bands vibration, low sub-glottal pressure, low mean flow
	Voice registers	
Creak	Very low frequency, periodic air pulses	Mechanism 0 of vocal folds vibration. Thick and heavy vocal folds, low sub-glottal pressure, low mean flow
Modal	Usual voice for most males and low-pitched females, low to medium F0 register.	Mechanism 1 of vocal folds vibration. Thick and heavy vocal folds vibrating along their whole lengths
Falsetto	Usual voice for high pitched females, high F0 register	Mechanism 2 of vocal fold vibration. Thin and light vocal folds, vibrating along about 2/3 of their anterior lengths

Aperiodicities		
Breath phonation	Unvoiced speech	Glottis wide open, high mean flow
Breathy voice	A mixture of breath and voice	Incomplete folds closure. High mean flow. Glottal chink
Whisper	Unvoiced speech	Narrowed opening compared to breath phonation, low mean flow
Whispery voice	A mixture of whisper and voice	Incomplete folds closure. Low mean flow. Narrow glottal chink.
Hoarse voice	Irregular, rough quality	A voice with structural aperiodicities, jitter or shimmer
Multiphony	A voice with multiple F0 and/or sub-harmonics	Dissymmetric vibration of the vocal folds, or combination of ventricular and voiced vibrations
Lax-tense dimension		
Tense	A hard or sharp quality, audible glottal formant	Adduction of the posterior part of vocal folds
Lax	A relaxed, soft voice quality	Abduction of the posterior part of vocal folds
Vocal effort dimension		
Loud	A strong voice, with much vocal force	High sub-glottal pressure, high tension of the vocal folds, moderate flow, high voicing amplitude
Flow voice	A strong voice, with high amplitude of voicing and flow.	Normal sub-glottal pressure, tension of the vocal folds, high flow, high voicing amplitude
Weak	A weak voice, without vocal force	Low sub-glottal pressure, low tension of the vocal folds, low flow, low voicing amplitude

Table 1: Phonation types and voice quality dimensions

Ideally, a voice analysis method should be able to deal with the perceived voice qualities or physiological descriptions that have been defined and are actually used by voice professionals. Unfortunately, our knowledge about the acoustic correlates of perceived or physiological qualities is still rather shallow. It seems difficult to find simple acoustic correlates for seemingly simple concepts like voice registers or vocal effort. A more realistic goal for signal analysis methods would thus be to process voice source parameters such as F0, periodic/aperiodic ratio in the voice source, and open quotient of the glottal signal. Voice quality dimensions are then analysed with the help of the voice source model parameters. It is therefore necessary to review glottal flow models in some detail.

3 Voice source models

Voice quality dimensions were proposed in the previous section. Unfortunately, these dimensions cannot directly be extracted from speech. An intermediate level between speech and voice quality dimensions is needed. This level is the level of voice source models. First, glottal flow models are presented, followed by spectral models of the voice source and models of the aperiodic component.

At the end of this section, Table 2 summarizes the relationships between voice source parameters (i.e. parameters of models of the voice source) and voice quality dimensions.

3.1 Time domain glottal flow models

In most acoustic models of speech production, the effect of the voice source is represented by the acoustic flow passing through the glottis, which varies over time. When the vocal folds regularly oscillate (voiced speech), the glottal flow can be represented using a glottal flow model. Several glottal flow models have been proposed so far (Fant et al. 1985, Klatt and Klatt, 1990, Rosenberg et al. 1971, Fujisaki et al., 1986), the most widely used being the Liljencrants-Fant (LF) model (Fant et al., 1985). The glottal flow is the air stream moving from the lungs through the trachea and pulsed by the glottal vibration. All the glottal flow models are indeed pulse like, positive (except in the case of ingressive speech), quasi-periodic, continuous, and differentiable (except at closure). The acoustic radiation of speech at the mouth opening can be approximated as a derivation of the glottal flow. Therefore, the glottal flow derivative is often studied in place of the glottal flow itself.

The form of the glottal flow derivative can often be recognized in the speech waveform, with additional formant ripples. The time-domain glottal flow models can be described by equivalent sets of 5 parameters (Doval and d'Alessandro, 1999). An example of such a set is given by:

- Av: peak amplitude of the glottal flow, or amplitude of voicing;
- T0: fundamental period (inverse of F0);
- Oq: open quotient, defined as the ratio between the glottal open time and the fundamental period. This quotient also defines the glottal closure instant at time Oq*T0;
- Am: asymmetry coefficient, defined as the ratio between the flow opening time and the open time. This quotient also defines the in-

stant Tm of maximum glottal flow, relative to T0 and Oq (Tm = Am*Oq*T0). Another equivalent parameter is the speed quotient Sq, defined as the ratio between opening and closing times, Am = Sq / (1 +Sq);

- Qa: the return phase quotient, defined as the ratio between the effective return phase duration (i.e. the duration between the glottal closure instant, and effective closure), and the closed phase duration. In case of abrupt closure Qa = 0.

