The Analysis of Voice Quality in Speech Processing

Eric Keller

Informatique et méthodes mathématiques (IMM), Faculté des Lettres,
Université de Lausanne, 1015 Lausanne, Switzerland
eric.keller@unil.ch

Abstract. Voice quality has been defined as the characteristic auditory colouring of an individual's voice, derived from a variety of laryngeal and supralaryngeal features and running continuously through the individual's speech. The distinctive tone of speech sounds produced by a particular person yields a particular voice. Voice quality is at the centre of several speech processing issues. In speech recognition, voice differences, particularly extreme divergences from the norm, are responsible for known performance degradations. In speech synthesis on the other hand, voice quality is a desirable modelling parameter, with millions of voice types that can be distinguished theoretically. This article reviews the experimental derivation of voice quality markers. Specifically, the use of perceptual judgements, the long-term averaged spectrum (LTAS) and prosodic markers is examined, as well as inverse filtering for the extraction of the glottal source waveform. This review suggests that voice quality is best investigated as a multi-dimensional parameter space involving a combination of factors involving individual prosody, temporally structured speech characteristics, spectral divergence and voice source features, and that it could profitably complement simple linguistic prosodic model processing in speech synthesis.

1 Introduction

The study of voice quality has recently gained considerable importance in speech processing. It bears a direct relationship to the naturalness of speech synthesis systems, and it is part of the natural language constraints considered in speech recognition.

Specifically, current speech synthesis systems are used in an extensive range of applications, encompassing a wide variety of individual speech styles. This brings with it a call for greater authenticity in voice quality. An automatised product description, for example, should be produced in a clearly audible and informative speech style when the context calls for an efficient transmission of factual information, e.g., by telephone. Yet in a promotional context, essentially the same information should

be provided in a more "peppy", more engaging style in order to capture the client's interest. Similarly, contexts can be imagined where low, raspy voices, strong, commanding voices, or quiet, retiring voices could advantageously replace the "standard-fare", neutral and somewhat monotonous voices that are typically provided in current speech synthesis systems.

To do this, two major aspects of an individual speech style must be controlled: first, its *prosodics*, involving the timing, fundamental frequency and amplitude of various elements in the speech signal (typically syllables), and second, its *voice quality*, involving control over a whole series of further aspects of the voice signal, primarily *divergence from spectral distributions*, *voice source features* and *temporally structured features* (e.g., voice on-/offsets, jitter [cycle-to-cycle durational variation]). Although it will be seen that prosodic and voice quality features interact closely, it is useful to distinguish them for historical, explanatory, and algorithmic purposes.

Since there are systematic divergences in signals produced with distinct voice types, the differentiating potential of voice quality can also be exploited for various speech recognition purposes. For example, such divergences are part of an individual biological identification pattern and can be used as one component of a larger speaker identification system for information access control. Also, acoustic properties relating to habitual articulatory configurations and to typical speech motor activity can to a large extent be obtained automatically, and can be exploited as part of a wider feature set in general speech recognition systems.

In line with this logic, this introductory review article explores the methodology of obtaining different types of voice quality from the perspectives of *articulatory determinants* and their *acoustic signal realization*. Certain *articulatory settings* in the vocal tract have been associated with particular acoustic feature sets found with specific speech styles, and thus underpin differentiations formulated at the acoustic level. At the *acoustic signal* level, a variety of signal features have been associated with various aspects of voice quality which are directly relevant to speech processing systems that must deal with *voice quality modification* in the case of synthesis systems, and with *voice quality differentiation* in the case of recognition systems. By briefly reviewing both articulatory and acoustic indicators, we wish to clarify the core of the work to be done in speech synthesis and speech recognition in the area of voice quality. In this paper, we are *not* concerned with detailed algorithmic content. We are mainly interested in understanding the *conceptual framework* of doing research on voice quality[1].

2 Laver's Articulatory and Acoustic Voice Quality Schema

We begin by examining the articulatory and acoustic correlates of voice quality in terms of an initial unified scheme. Our point of departure is the influential classification scheme proposed by Laver [2, 3], which is based on perceptual ratings of articu-

[1] A reader familiar with basic information about the articulatory and acoustic aspects of speech production is supposed here. Much of the basic information not covered here, as well as many additional sources and explanations are furnished in [1].

latorily defined voice quality modifications. This will introduce concepts and a useful initial terminology for voice quality and its acoustic measures.

In Laver's approach, the definitional process issues from a "neutral setting" of the articulatory apparatus. In this setting, the speaker is assumed to produce speech with articulatory organs that show equilibrated muscular tension throughout the vocal tract. At the supralaryngeal level, this means primarily that the jaw is neither lowered nor raised, the tongue root is neither particularly advanced nor retracted, and the lips are not protruded ([2] p. 14). At the laryngeal level, this means that "vibration of the true vocal folds is periodic, efficient and without audible friction" ([2] p. 95). This articulatory configuration produces a "neutral voice" or "modal voice" for a given speaker. Voice quality modification, or "voice modulation", can then be defined as the effect of departures from this habitual setting. This makes it possible to examine the selective modifications of labial, lingual, velar and laryngeal settings in the articulatory tract.

2.1. Summary of Laver's Voice Quality Schema

Laver's schema is summarized in the following listing. As is common, the vocal tract is divided into a supralaryngeal portion on the one hand (lips, tongue, palate, velum, pharynx), and a laryngeal portion on the other (larynx and the associated musculature). Respiration, though relevant to voice quality, is subsumed. The schema (which is based on an extensive literature review prior to 1980) summarizes articulatory settings in terms of an integrated description of active and largely independent motor components of articulatory functioning. At various points indicated in the original text, Laver's account of articulatory descriptions diverges marginally from accounts given by other authors. Settings for different articulatory structures are given in regular font, and the acoustic consequences of non-neutral ("non-modal") settings are given in italics.

1. **Supralaryngeal settings,** defined in terms of *longitudinal*, *latitudinal* and *velopharyngeal* settings of the vocal tract.
    a. **Longitudinal settings:** relative lengthening and shortening of vocal tract.
        i. **Raised larynx and lowered larynx voice ("high" *vs.* "low" voice).** *Raised: high F0, lowered: low F0, often breathy, i.e., presence of noise in all speech frequencies*.
        ii. **Labial protrusion.** *Lowers all formant frequencies, particularly higher formants.*
    b. **Latitudinal settings:** relative widening and constricting of vocal tract at various levels.
        i. **Labial settings:** constrictions and expansions of the lip opening. *Acoustic effect not specified.*
        ii. **Lingual settings:** displacement of tongue body towards the anterior, central or posterior portion of the palate or the back wall of the pharynx. *Displaces formant 1/2 vowel triangle as a*

whole, provides general acoustic coloring in accordance with the vowel towards which the triangle is displaced.

iii. **Faucal settings:** constriction of the passage between the oral and the pharyngeal cavities. *Acoustic effect not specified.*

iv. **Pharyngeal settings:** constriction of the pharyngeal cavity. *Acoustic effect not specified.*

v. **Mandibular settings:** close/open jaw position, protruded jaw position. *Major effect: raising of formant 1 with lowering of jaw, lowering of formant 2 with closing of jaw.*

c. **Velopharyngeal settings:** abnormal degrees of nasality ("nasal twang") or lack of nasality ("denasality"). *With nasality, appearance of one or more nasal formants (200-300 Hz, 1 kHz and 2 kHz) and of one or more anti-formants (anti-resonances[2]), plus overall loss of acoustic power, especially in first formant and in higher frequencies, accompanied by a flattening and widening of formants.*

2. **Laryngeal settings,** defined in terms of *adductive tension* approaching the arytenoid cartilages, *medial compression* on the vocal processes of the arytenoid cartilages, and *longitudinal tension* along the vocal folds ([2] pp. 108-109).

a. **Modal voice:** moderate adductive tension, medial compression and longitudinal tension. *Acoustic norms for neutral or modal voice are summarized in [2], Chpt. 1. They correspond to the acoustic indicators obtainable from a standard corpus of neutral, declarative speech.*

b. **Falsetto:** alternative to modal voice, high adductive tension, large medial compression and high longitudinal tension. *F0 two to three times as high as for modal voice, simplified waveform, rapid spectral falloff in high frequencies.*

c. Whisper: can be combined with modal or falsetto voice; low adductive tension, moderate to high medial compression, variable longitudinal tension, producing a triangular opening of vocal folds of variable size. *Addition of considerable noise in all frequencies, particularly in the higher ranges.*

d. Creak: can be combined with modal or falsetto voice; also called "vocal/glottal fry"; emission of short vocal pulses at a frequency and with a degree of inter-pulse damping that permits their perception as "separate taps". *Low f0 (20-90 Hz), separate, potentially complex vocal pulses.*

e. Harshness: can be combined with other voice-forms; other terms: "raspy", "rough"; due to great adductive tension and great medial compression, yet with indifferent longitudinal tension, inducing excessive approximation of vocal folds. *Aperiodicity in f0, "jitter".*

f. **Breathiness:** can be combined with other voice-forms; low adductive tension, low longitudinal tension, in comparison with whisper, low medial compression. *Addition of moderate degrees of noise, reduction of higher frequencies.*

3. **Tension Settings.** Defined by a general tensing or laxing of the entire vocal tract musculature, giving "tense", "sharp", "shrill", "metallic" or "strident" voices on

[2] Reduction of spectral amplitude at a certain resonance frequency.

the one hand, and "lax", "soft", "dull", "guttural" or "mellow" voices on the other. *Distinguished primarily by relative amounts of energy in the upper and lower harmonics, where the limit between the two is set at about 1 kHz. Secondarily, tense voices tend to show higher overall amplitude than lax voices.*

2.2. Comments on Laver's Voice Quality Schema

A number of comments are in order with respect to Laver's approach to the analysis of voice quality. As indicated, the point of departure is the notion of *vocal tract setting*. Laver ([2] p. 13) says that "a preliminary way of envisaging an articulatory setting is to imagine a cineradiographic film being taken of the vocal apparatus in action over, say, 30 seconds. If the individual frames of the film were superimposed on top of each other, a composite picture might emerge which would represent the long-term average configuration of the vocal organs. This configuration constitutes the setting underlying the more momentary segmental articulations..." In Laver's approach, a setting is thus an average state of the vocal tract, and a given voice quality can be thought as an acoustic condition resulting from such an average articulatory configuration. As a reflection of this, one traditional acoustic measure of voice quality settings has been the long-term (average) spectrum (LTS or LTAS), a spectrum derived from a set of spectra taken over a given time period (typically 30+ seconds of speech).

The articulatory definition of a setting has a number of important implications. The first is that certain departures from a neutral setting in one part of the vocal tract can be combined with departures in another part of the speech apparatus, while other combinations are impossible because of articulatory linkage. This reduction in degrees of freedom has the effect of reducing the total number of voice quality states that can be either produced or perceived. For example, the laryngeal production of falsetto voice appears to be quite different from that used for modal voice, and a combination of falsetto and modal voice is thus impossible (although a rapid alternation between the two is). As a result, the combination of modal and falsetto is impossible, while some other combinations (such as modal and creak) are possible.

Also, some voice quality states may be articulatorily and acoustically similar to each other, such as whispered and breathy speech, while others are strongly distinctive in both respects. However since the intention behind whispered speech (typically the wish not to be heard by others) is generally different from that which leads to breathy speech (typically a secondary effect of relaxed or affective speech), it remains important to distinguish the two types of voice quality, even though in terms of an articulatory description, the two types of voice form a continuum ([2] p. 133).

A limitation of the Laver scheme is rooted in the observability of articulatory events. For example, some pathological and some less common voice styles are characterized by the prominent presence of mucus and saliva ("wet voices"). This produces audible acoustic modifications in the voice. However, such voices are not distinguished in Laver's scheme, since the presence of oral humidity was rarely measured in pre-1980 studies and even today, the degree of oral humidity is not normally assessed in voice studies.

Further, a given articulatory setting may in fact correspond to a linguistically distinctive state in a given language. For example, nasality is distinctive in French or

Portuguese, and in these languages, acoustic indicators of nasality are primarily associated with the distinctive feature set of the language and not with voice quality. In English, on the other hand, nasality is common among certain speakers of U.S. English, particularly in vowels preceding nasal consonants (a "nasal twang"). In these cases, it is appropriate to speak of nasality as a type of voice quality.

Finally, there are a number of problems associated with the identification of an articulatory setting. As we have seen, this notion supposes a temporary or habitual modification of the vocal tract in a given individual. A temporary modification can be empirically verified in a given speaker. But it is difficult to do so with a habitual modification, since the vocal tract is rarely or never in a neutral setting. In such cases, the underlying norm is derived from an appropriate sample of similar speakers, which in turn introduces the difficulty of establishing what constitutes a "similar speaker". Given the small anatomical and physiological modifications that are responsible for what are often rather elusive acoustic differences, empirical verification of the notions presented here is therefore not always easy.

A related empirical difficulty resides in the fact that Laver's schema is defined in terms of a given *individual's* articulatory settings (neutral and otherwise), while voice quality as generally understood concerns the use of voice by the *generality* of speakers. The definition, distinction and classification of non-neutral voice quality, as well as its application to speech processing, usually concerns *groups* of speakers[3], yet in Laver's approach, it must be based on (sometimes only supposed) articulatory settings in individual speakers. These group associations are very difficult to verify empirically in an articulatory framework. Without suggesting that Laver's approach is ill-founded or that its schema is irrelevant to current speech processing research, it would clearly be helpful to have a set of *acoustic* features that reliably link all speakers showing a given voice quality ***x*** or ***y***. It is with this goal in mind that we turn to a closer look at the acoustic measures used in the analysis of voice quality, to see if such features can indeed be identified in the acoustic waveform.