The peak amplitude of the glottal flow derivative is in most cases negative. This is because the closing phase is usually shorter than the opening phase. Then its slope is steeper, and its derivative larger. All time-domain parameters are equivalent for the glottal flow and its derivative except the following amplitude parameter:

- E: peak amplitude of the derivative, or maximum closure speed of the glottal flow. Note that E is situated at Oq*T0, or glottal closure instant. It is often assumed that E represents the maximum acoustic excitation of the vocal tract.

Figure 1 gives an example of glottal flow and its derivative (Henrich et al., 2001).

There is no one-to-one correspondence between voice quality and glottal flow parameters. Their relationships are the subject of a large body of work (see e.g. Childers, 1991; Hanson, 1997; Hanson and Chuang, 1999; Klatt and Klatt, 1990; Gobl, 2003), and can be sketched as follows:

- F0 describes melody, and is important for the register dimension. A very low F0 generally signals creaky voice and a high F0 generally signals falsetto voice.
- Oq describes mainly the lax-tense dimension. Oq is close to 1 for a lax voice, and may be as low as 0.3 for very pressed or tense phonation. Qa also correlates with the effort dimension. When Qa = 0 the vocal cords close abruptly. Then E is generally large, as is the asymmetry Am. Vocal effort is high. Conversely, large values of Qa (0.05-0.2) give birth to a smooth glottal closure. Vocal effort is low.
- Av represents the maximum flow, and is an indication of flow voice; it may help for analysis of the vocal effort dimension.
- E correlates well with sound intensity (SPL: Sound Pressure Level, Gauffin and Sundberg, 1989).
- The asymmetry coefficient Am has an effect on both the lax-tense dimension (asymmetry is close to 0.5 for a lax voice, and higher

for a tense voice) and the vocal effort dimension (asymmetry generally increases when the vocal effort increases).

The number of glottal flow parameters seems larger than the number of perceptive dimensions described by these parameters, whereby some parameters should be linked. A global parameter Aq (amplitude quotient), defined as the ratio of Av and E: Aq = Av/E, has been proposed (Alku and Vilkman, 1996). This quotient has a number of advantages in terms of estimation. The authors argue that a good correlation of this parameter with the lax-tense dimension can be found in their data. On the basis of a large set of measurements for male and female voices, a similar parameter Rd = Av F0/ 100 E = Aq F0/E has been proposed (Fant, 1995). The difference is that normalization by F0 is added. This parameter seems to be able to describe vocal tension for a wide variety of conditions.

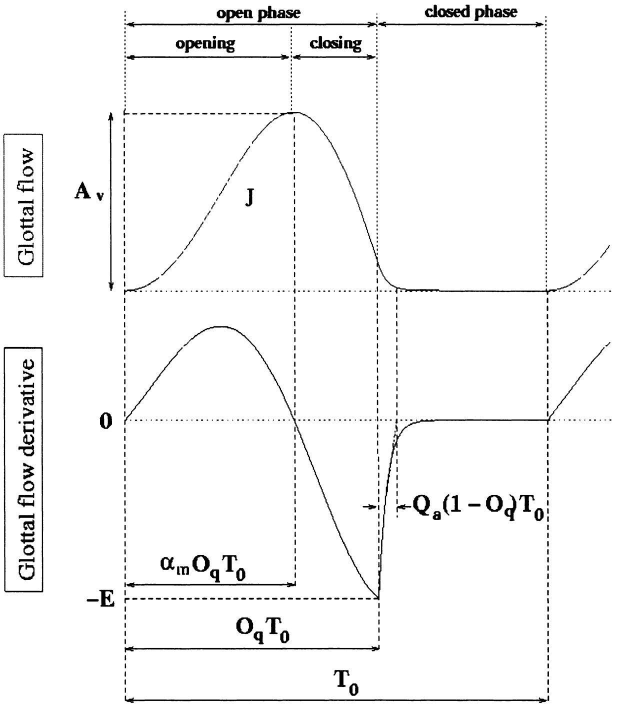

Figure 1: Glottal flow and glottal flow derivative (Henrich et al., 2001)

3.2 Voice source spectrum

Spectral glottal flow models present another point of view on glottal activity. These models are useful because the spectral description of sound is closer to auditory perception (we speak in terms of frequency rather than in terms of periods when describing a sound). For instance, vocal effort is mainly perceived as spectral variations (its effects in the time domain are more difficult to analyse). Voice tenseness is also easily described in terms of thẹ amplitude of the first harmonics. Of course, time-domain and frequency-domain descriptions of the glottal flow are two sides of a same coin, and are mathematically equivalent.