3 Acoustic Voice Quality Measures

Acoustic measures of voice quality must satisfy a series of requirements:

- Perceived differentiations of voice should be reflected in predictable variations in the signal waveform or in one or several of its derivatives.
- The measures should reflect states or a set of states of the vocal tract typical of a certain individual, and should be separable from states that are shared by large numbers of speakers and that are relevant to the production of phonetic segments or of prosodic features in a community of speakers ("linguistic features").
- Since perceived voice quality reflects supra-laryngeal as well as laryngeal and sublaryngeal (respiratory) vocal tract settings, measures of the acoustic

[3] E.g., sportscasters, priests or ministers in a church service, mothers interacting with young children, etc.

speech waveform should capture all of these types of information, and should separate them if possible.

It is not easy to satisfy all of these requirements with a single measure. As will be seen, measures that satisfy one requirement tend to fail on another, and the assessment of voice quality probably ultimately requires the parallel application of a whole series of measures. Let us review the most important measures, beginning with the long-term average spectrum.

3.1. Spectral divergence and the long-term average spectrum (LTAS)

As briefly mentioned above, one traditional measure of voice quality has been the long-term average spectrum. In this approach, power spectra are taken at a given frequency (e.g. one every *ms*) throughout a given stretch of speech, are averaged and are optionally summarized as a set of spectral bands. Average differences between spectral profiles on the same stretch of speech presumably reflect long-term settings and are ideally expected to capture the essence of voice quality differences. The LTAS typically stabilizes after about 40 seconds [4], cited in [5]. Also, the LTAS reliably identifies quasi-constant characteristics such as a singer's formant [6], cited in [5].

However, this approach is subject to four major limitations. First, long term averages mix spectral features relevant to segmental information with those more directly related to voice. This makes it difficult to compare *different* stretches of speech, or even, stretches that are lexically the same, but pronounced somewhat differently. As a consequence, there has been a tendency to replace the simple LTAS by more localized measures recently. Klasmeyer [7] for example suggests performing LTASs on vowel nuclei only, and Keller [8] replaced LTASs by averages based on a large number of spectra obtained from the centre of vowel nuclei, as identified in a large corpus.

The second problem is that averaging neglects temporal dynamics that often contribute to the definition of a given voice quality. Creak, for example, is defined in Laver's scheme by the fact that voice pulses are clearly spaced in the temporal domain. Since fundamental frequency values between 70 and 90 Hz are common, and since spaced f0 pulses with frequencies up to 90 Hz can be perceived as creak ([2], p. 124), it is this temporal spacing, not the absolute fundamental frequency value, that is responsible for the perceptual impression of creaky voice. An LTAS neglects this type of information, just as it neglects some other prominent individual temporal features such as jitter, i.e., cycle-to-cycle variation, or particular temporal evolutions at vowel-consonant transitions.

Third, LTAS profiles obtained from high-amplitude signals show considerable divergences from LTAS profiles recorded from low-amplitude signals [5]. Increases in vocal loudness cause a larger increase of LTAS at 3 kHz than at 0.5 kHz, which complicates even comparisons of multiple recordings of the same text from a single speaker. Below 4 kHz, the difference between high- and low-amplitude LTAS profiles is predictable from overall sound level to a reasonable degree (within 2-3 dB), but above 4 kHz, the individual variation is too great to be modelled [5]. It is thus recommended to take at least three recordings at different loudness levels to calculate LTAS profiles.

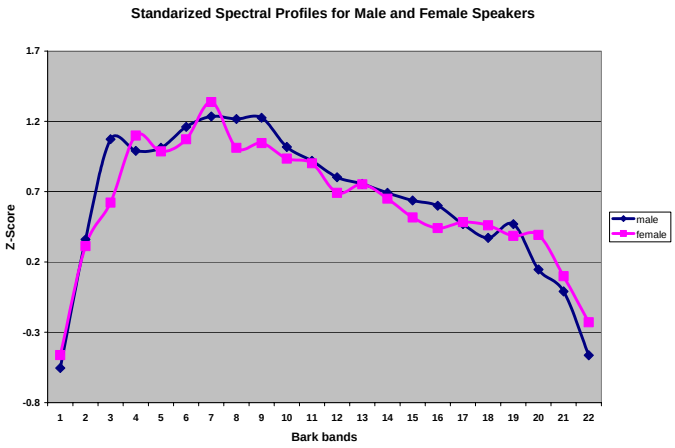

Fig. 1. Averaged and standardized *gender* profiles for some 30'000 vowel nuclei identified in the MARSEC corpus. Although all differences were significant (except for bark band 1), the only major differences occur around Bark bands 3 and 8-9.

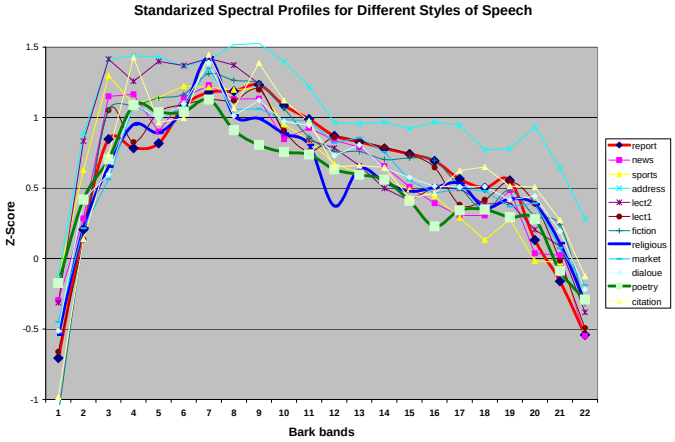

Fig. 2. Averaged and standardized *speech style* profiles for some 30'000 vowel nuclei identified in the MARSEC corpus. Although all differences were significant, the differences tend to be minor and are barely audible when resynthesized by a spectral synthesis method such as HNM (harmonics and noise modeling).

Finally, distinctions obtained through averaging methods have been regrettably weak. In the Keller study just referred to ([8]) which involved the examination of

some 30'000 vowel nuclei from 56 speakers of the MARSEC corpus[4], only relatively minor systematic spectral differences were identified in standardized spectra for gender and speech style (Figs. 1, 2). Attempts to use these differences to modify speech resynthesized with a harmonics-and-noise model were not successful, presumably because the averaging technique removed acoustic and temporal indices that are important for the differentiation of gender and speech style.

Altogether, while LTASs undeniably illustrate certain voice quality differences, they are apparently of insufficient power to effect voice quality modification or speaker distinctiveness in speech synthesis when used alone. It must be assumed that a number of dynamically patterned features in the acoustic waveform make further important contributions to the perception of given voice quality. We next turn to an important subgroup of such features which are the prosodic markers.