The voice source in the spectral domain can be approximated by a low-pass system. It means that the energy of the voice source is mainly concentrated in low frequencies (recall that only frequencies below 3.5 kHz were used in wired phones) and decreases rapidly when frequency increases. The spectral slope, or spectral tilt, in the radiated speech spectrum (which is strongly related to the source derivative) is, at most, -6 dB/octave for high frequencies. As this slope is of +6 dB/octave at frequency 0, the overall shape of the spectrum is a broad spectral peak. This peak has a maximum, rather similar in shape to vocal tract resonance peaks (but different in nature). This peak is called the "glottal formant". This formant is often noticeable in speech spectrograms, where it is referred at as the "voice bar", or glottal formant below the first vocal tract formant.

Spectral properties of the source can then be studied in terms of properties of this glottal formant. These properties are: 1. the position of the glottal formant relative to F0; 2. the width of the glottal formant; 3. the high frequency slope of the glottal formant, or "spectral tilt"; 4. the amplitude of the glottal formant.

One can show that the frequency of the glottal formant is inversely proportional to the open quotient Oq (Doval and d'Alessandro, 1997). It means that the glottal formant is low for a lax voice, with a high open quotient. Conversely, a tense voice has a high glottal formant, because the open quotient is low.

The glottal formant amplitude is directly proportional to the voicing amplitude. The width of the glottal formant is linked to the asymmetry of the glottal waveform. The relation is not simple, but one can assume that a symmetric waveform (a low Sq) results is a narrower and lower glottal formant. Conversely, a higher asymmetry results in a broader and higher glottal formant.

The glottal formant is located slightly below or close to the first harmonic (H1 = F0) for typical values of the asymmetry coefficient (e.g. 2/3) and open quotient (e.g. between 0.5 and 1). It can be much higher, and reach, for instance, the fourth harmonic for Oq = 0.4 and Am = 0.9.

Up to now, we have assumed an abrupt closure of the vocal folds. A positive Qa in time domain produces a smooth glottal fold closure. In the spectral domain, the effect of a smooth closure is to increase spectral tilt. The cut-off frequency of this additional attenuation is inversely proportional to Qa. For a low Qa, attenuation affects only high frequencies, because the corresponding cut-off frequency is high. For a high Qa, the cut-off frequency is located lower in the spectrum.

Glottal flow models can be considered to be low-pass filters. Glottal flow derivatives can then be considered as band-pass filters. The source spectrum can be stylized as 3 linear segments with slopes of +6dB/octave, -6dB/octave and -12dB/octave (or sometimes -18dB/oct) respectively. The two spectrum breakpoints correspond to the glottal spectral peak and the spectral tilt cut-off frequency.

An example of displaying linear stylization of the glottal spectrum envelope in a log representation is given in Figure 2, top panel. In this figure the stylized envelope is applied to the spectrum of a French vowel /a/ for a male voice. Note that the glottal formant is clearly visible for an /a/ because the first formant is high. The situation is less clear when the first formant is lower.

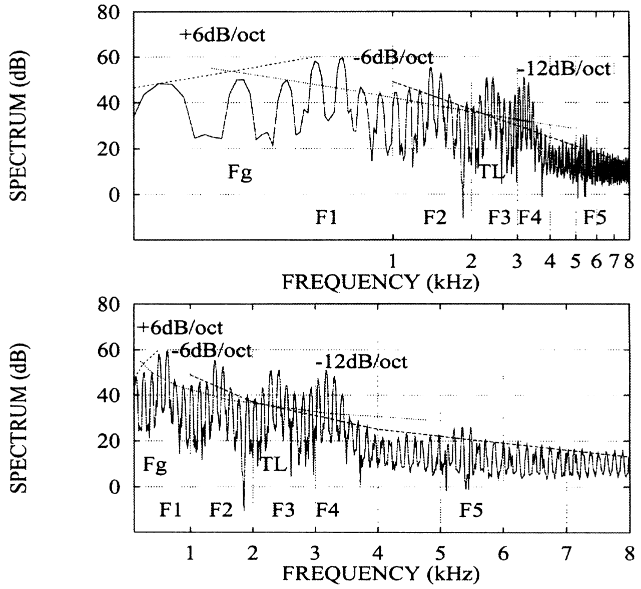

Figure 2: Spectrum of a vowel. The spectral slopes due to the voice source are indicated (d'Alessandro and Doval, 2003).

This global spectral description of the source spectrum shows that the two main effects of the source affect the two sides of the frequency axis. The low-frequency effect of the source, related to the lax-tense dimension is often described in terms of the first harmonic amplitudes H1 and H2 (Hanson, 1997; Klatt and Klatt, 1990) or in terms of the low frequency spectral envelope (Alku and Vilkman, 1997). A pressed voice has a higher H2 compared to H1, and conversely a lax voice has a higher H1 compared to H2. The effort dimension is often described in terms of spectral tilt (ST). A louder voice has a lower spectral tilt, and spectral tilt increases when loudness decreases (Fant and Lin, 1988; Hanson, 1997).

3.3 Aperiodicities

Purely periodic speech is no the most common situation in real-world corpora. Variations of the degree of noise in the speech signal are important indices for prosody. A degree of voicing should be introduced rather than a binary decision between voiced and unvoiced speech. Speech can be unvoiced, strongly voiced (with no or little aperiodicities) or weakly voiced (with more aperiodicities). The signal resulting from the quasi-periodic vibration of the vocal cords is referred to as the "periodic component" or "harmonic component" of speech, and the sound resulting from aperiodic excitation is referred to as the "aperiodic component" or "noise component" of speech.

It is acknowledged that the energy ratio of the periodic and aperiodic components is a useful parameter in quantifying voice hoarseness or breathiness (Kojima et al., 1980; Hillenbrand, 1987). This can be extended to prosodic variations as well.

In the spectral domain, additive aperiodicities generally dominate the higher part of the spectrum above a given frequency. Some models make use of a frequency position, the limit of voicing (Laroche et al., 1993), which defines the point in the spectrum where the aperiodic component dominates the periodic component (LoV). The spectral envelope of the aperiodic component or at least its spectral tilt is also an important characteristic of the voice (noise spectral tilt NTL). The effect of structural aperiodicities in the spectral domain is different. It results in a broadening of voiced harmonics rather than noise between the harmonics. In summary, several parameters can be helpful for characterising aperiodicities. They are:

1. jitter (as a percentage of F0)
2. shimmer (as an amplitude variation in dB)
3. PAPR: periodic to aperiodic energy ratio, or signal to noise energy ratio
4. noise spectral envelope, or noise spectral tilt
5. limit of voicing

In the spectrum, additive noise appears mainly between harmonics. A measure of inter-harmonic noise (IHN) has also been proposed (Childers and Lee, 1991).

Table 2 summarizes the acoustic parameters used for modelling the voice source (in time domain and spectral domain). The main effects of each parameter on phonation are described. Some parameters are almost equivalent, because of the duality between time domain and frequency domains (e.g. Qa and Fa) or because they represent the same underlying dimension (e.g. Av and E represent the amplitude of voicing). These correspondences are indicated in Table 2 in the column "Duality".

Parameter	**Description**	**Duality**	**Main effect on phonation**
	Time domain parameters		
Av	Amplitude of voicing	E, Ags	Flow
Oq	Open quotient	Fg	Tenseness
Am	Asymmetry	Bg, Sq	Tenseness, loudness
Qa	Return phase	Fa	Loudness
	Alternative time domain parameters		
E	Derivative peak	Av	Flow, loudness
SPL	Sound pressure level	Av, E	Flow, loudness
Sq	Speed quotient	Bg, Am	Tenseness, loudness
Rd	Amplitude quotient (Fant)	AV, E, F0	Loudness, tenseness
Aq	Amplitude quotient (Alku)	AV, E	Loudness, tenseness
	Spectral parameters		
Fg	Glottal formant frequency	Oq	Tenseness
Bg	Glottal formant bandwidth	Sq, Am	Tenseness, loudness
Fa	Spectral tilt frequency	Qa, Tl	Loudness
Ags	Glottal formant amplitude	Av, SPL	Flow
	Alternative spectral parameters		
H1*-H2*	1st and 2nd Harmonic amplitude differences	Oq, Am	Tenseness
H1*-F3*	1st harmonic to 3rd formant amplitude difference	Tl, Qa,Fa	Loudness
Tl	Spectral tilt	Qa	Loudness
HRF	Harmonic richness factor	Qa	Loudness
	Aperiodicities		
Jitter	Period-to-period frequency variation		Roughness
Shimmer	Period-to-period amplitude variation		Roughness
PAPR	Periodic-aperiodic ratio		Breathiness, whisper
LoV	Limit of voicing		Breathiness, whisper
NTL	Noise spectral tilt		Breathiness, whisper
IHN	Inter harmonic noise		Breathiness, whisper

Table 2: Voice source parameters and phonation types. "Duality" means that parameters have a dual effect in time and frequency, or that they are equivalent. All the parameters are defined in the text.

4 Voice source parameter analyses

Signal analysis methods for the estimation or evaluation of the voice dimensions described above are reviewed in this section. As non-invasive methods are only of practical use when human subjects are involved, only a few types of signals are available. The most widely used signal is the acoustic signal recorded in the acoustic field at some distance of the head. This signal measures the acoustic pressure (or the acoustic speed). Another signal of interest is the ElectroGlottoGraphic (EGG) signal (Childers et al., 1986, Holmberg et al., 1995). This signal measures the tension between two electrodes positioned on both sides of the larynx. EGG allows for the accurate estimation of F0 and the open quotient as displayed in Figure 3. Some prosodic databases are proposing simultaneous recordings of the acoustic and EGG signals. Other physiological measurements exist, but they are either impractical in fluent speech (e.g. high speech cinematography) or vary slowly, and thus of only little use for prosody analysis (e.g. nasal or oral flow).