3.2. Acoustic Measures Related to Prosody

From the previous observation, it appears unlikely that voice quality can be defined solely in terms of static acoustic features such as the LTAS. As indicated in Laver's schema, several temporally structured elements enter into consideration as well, together with static acoustic indicators, as well as several parameters that are traditionally associated with *prosody*, notably fundamental frequency (f0) for perceived pitch, intensity (dB) for perceived loudness, as well as duration (typically of vowels or syllables). Raised and lowered larynx voices, for example, were distinguished primarily on the basis of pitch in Laver's schema. Also, some voices are distinctively loud.

At the same time, it is well-known that pitch and loudness have several linguistic functions, such as the declarative/interrogative distinction and accentuation found in European languages. A local combination of high f0 and high dB value can thus be indicative either of a high voice, or of a question intonation, or both. The assignment of prosodic and/or voice quality values depends in part on how voice quality is defined, and in part on the context. In Laver's "stable articulatory setting" approach, voice quality is the underlying, relatively unchanging base for a parameter like f0, and linguistic significance is decided by different degrees of departure from this line. However, this decisional process becomes more complex as the notion of voice quality is extended to cover a wider variety of concepts.

Suppose for example a combination of high f0, dB and duration values on the first syllable of the word *really*. Given sufficient contextual information, a human listener can deduct the relevant linguistic and social information from various acoustic components of this syllable. Suppose for example (1) that the word is contextually identifiable as part of a question, (2) that it is spoken with a timbre assignable to a male voice (whatever "timbre" might be in this context), (3) that its f0, dB, and duration values are excessively high for the given speaker, language and context, and (4) that phrasal on- and offsets show a high degree of jitter. Taken together, these indicators not only suggest to another speaker of English that the word is part of a question, but also that the speaker is male and possibly anxious. This decisional process involves minimally a fairly complex and language-specific set of thresholds for f0, duration,

[4] Machine-Readable Spoken English Corpus, available from www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec

dB and jitter, plus knowledge about spectral profiles corresponding to male vs. female vocal tracts, the pragmatic and linguistic contexts that permit the interpretation of the word as part of a question, and possibly knowledge about previous speech performances by the same speaker.

In speech recognition and speech synthesis, analogue decompositional, decisional and combinatorial efforts must be undertaken. In speech recognition, success of linguistic, paralinguistic and extralinguistic[5] interpretations of a given combination of f0, dB or jitter values depends on the adequacy of the multi-tiered statistical models to which incoming utterances are compared. When integrating such information in speech synthesis, all contributing elements have to be furnished with a credible temporal profile, and must be combined satisfactorily to evoke the intended utterance. This supposes not only a well-defined set of correspondences between acoustic and linguistic, paralinguistic and extralinguistic elements, but it also involves knowledge concerning the value and linearity of the various contributional ratios. For example, are linguistic, paralinguistic and extralinguistic models strictly additive? If not, to which proportion does each level contribute to the overall measure of f0, dB or duration? And are these proportional contributions linear throughout the usage range?

Furthermore to synthesize natural-sounding speech, prosodic parameterizations must be combined with appropriate voice quality manipulations. In a study where fundamental frequency and voice quality parameters were manipulated separately in the synthesis of various types of voices[6], voice quality manipulations contributed in important fashion to the communication of affect [9]. Only the signals that included adjustments for voice quality succeeded in communicating the emotions in question reliably. By contrast, signals combining fundamental frequency manipulations with voice quality appropriate to a neutral voice (i.e., signals similar to those used in current speech synthesis systems) were judged much less expressive by comparison.

In summary, the prosodic parameters of fundamental frequency, duration and intensity are clearly of relevance in judgments of voice quality. However, in contrast to other areas of research on voice quality, particularly voice source analysis, the relationship between linguistic and individual components of prosody has been examined far less systematically. It is evident that future paralinguistic interpretation of these indicators will have to involve a parallel clarification of the speaker's linguistic and pragmatic situation. Various extralinguistic markers in the prosodic domain are also indicative of the speaker's individuality, age, sex and various psychological attributes.

Further, no study has to our knowledge systematically explored the relationship between prosodic parameters and personality style. Yet it is a common observation that speakers, apart from their voices, differ tremendously regarding their prosodic style. With respect to intonation and rhythm alone, for example, one may describe a

---

[5] It is useful to distinguish linguistic, paralinguistic and extralinguistic aspects of voice quality. The *linguistic* component communicates semantic and distinctive information that is part of the speaker's *language*. The *paralinguistic* component communicates the speaker's affective, attitudinal or emotional states, his/her sociolect and regional dialect, as well as aspects of turn-taking in conversation. This components is to a large degree specific to a given language or language group. The *extralinguistic* component communicates the speaker's individuality, gender and age, i.e., the characteristics of a certain speaker. It can be judged independently of the speaker's language.

[6] Modal (neutral), breathy, whispery, creaky, lax-creaky, modal, tense and harsh voice.

speaker's expression as fluent, lively, constrained, relaxed, etc. Intonation, rhythm and breathing are apparently the main parameters that convey these styles, in some specific combination that still needs to be established. These are the parameters that are often imitated by impersonators, whereby the imitation of the vocal component is more difficult to perform [10, 11].

Also, studies on twins are interesting in this context because of their morphological similarity (see e.g., [12]). Loakes [13] found acoustic differences in the speech of twins, despite closeness as judged perceptually. Speech samples from twins were compared by focusing on variables which have a high degree of speaker variation (e.g. consonant sequences /stR/, /tR/ and /tS/), and also variables that show minimal variation (e.g. mid-vowels such as /E/). Within- and between-speaker differences were identified using both auditory and acoustic methods of analysis. Results indicate that *similar-sounding voices* that originated from vocal tracts with minimal differences can be discriminated acoustically in the Formant 4 region. Prosodic parameters were not explored, and one may expect that some fine prosodic differences could also be identified.

Zellner Keller is currently investigating relationships between certain prosodic styles (e.g., very expressive *vs.* barely expressive), signalled by specific combinations of acoustic parameters, and specific personality styles. The aim of the study is to examine if and to which degree some personality styles can be reliably associated with typical strategies of prosodic expression, considering that this level of expression will be more or less obfuscated, due to the other superimposed social, linguistic, emotional and attitudinal encodings. In first results, it was found that listeners can and do associate speech with personality traits [14]. These attributions are remarkably consistent across listeners, even when they have a different language background (French/German). The significant correlation between personality and prosody clusters can be explained by supposing that listeners attribute personality traits on the basis of prosodic features of speech.

For the assessment and the automatic processing of individual voice quality, it is thus important to examine the contribution of the three classical prosodic parameters in interaction with indicators of their linguistic and paralinguistic significance. These parameters must also be combined with appropriate voice quality parameters to effect synthesis that is natural-sounding with respect to affect. This is the issue we turn to next.