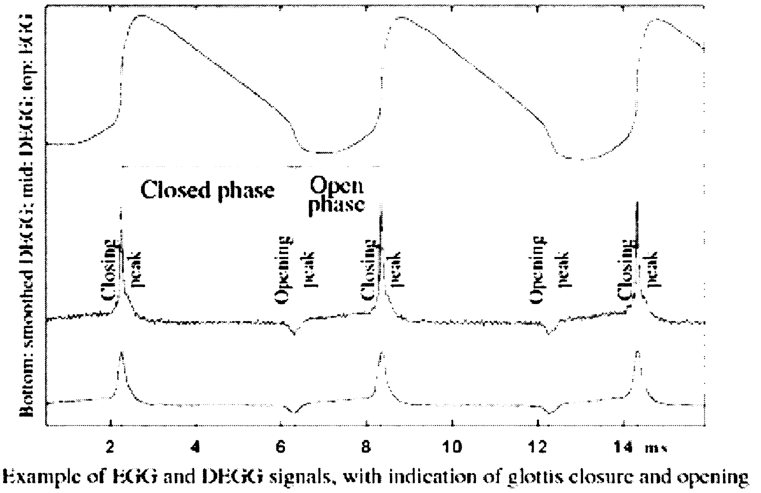

Figure 3: Measurements of voice source parameters using EGG signals (after A. Michaud, http://voiceresearch.free.fr/egg/)

4.1 Analysis of aperiodicities

Noise measurement in speech is widely used for voice quality characterization in clinical voice analysis. In voice pathology there is generally no need to explicitly separate the signal into two components. Long-term measurements of harmonic to noise ratio seem sufficient (see e.g. Kojima et al., 1980; Klingholz, 1987; Hillenbrand, 1987).

On the contrary, short-term measurement methods are needed for prosodic analysis. As explained earlier, the amount of noise in the voice, as it varies over time, is one of the acoustic features of the prosodic domain. Therefore, the dynamics of noise variations need to be traced in the speech signal. Cepstral analysis can be used for short-term harmonic/noise ratio measurement (de Krom, 1993). Methods for explicit decomposition into two components are often based on the speech sinusoidal model (Serra and Smith, 1990). In the method proposed in Yegnanarayana et al. (1998), the cepstrum and sinusoidal approach are combined. The speech signal is decomposed into short-term overlapping frames. Then cepstral analysis is performed in order to select an initial estimate of the noise regions and the voiced regions in the spectrum. A first decomposition is achieved in the short-term spectral domain. Then the decomposition is refined by iterations between the spectral and time domains. This method is specifically designed for extracting the additive noise in speech, and thus provides a meaningful aperiodic component. Moreover, the ability of the algorithm for decomposing the additive noise and voiced excitation has been tested and demonstrated on both natural and synthetic signals containing a mixture of quasi-periodic excitation, additive noise excitation, and structural aperiodicities (d'Alessandro et al., 1998).

The periodic-aperiodic decomposition algorithm is able to separate additive random noise and periodic voicing for a wide range of F0 variations. The dynamic range obtained (i.e. the average difference between the periodic component and the computational noise power spectra) is greater than 30 dB in all cases. The algorithm is able to separate continuous noise as well as pulsed noise. In the case of large jitter or shimmer values, both additive noise and structural noise are merged into the aperiodic component. Whilst it is still possible to achieve separation of a periodic and an aperiodic component, it seems difficult in this case to isolate the various production mechanisms of the aperiodic component. As such, the aperiodic component may be a useful parameter in the analysis of global voice quality, although it cannot be directly interpreted in terms of each underlying speech production parameter. It seems that perceptual discrimination of different kinds of aperiodicities is difficult also for most human subjects. They rather perceive a given amount of noise in speech, but they are unable to distinguish between structural noise and additive noise (Kreiman and Gerratt, 2003). Then both algorithms and subjects may exhibit similar behaviour, and a global measure of aperiodicity as such may be useful.

Analysis of structural aperiodicities can be performed in the time domain. Jitter can be measured by the period-to-period F0 fluctuations. Measurement of shimmer is obtained by period-to-period amplitude fluctuation. Jitter and shimmer variation are indications of large voice quality changes: they are not likely to occur in the prosodic domain because they are in fact long term individual characteristics.

4.2 Analysis of voice pressure

In time domain, voice tenseness or pressure is associated with the voice open quotient. Variations of the open quotient in the time domain result in a variation of the glottal formant position in the spectral domain.

A convenient approach for open quotient measurement is given by the Electro Glottogaphic signals. An algorithm for the measurement of the open quotient in singing using the EGG signal has recently been proposed (Henrich et al., 2004). Figure 3 explains the principle of this algorithm. As the EGG signal provides information about the vocal fold contact area, a sudden variation in the contact leads to noticeable peaks in the derivative. These peaks can accurately be related to the glottal opening (or closing) instants, which are defined as the instants for which the glottal flow starts to increase greatly from the baseline (or decrease greatly towards the baseline). The fundamental period can thus be derived from a DEGG signal by measuring the duration between two consecutive glottal closing instants. The duration between a glottal opening instant and the consecutive glottal closing instant corresponds to the open time. The open quotient can be derived from these two measures as the ratio between the open time and the fundamental period.

If only the acoustic signal is available, inverse filtering and glottal model fitting can be used for open quotient estimation (Fant, 1995, Fujisaki and Ljungqvist, 1986; Alku, 1992).

Another path of thinking is based on the spectral analysis of the open quotient. In many situations, a strong correlation can be found between the open quotient (in the time domain) and the amplitude difference of the first two harmonics (H1-H2). Amplitude differences have to be corrected according to the position of the first formant H1*-H2* (Hanson, 1997). One can show (Henrich et al., 2001) that the relation between the open quotient and H1*-H2* does not hold for all possible variation of the open quotient. The exact relation would also take into account glottal flow asymmetry. However, in many cases this measure can serve for the analysis of the tense-lax dimension.

A new method for the measurement of the open quotient using the position of the glottal formant has recently been proposed (Bozkurt et al., 2005). This method takes advantage of the phase spectrum of the glottal flow.

Depending on the signals available (acoustic and/or EGG), one or the other technique cited above could be of use in tenseness analysis. As almost no comparative study exists, it is difficult to figure out the relative merit of the proposed algorithms. If one is looking for an accurate and robust method of open quotient estimation, simultaneous recordings of the EGG signal are highly desirable. If only acoustic signals are available, I would suggest using spectral methods rather than inverse filtering, as far as robustness is concerned.

4.3 Analysis of vocal effort

The difference in the shape of the glottal flow due to vocal effort mainly affects the higher part of the spectrum. This spectral variation is in turn perceived as a difference in loudness. The accurate time domain measurement of vocal effort is very difficult if not impossible. This is because vocal effort changes in the time domain are associated with very localized time events, like the continuity of the derivative at glottal closure, dissymmetry, and return phase. However, methods based on inverse filtering and glottal flow model matching have in fact been used for vocal effort analysis (Oliveira, 1993). Most of the vocal effort analysis methods take into account the spectral correlate of vocal effort, i.e. the increase of energy within higher frequency bands or the spectral tilt.

Spectral tilt measurement is not as simple as one could think. This is because when the voice source is filtered by the vocal tract, the resulting spectrum is a composite spectrum. Depending on the vowels, a larger or smaller amount of spectral energy will remain in higher frequency bands. In some vocal styles, the vocal tract is adjusted for increasing energy in higher frequency bands (e.g. the singing formant used in classical western singing, that does not depend much on the voice source, but rather on specific vocal tract gestures).

In most spectral tilt measurement methods, some kind of vowel normalization or inverse filtering used to compensate for the effect of the vocal tract is performed. The simplest method is given by measurements of energy in octave bands (or third octave bands) of inverse filtered speech.

A method for spectral tilt measurement is provided by the amplitude difference between the first harmonic amplitude and the third formant amplitude. This takes into account vowel normalization. A correction to compensate for the vocal tract effect must be introduced (Hanson, 1997).

Another measure is provided by the harmonic richness factor HRF (Childers and Lee, 1991), the ratio between energy in the first harmonic and higher harmonics.

It must be pointed out that a reliable and automatic method for spectral tilt measurement is still to be found. Almost everybody agrees on the nature of spectral tilt, but this apparently simple concept is still rather difficult to derive automatically from continuous speech.

4.4 Analysis of voice registers

The analysis of voice register on the basis of acoustic signals alone is difficult. It seems that EGG signals could be more successful, because voice register changes imply vocal setting changes, and can therefore be detected in the glottal activity analysis.

Note that as the voicing period is very long in creaky voice, all the parameters defined relative to the voicing period (Oq, Qa) reach extreme values.

The analysis of voice registers is a challenging problem in speech processing, and particularly in prosodic research. Unfortunately, this problem has received only marginal consideration up to now.

5 Conclusion

The need to consider the quality of voice has emerged in various fields of prosodic research. Voice quality is an integral part of the prosodic description of some languages that have been only recently studied with the help of laboratory equipment. For instance, voice quality tones are needed for a tonal description of Vietnamese. In speech synthesis, F0 and duration modelling reached a limit in terms of quality and naturalness. Prosodic voice quality variation seems needed for expressive speech synthesis. As research in expressive speech processing emerges, new data and new methods are needed to deal with a type of speech where voice quality is much more relevant than laboratory speech read from prepared texts.