3.3. Source Modeling

While all parts of the articulatory tract contribute to some degree to voice quality, the research community largely agrees that conditions affecting the air flow at the glottis (i.e., laryngeal or *source* settings) are responsible for particularly salient aspects of this speech component[7]. Conditions relevant to voice quality can be transitory (as in the case of voice on- and offsets), short-term (e.g., over the duration of a vowel) and

[7] Indeed, some researchers reserve the term "voice quality" uniquely for aspects of the sound produced by the larynx, i.e., the glottal waveform. We did not follow this tradition here, because for speech synthesis and speech recognition processing, *all* vocal tract effects on voice quality must be considered and modelled.

longer-term (i.e., affecting all voiced components on an individual's speech). Much research of the past 20 years has thus been directed at obtaining the glottal waveform reliably and automatically, with a minimum of discomfort to the speaker.

Unfortunately, the source waveform is difficult to recuperate, since even intraoral recordings performed directly above the larynx show effects of resonator coupling from the supraglottal cavities. Approximations to the "pure" glottal waveform can generally only be obtained through recordings performed with specially designed pneumotachograph mask for recording oral airflow at the mouth, the "Rothenberg mask" [15], from more indirect evidence such as glottal pulse wave trains obtained with electroglottographs[8,9] [16], or from calculations and theoretical inductions performed on the acoustic speech waveform, that is, from so-called *inverse filtering* methods. This latter approach is of particular interest to persons working in speech processing, since it can be applied to standard sound recordings performed with a microphone outside the mouth.

In inverse filtering, the glottal waveform is extracted from the speech waveform by separating the respiratory and glottal *source component* of the speech waveform from the *filter component* (corresponding to the supraglottal vocal tract resonator contribution), plus the radiation loading at the lips. This source-filter model of speech production (originally formulated by Fant in [17]) treats the glottal source and the supraglottal filter as independent components. Although more recent research has documented various interactions between the glottal source and the vocal tract resonances [18], Fant's original theory of speech production is still a good point of departure, particularly with respect to voice quality transmitted in signal portions where the airflow is much more strongly impeded at the glottis than in the supraglottal vocal tract, a condition that characterizes most vowels. In this part of the chapter, we follow the excellent general introductions to this topic by Ní Chisaide and Gobl [20] and by Gobl [21][10].

3.3.1. Manual vs. Automatic Inverse Filtering Methods

Consider the top part of Figure 3 (adapted from [20]) showing the effects of vocal tract filtering on an idealized source spectrum. The source spectrum (representing typical mid-vowel glottal flow with a normal voice) simply reflects the harmonic components of the glottal wave with a constant slope of 12 dB fall-off for every doubling of the frequency[11]. It can be seen in the speech output spectrum that the vocal tract cavities and their resonances (formants F1-F5) in effect impose a filter on the

---

[8] Electroglottography is a non-invasive method of measuring vocal fold contact during the production of voiced sounds. An electroglottograph (EGG) measures the variation in impedance to a small electrical current between a pair of electrodes placed on the two sides of the neck, as the area of vocal fold contact changes during voicing.

[9] An excellent survey of empirical methods of obtaining and measuring the glottal waveform is found on the following University of Stuttgart webpages: http://www.ims.uni-stuttgart.de/phonetik/EGG/page1.htm

[10] See also Emir Turajlic's webpages on glottal pulse modelling at: www.brunel.ac.uk/depts/ee/Research_Programme/COM/Home_Emir_Turajlic/index.htm

[11] The true source spectrum shows various dips and does not have a constant slope ([20], p. 427).

source spectrum[12]. Inverse filtering (see middle portion of Figure 3) consists of designing a filter of *antiresonances* (opposite-valued resonances) in such a manner that the vocal-tract filtering effect is cancelled. The effect of a well-designed filter on the speech waveform is seen in the bottom portion of Figure 3. The oral airflow U(t) (essentially, the speech waveform recorded by the microphone) is transformed by inverse filtering into glottal airflow $U_g(t)$, which is generally shown in its differentiated form as the *differentiated glottal airflow* $U'_g(t)$. Various aspects of this differentiated glottal airflow have been related to voice quality (see below).

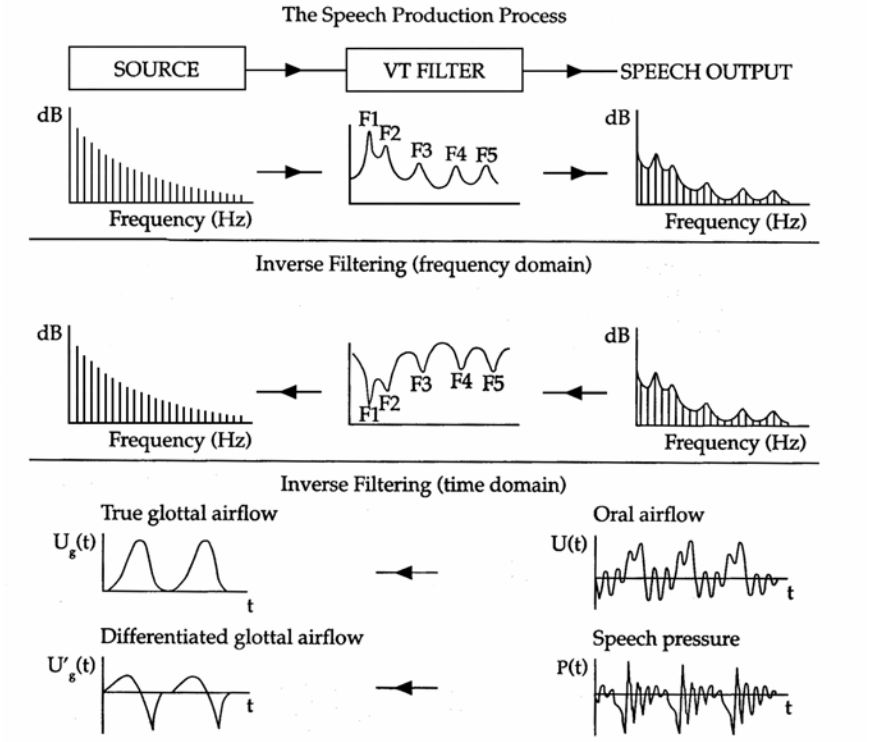

Fig. 3. Source-filter decomposition in frequency and time domains. *Top:* A smoothly decaying theoretical voice spectrum is modified by an idealized vocal tract filter to produce the measurable speech output for a vowel. *Middle:* The speech output spectrum is inverse-filtered to extract the underlying voice spectrum. *Bottom-top:* The effect of converting from oral airflow to glottal airflow. *Bottom-bottom:* The effect of the conversion from speech pressure (the differentiated oral airflow) to a differentiated glottal airflow, which is the habitual manner of representing the voice waveform in the time domain. The differentiated representation facilitates the standardized identification of voice quality markers. Figure from [20] reproduced with permission.

[12] For a recent calculation of vocal tract resonances (formants) for 18 Swedish vowel states measured with X-ray, see [19].

It can readily be seen that the success of the inverse filtering procedure is directly related to the nature of the input waveform. When the waveform provides insufficient digitalizable amplitude, an inverse filter is impossible. Inverse filtering methods are thus preferentially applied to voiced portions of speech with sufficient amplitude to permit a reliable identification of formants. Further, substantial problems arise when inverse filtering is performed by automatic procedures. Formants can change rapidly as a result of changes in resonator cavity size, particularly in the neighborhood of consonants, which tend to "throw off" relatively simple formant tracking mechanisms. Also, inverse filter modeling frequently suffers interference from source-filter interactions at weak glottal flows [23], or in places where a non-modal voice is used. Finally, there is the possibility of phase distortion with tape-recorded material. Ideally, inverse filtering is thus performed manually and interactively, on selected vowel nuclei with normal or high intensity, best of all with material recorded directly to computer. In 1997, Ní Chisaide and Gobl concluded that the most accurate source signal is obtained by interactively (manually) fine-tuning the formant frequencies and bandwidths of the inverse filter [20].

Some more recent research suggests that at least for signal portions with sufficient amplitudes and speech falling within a reasonable range of predictability, particularly in male modal voices, automatic inverse filtering methods can show reliable performance, thus facilitating the acquisition of the considerable amount of data required in voice quality research. To understand the challenges involved in performing this type of operation for a wide variety of voices, we must examine the overall process of performing voice source analysis.

3.3.2. Voice Source Extraction: Automatic Formant Tracking

As indicated above, the extraction of the voice source involves two main steps, *inverse filtering* and *source modeling*. At the inverse filtering level, the essential difficulty consists in tracking the formant frequencies in the speech waveform, establishing their bandwidth, and assuring that classical voice parameters can be identified in the reconstituted source waveform. At the source modeling level, the challenge is to design a numeric model that captures the waveform modifications (i.e., the "classical voice parameters") that are associated with perceptible variations in voice quality.

The tracking of formants has traditionally been handled by an LPC that specifies the frequencies (and indirectly, the bandwidths) of the antiresonators required to cancel the formants. The average spacing between the poles is determined by the vocal tract length: for a typical male with a vocal tract of 17.5 cm, there is on average one formant per kHz. It is crucial to obtain the right number of poles and bandwidths, particularly in the lower frequency domain (Formant 1), while minor errors in the higher formants have less effect on the source pulse shape or the source frequency spectrum ([22], cited in [20]). LPC-type all-pole functions are adequate for many sounds such as vowels, yet for certain sounds such as nasals and laterals, the spectrum contains zeros as well as poles. While these zeros should theoretically be cancelled by the inclusion of corresponding poles in the inverse filter, they tend to be difficult to calculate and most researchers use all-pole models for all sounds.

A number of methods have been exploited for improving the reliability of formant tracking, revolving primarily around the concepts of the exploitation of contextual

phonetic information, pitch-synchronous analysis, filtering and optimized voice source matching techniques.

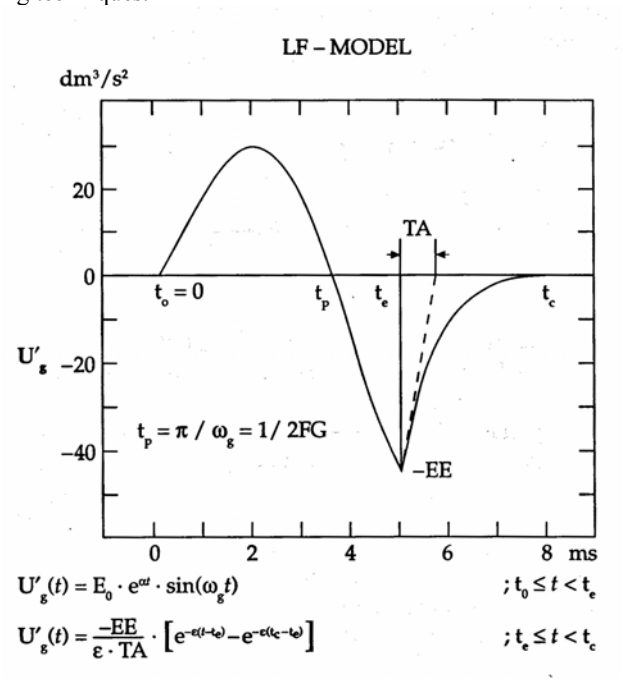

Fig. 4. A glottal pulse approximated by the LF model. Features prominently associated with voice quality are either marked in the figure or can be identified from features in the waveform: EE excitation energy (-EE), RA "return time" (TA), RG "glottal frequency" ((1/2Tp)/f0, or the inverse of twice the opening phase T_p normalized to fundamental frequency), RK "glottal asymmetry" ($t_p/(t_e-t_p)$, or the relationship between the opening and closing branches of the glottal pulses) and OQ "open quotient" (t_e/t_c, or the proportion of the glottal cycle during which the glottis is open). Figure reproduced with permission from [20].

McKenna [24] recalls that the quality of conventional fixed-frame pitch-asynchronous LPC (typically using the autocorrelation method) depends on the assumption that both formants and underlying articulatory movements are smooth and slow-evolving. However, the acoustic coupling of glottal and subglottal space during the open phase of the glottal period introduces momentary drops in formant frequencies that are reflected in pitch-asynchronous analysis as a general lowering of formant frequencies and an increase of formant bandwidths. By contrast, *pitch-synchronous analysis* improves the "sharpness" of formant tracking and thus contributes substantially to the quality of the inverse filtering process. Another benefit of pitch-synchronous analysis is that source data can be gathered specifically from the closed-glottis portions of the signal, which minimizes the effects of acoustic coupling between subglottal and supraglottal space.

It is noted however that this approach implies a considerable reduction of data points available for analysis in any specific voice period. This can render it inappro-

priate for female and children's voices which tend to provide a reduced number of data points, a disadvantage that can be overcome through Kalman filtering [24]. This recursive technique uses estimates based on measures of previous pitch periods to improve the modeling of relevant portions of the source waveform.

3.3.3. Voice Source Extraction: Glottal Pulse Modeling

Once the source waveform has been reconstructed through inverse filtering, it can be analyzed for voice quality features. The *differentiated* glottal waveform rather than the true glottal waveform is generally used for this purpose because of the greater ease of identifying relevant parts of the glottal cycle in the differentiated glottal pulse (see Figure 4). It also conveniently turns out that the radiation at the lips is approximately 6 dB per octave in the speech midrange frequencies, which corresponds to a filter that can be approximated by a first order differentiation of the output signal. The differentiation step thus serves at the same time to cancel lip radiation. The net effect of this differentiation and the canceling of lip radiation is a boosting of the higher frequencies, which improves the modeling in this frequency range. As in the case of inverse filtering, the robustness of source modeling can be improved through optimization procedures. Fu and Murphy [25] for example describe a multi-parameter nonlinear optimization procedure performed in two passes, where the first pass initializes the glottal source and the vocal tract models, in order to provide robust initial parameters to the subsequent joint optimization procedure, which iteratively improves the accuracy of the model estimation.