Voice quality research remains a challenge as several open problems remain. Acoustic models of the voice source are not fully satisfactory. On the one hand, they can be too simple, and nothing guarantees that they are able to represent all the possible voice source signals (e.g. there is some evidence that they are insufficient for singing voice). On the other hand, they can be too complex, because parameters are linked, and the exact dimensions of the models are not known. Physiological or mechanical models are even more complex and describe only a part of the story. For instance, no mechanical model is able to produce controlled aperiodicities. Analysis of voice quality acoustic parameters can be very difficult in many situations. No fully automatic and reliable methods exist for matters such as open quotient and spectral tilt measurement. The perceptual description of voice quality is also still rather incomplete. How many dimensions should be considered? How do these dimensions correlate with acoustic dimensions?

Despite these difficulties, the importance of voice quality parameters for prosodic research has been recognized for a long time. One can find detailed albeit impressionistic descriptions of voice quality effects on prosody in the pioneering works of Fónagy (1983), Catford (1977), and Léon (1993). Léon mainly addresses the questions of speech "styles", or phono-stylistics. Psychoanalysis and phonetics are associated in the work of Fónagy. He addresses the effects of emotions and affects on articulation and prosody. Catford reviews the wide variety of sounds used in languages (and particularly tone languages). Building on previous research and my own work, I propose the following four prosodic dimensions linked to voice quality variations:

1. the noise dimension
2. the voice register dimension
3. the pressed/lax dimension
4. the effort dimension

These dimensions are relatively independent. Their effects on prosody can be sketched as follows. The noise dimension seems important in signalling the speaking style (and its possible pragmatic meaning), distance to the speaker, intimacy between speaker and listener, emotions, etc.

The pressed/lax dimension is obviously important for the speaking style and emotions. This dimension is also used for tonal distinctions in some languages (e.g. Vietnamese, that makes use of "strangled" tones, i.e. pressed voice quality tones).

The effort dimension is linked to accentuation, together with intonation. Obviously, vocal effort is also important in signalling the expression of affects and emotion.

Register changes are mechanical changes in the vocal activity. Registers have received very little attention in speech research, contrary to singing research, where voice quality is of primary importance. Register changes are easily perceived and also play a role in signalling various emotions, attitudes, and other expressive features in speech.

In this short paper, it has not been possible to describe in detail voice source models or analysis methods. Moreover, this description is by no means complete and exhaustive, but is rather a tentative organization of widespread material. The author hopes that the reader will be able to find enough information via the following reference list in order to enter the world of voice quality in experimental prosodic research.

References

Alku P. (1992): Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11, 109-118.

Alku, P. and E. Vilkman (1996): Amplitude Domain Quotient for Characterization of the Glottal Volume Velocity Waveform Estimated by Inverse Filtering. Speech Communication 18(2), 131-138.

Alku, P., H. Strik, and E. Vilkman (1997): Parabolic spectral parameter. A new method for quantification of the glottal flow. Speech Communication 22, 67-79.

d'Alessandro, C., V. Darsinos, and B. Yegnanarayana (1998): Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. IEEE transactions on Speech and Audio Processing 6, 12-23.

d'Alessandro C. and B. Doval (2003): Voice quality modification for emotional speech synthesis. Proceedings of ISCA-EUROSPEECH'03. Geneva, 1653-1656.

Bozkurt, B., B. Doval, C. d'Alessandro, and T. Dutoit (2005): Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech. IEEE Signal Processing Letters 12(4), 344-347.

Catford, J.C. (1977): Fundamental problems in phonetics. Edinburgh: Edinburgh University Press.

Childers, D.G., D.M. Hicks, G.P. Moore, and Y.A. Alsaka (1986): A model for vocal fold vibratory motion, contact area, and the electroglottogram. Journal of the Acoustical Society of America 80, 1309-1320.

Childers, D.G. and C.K. Lee (1991): Vocal quality factors: Analysis, synthesis, and perception. Journal of the Acoustical Society America 90, 2394-2410.

De Jong, K. (1995): The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. Journal of the Acoustical Society America 97, 491-504.

Doval B. and C. d'Alessandro (1997): Spectral correlates of glottal waveform models: An analytic study. In: Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 97, Munich, 446-452.

Doval, B. and C. d'Alessandro (1999): The spectrum of glottal flow models. Notes et Documents LIMSI 99-07.

Eide, E. M., R. Bakis, W. Hamza, and J.F. Pitrelli (2004): Towards Synthesizing Expressive Speech. In: S. Narayanan and A. Alwan (eds.): Text to Speech Synthesis: New Paradigms and Advances. Prentice Hall, 219-248.

Epstein, M.A. (2003): Voice quality and prosody in English. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 2405-2408.

Fant, G. (1995): The LF-model revisited. Transformation and frequency domain analysis. Speech Transmission Laboratory Quarterly Progress Scientific Report 1995(2-3), Stockholm, 119-155.

Fant, G., J. Liljencrants, and Q. Lin (1985): A four-parameter model of glottal flow. Speech Transmission Laboratory Quarterly Progress Scientific Report 1985(2), Stockholm, 1-13.

Fant, G. and Q. Lin (1988): Frequency domain interpretation and derivation of glottal flow parameters. Speech Transmission Laboratory Quarterly Progress Scientific Report 1988(2-3), 1-21.

Fant, G. and A. Kruckenberg (1995): The voice source in prosody. Proccedings of the 13th International Congress of Phonetic Sciences, Stockholm, vol. 2, 622-625.

Fónagy, I. (1983): La vive voix, Paris: Payot.

Fougeron C. and P. Keating (1997): Articulatory strengthening at edges of prosodic domain. Journal of the Acoustical Society of America 106 (6), 3728-3740.

Fujisaki, H. and M. Ljungqvist (1986): Proposal and evaluation of models for the glottal source waveform. In: Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing. ICASSP'86, 31.2.1-31.2.4.

Gauffin J. and J. Sundberg (1989): Spectral correlates of glottal voice source waveform characteristics. Journal of Speech and Hearing Research 32, 556-565.

Gobl C. (2003): The voice source in speech communication. Production and perception experiments involving inverse filtering and synthesis. Doctoral dissertation. Stockholm: KTH.

Hanson, H.M. (1997): Glottal characteristics of female speakers: Acoustic correlates. Journal of the Acoustical Society America 101, 466-481.

Hanson, H.M. and E.S. Chuang (1999): Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. Journal of the Acoustical Society America 106(2), 1064-1077.

Henrich, N., C. d'Alessandro, and B. Doval. B. (2001): Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data. Proceedings of ISCA-EUROSPEECH'01. Aalborg, 47-50.

Henrich N., C. d'Alessandro, M. Castellengo, and B. Doval (2004): On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. Journal of the Acoustical Society of America 115(3), 1321-1332.

Hillenbrand, J. (1987): A methodological study of perturbation and additive noise in synthetically generated voice signals. Journal of Speech and Hearing Research 30, 448-461.

Holmberg, E.B., R.E. Hillman, J.S. Perkell, P.C. Guiod, and S.L. Goldman (1995): Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. Journal of Speech and Hearing Research 38, 1212-1223.

Klatt D. and L. Klatt (1990): Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America 87, 820-857.

Klingholz, F. (1987): The measurement of the Signal-to-Noise Ratio (SNR) in Continuous Speech. Speech Communication 6, 15-26.

Kojima, H., W.J. Gould, A. Lambiase, and N. Isshiki, (1980): Computer analysis of hoarseness. Acta Oto-laryngologica 89(5-6), 547-554.

Kreiman, J. and B. Gerratt (2003): Jitter, Shimmer and Noise in pathological voice quality. In: C. d'Alessandro, B. Doval, and K. Scherer (eds.): proceedings of ISCA Tutorial and Research Workshop on Voice Quality, VOQUAL'03. Geneva, 57-62.

de Krom, G. (1993): A cepstrum-Based Technique for Determining a Harmonics-to Noise Ratio in Speech Signals. Journal of Speech and Hearing Research, 36, 254-266.

Laroche, J., Y. Stylianou, and E. Moulines (1993): HNS: Speech modification based on a harmonic + noise model. In: Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP'93, Minneapolis, 550-553.

Laver, J. (1980): The phonetic description of voice quality, Cambridge: Cambridge University Press.

Léon, P. (1993): Précis de phonostylistique, parole et expressivité. Paris: Nathan Université.

Ní Chasaide, A. and C. Gobl (2004): Voice quality and F0 in prosody: Towards a holistic account. In: Proceedings of International Conference on Speech Prosody. Nara, 189-196.

Oliveira, L.C. (1993). Estimation of source parameters by frequency analysis. Proceedings of ISCA-EUROSPEECH'93, 99-102.

Rosenberg, A.E. (1971): Effect of Glottal Pulse Shape on the Quality of Natural Vowels. Journal of the Acoustical Society of America 49, 583-590.

Roubeau B, C. Chevrie-Muller, and C. Arabia-Guidet (1987): Electroglottographic study of the changes of voice registers. Folia Phoniatrica 39(6), 280-289.

Roubeau B., C. Chevrie-Muller, and J. Lacau Saint Guily (1997): Electromyographic activity of strap and cricothyroid muscles in pitch change. Acta Oto-laryngologica. 117(3), 459-64.

Scherer, K.R. (1986): Vocal affect expression: A review and a model for future research. Psychological Bulletin 99, 143-165.

Scherer, K.R. (2003): Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227-256.

Serra, X. and J. Smith (1990): Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition. Computer Music Journal, 14(4), 12-24.

Sluijter, A. and V.J. van Heuven (1996): Spectral balance as an acoustic correlate of linguistic stress. Journal of the Acoustical Society of America 100, 2471-2485.

Sluijter, A., V.J. van Heuven, and J.J.A. Pacilly (1997): Spectral balance as a cue in the perception of linguistic stress. Journal of the Acoustical Society of America 101, 503-513.

Yegnanarayana, B., C. d'Alessandro, and V. Darsinos (1998): An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. IEEE Transactions on Speech and Audio Processing 6, 1-11.