Various portions of the glottal waveform obtained in this manner have been shown to be of direct relevance to voice quality. To identify these features reliably throughout one or several databases, the source waveform is generally matched by an approximate mathematical description of the waveform. Liljencrants and Fant [26] have provided the best-known such formulation, known as the LF-model[13]. Figure 4 shows a differentiated source cycle described by the LF model. The cycle is approximated by two different equations for its two branches that extend from $\mathbf{t_0}$ to $\mathbf{t_e}$ (the maximal excitation point) and from $\mathbf{t_e}$ to $\mathbf{t_c}$, where $\mathbf{t_c}$ corresponds to point $\mathbf{t_0}$ in the next cycle.

Features prominently associated with voice quality are either marked in figure 4 or can be identified from features in the waveform. They are as follows (terminology as in [21]:

EE excitation energy [shown as -EE]. This is the main source parameter, showing the overall strength of source excitation.

RA "return time" or "dynamic leakage" [shown as TA]. This parameter measures the sharpness of glottal closure, i.e., the time that the glottal folds take to accomplish closure. This in turn has a major effect on the slope of the glottal spectrum. Sharp closures are associated with increases of spectral amplitudes in the high frequencies ([20], p. 440).

---

[13] See [20] p. 437 for a comparison and [21], p. 7, for references of other mathematical approximations to the glottal cycle waveform. Alternative models share many common features with the LF model, but they can generally be described by three to five parameters, plus fundamental frequency.

RG "glottal frequency" [$(1/2T_p)$/f0, or the inverse of twice the opening phase T_p normalized to fundamental frequency]. This parameter estimates the degree of boosting found with some voices in the areas of the first and the second harmonic.

RK "glottal asymmetry" [$(t_e-t_p)/(t_p-t_0)$, or the relationship between the opening and closing branches of the glottal pulses]. In general, glottal pulses tend to be right-skewed, and increased symmetry results in a boosting of lower frequencies and a deepening of spectral dips.

OQ "open quotient" [t_e/t_c, or the proportion of the glottal cycle during which the glottis is open]. Increased degrees of OQ result in a boosting of the lowest harmonics of the voice spectrum.

AS "aspiration". This acoustically important non-periodic parameter can unfortunately not be derived from the LF model, and its algorithmic separation from periodic source information is a non-trivial challenge. However for synthesis purposes, appropriately-filtered pseudo-random noise can be generated to compensate for the absence of empirically-derived estimates of aspiration noise at the glottis.

Since some of these parameters have been found to co-vary frequently, some further simplification of this parameter list may be possible, at least in the case of some voices (see discussion in [20] pp. 441).

3.3.4. Source Parameters and Specific Voice Types

Voice quality modifications associated with variations in the described parameters have been widely examined. In this presentation, we limit ourselves to a short review of commonly occurring voice types. For this, we summarize Ní Chisaide and Gobl's observations in [20] of four key voice types defined in Laver's descriptive scheme (see above):

- *Modal voice.* In tune with the occurrence of normal, efficient and frequently complete closures at the glottis, glottal cycles take on the standard source waveform for modal voice. The spectral slope is particularly steep. Gradual or abrupt changes to other voice modes (e.g., breathy or creaky voice) are frequent.
- *Breathy voice.* It will be recalled that in Laver's scheme, glottal articulation for breathy voice was characterized as a general lack of tension. Vocal fold vibrations are inefficient and never show complete closure. Acoustically, this translates into audible aspiration and slow RA (time of "glottal return") values. High RK values demonstrate relatively greater symmetry in the glottal pulse, and high open quotient values (OQ) reflect the looseness and gradualness of the glottal gesture.
- *Whispery voice.* Like breathy voice, this type of voice is characterized by low tension in the glottis, with the exception that there is moderate to high medial compression and moderate longitudinal tension. This tension pattern creates a triangular glottal opening whose size varies inversely with the degree of medial compression. Acoustically, this inefficient mode of voice production translates into high aspiration levels and generally more extreme deviations from modal values than those seen with breathy voice. Whispery voice differs mainly from breathy voice by its lower RK values (greater pulse asymmetry due to a shorter closing branch) and a lower "open quotient" OQ, i.e., a lower proportion of cycle time spent in an open state.

- *Creaky voice.* According to Laver, creak results from high adductive tension and medial compression, but little longitudinal tension. It is generally associated with very low pitch. However, f0 and the amplitude of consecutive glottal pulses are very irregular and frequently alternate with normal, non-creaky voice. Low OQ, low RK and a relatively high RG have been observed for creaky voice [20].

4. Conclusion

In many respects, the study of voice quality represents the "ultimate frontier" for speech processing research, due to considerable complexity at both functional and signal processing levels.

At the functional level, each speaker has his or her own characteristic *individual* voice quality. In addition, the *psychological* dimension of voice quality reaches from the distinction of personality types, via the communication of affect and emotion, to the communication of delicate nuances in conversational exchanges. *Sociologically*, certain types of voice quality serve as social markers (such as markers for position in a social hierarchy, or for homosexuality). Several aspects of voice quality have also been integrated into the *linguistic* coding system of certain languages[14]. These various functional strands interact and are partially superimposed. For a successful use of parameters related to voice quality, speech processing technologies must be rendered sensitive to and/or implement structuring related to these various predictor groupings.

At the signal processing level, the multidimensional complexity of voice quality takes another form. In this review, it was seen that systematic variations relating to voice quality can be documented as *spectral divergence* in long-term spectra, in *individual prosodic parameters* and in *voice source parameters*. Certain *temporally structured parameters*, such as jitter, creak, or individually distinctive manners of producing voice on- and offsets, are also of interest for voice quality control. Further parameters, particularly *high-frequency components of the glottal waveform*, in all probability also contribute to individual voice quality [27], but have so far remained under-researched.

It is evident that no single parameter corresponds to that elusive propensity that is "voice quality". The challenge is to understand and to learn to manipulate the strength and interactions between the various layers of the multidimensional parameter space that emerges here. While this seems to be a formidable task, there is some reason to take heart: while the parameters reviewed in this article are unlikely to exhaust the inventory of relevant indicators of voice quality, they are likely to play a major role in any future attempt to understand the voice quality pattern of human speech.

---

[14] See the following quote from [20]: "The contrastive use of voice quality for vowels or consonants is fairly common in South East Asian, South African, and Native American Languages, and these have been the focus of a number of studies carried out at UCLA. Although both vowels and consonants may employ voice quality contrasts in a given language, Ladefoged (1982) points out that it is very rare to find contrasts at more than one place in a syllable."

Finally, the arguments here suggest a manner of integrating voice quality into a larger speech processing system. We can illustrate this for voice quality control in speech synthesis. The traditional structure of a speech synthesis system consists of three processing levels: text processing, prosody processing, and signal generation (Figure 5). The close articulation between linguistic and individual prosodic parameters suggest that voice quality control should probably be implemented in conjunction with prosody processing. In a model expanded to handle voice quality, processing components for linguistic prosody, individual prosody, spectral divergence, voice source parameters, plus certain temporally structured parameters would probably take the place of the single traditional prosody module. Initially, this prosody-plus-voice quality tier could be conceived as an additive model with weights assigned to each component. In time, a more complex, integrated model is conceivable, capable of handling interactions between prosodic and voice quality components.

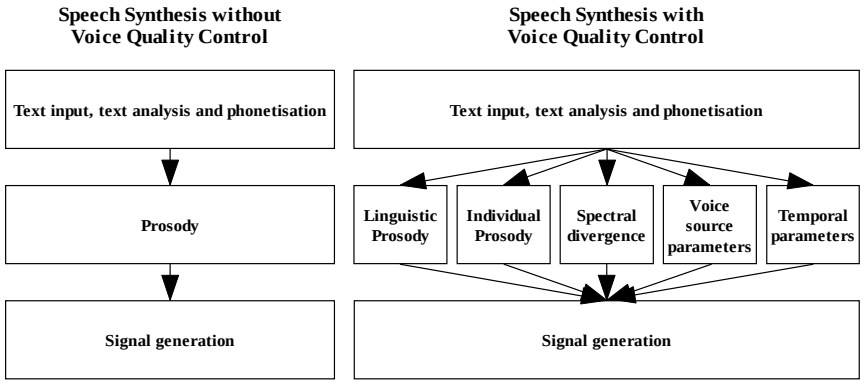

Fig. 5. Integration of voice quality control in speech synthesis. A set of processing components for linguistic prosody, individual prosody, spectral divergence, voice source parameters and temporally structured measures take the place of the single traditional prosody module. Initially, this could be conceived as an additive model with weights assigned to each component. In time, a more complex, integrated model is conceivable, capable of handling interactions between components.

Acknowledgements

The present research was supported by funding obtained from the Federal Office for Education and Science (BBW/OFES), Berne, Switzerland, in the framework of the European COST 277 project entitled “Non-linear Speech Processing”. Grateful acknowledgements are made to Dr. Brigitte Zellner Keller (Universities of Lausanne and Berne) for her critical reading of this contribution, and to Antonio Bonafonte (UPC Barcelona) for a number of stimulating suggestions.

References

1. Pittam, J.: Voice in Social Interaction: An Interdisciplinary Approach. Language and Language Behaviors, Volume 5. (1994)
2. Laver, J.: The Phonetic Description of Voice Quality. Cambridge University Press. (1980)
3. Laver, J.: The Description of Voice Quality in General Phonetic Theory. In: Laver, J. (eds.): The Gift of Speech. Edinburgh University Press (1991) 184-208
4. Fritzell, B., Hallén, O, Sundberg, J. Evaluation of Teflon Injection Procedures for Paralytic Dysphonia. Folia Phoniatrica, 26 (1974) 414-421.
5. Nordenberg, M., Sundberg, J. Effect on LTAS of Vocal Loudness Variation. TMH/QPSR, 1/2001. (2003). Available at: http://www.speech.kth.se/qpsr/tmh/2003/03-45-093-100.pdf.
6. Leino, T. Long-term Average Spectrum Study on Speaking Voice Quality in Male Actors. In: Friberg, A., Iwarsson, J., Jansson, E., Sundberg, J. (eds.) SMAC 93 (Proceedings of the Stockholm Music Acoustics Conference, 1993). Stockholm: Publication No. 79, Royal Swedish Academy of Music (1994) 206-210.
7. Klasmeyer, G.: An Automatic Description Tool for Time-contours and Long-term Average Voice Features in Large Emotional Speech Databases. SpeechEmotion-2000 (2000) 66-71
8. Keller, E.: Voice Characteristics of MARSEC Speakers. VOQUAL: Voice Quality: Functions, Analysis And Synthesis (2003).
9. Gobl, C., Bennet, E., Ní Chasaide, A. Expressive Synthesis: How Crucial is Voice Quality. Proceedings of the IEEE Workshop on Speech Synthesis. Santa Monica, CA (2002) Paper 52: 1-4.
10. Besacier, L. Un modèle parallèle pour la reconnaissance automatique du locuteur. Doctoral Thesis, University of Avignon, France (1998)
11. Zetterholm, E. A Comparative Survey of Phonetic Features of two Impersonators. Fonetik, 44 (2002) 129-132
12. Nolan, F., & Oh, T. Identical Twins, Different Voices. Forensic Linguistics 3 (1996) 39-49
13. Loakes, D. (2003) A Forensic Phonetic Investigation into the Speech Patterns of Identical and Non-Identical Twins. Proceedings of 15th ICPhS. Barcelona. ISBN 1-876346-48-5 (2003) 691- 694
14. Zellner Keller, B. Prosodic Styles and Personality Styles: are the two Interrelated? Proceedings of SP2004. Nara, Japan. (2004) 383-386
15. Rothenberg, M.: A New Inverse-filtering Technique for Deriving the Glottal Air Flow Waveform During Voicing. J. Acoust. Soc. Am., 53 (1973) 1632-1645.
16. Fourcin, A. Electrolaryngographic Assessment of Vocal Fold Function. Journal of Phonetics, 14 (1986) 435-442.
17. Fant G.: Acoustic Theory of Speech Production. The Hague: Mouton. (1960)
18. Fant, G.: Glottal Flow: Models and Interaction. Journal of Phonetics, 14, (1986) 393-399
19. Fant, G.: Swedish Vowels and a New Three-Parameter Model. TMH/QPSR, 1/2001. (2001). Available at: http://www.speech.kth.se/qpsr/tmh/2001/01-42-043-049.pdf
20. Ní Chasaide, A., Gobl, C.: Voice Source Variation. In W.J. Hardcastle, Laver, J. (eds.): The Handbook of Phonetic Sciences. Blackwell (1997) 427-461
21. Gobl, C.: The Voice Source in Speech Communication. Doctoral Thesis, KTH Stockholm, Sweden. (2003).
22. Gobl, C.: Speech Production. Voice Source Dynamics in Connected Speech. STL-QPSR 1/1988. (1988). 123-159.
23. Strik, H., Cranen, B., Boves, L.: Fitting a LF-model to Inverse Filter Signals. EUROSPEECH-93, Berlin, Vol. 1 (1993) 103-106
24. McKenna, J.G. Automatic Glottal Closed-Phase Location and Analysis by Kalman Filtering. 4th ISCA Tutorial and Research Workshop on Speech Synthesis, SSW4 Proceedings, Perthshire Scotland, 2001.

25. Fu, Q., & Murphy, P. A robust glottal source model estimation technique. 8th International Conference on Spoken Language Processing ICSLP, Jeju Island, Korea, 2004.
26. Fant, G., Liljencrants, J., Lin, Q.: A four-parameter model of glottal flow. STL-QPSR, No. 4/1985 (1985).
27. Plumpe, M. D., Quatieri, T. F., Reynolds D. A. Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification. IEEE Trans. on Speech and Audio Processing, Vol. 1 (1999) 569-586.