16. The Human Voice in Speech and Singing

This chapter describes various aspects of the human voice as a means of communication in speech and singing. From the point of view of function, vocal sounds can be regarded as the end result of a three stage process: (1) the compression of air in the respiratory system, which produces an exhalatory airstream, (2) the vibrating vocal folds' transformation of this air stream to an intermittent or pulsating air stream, which is a complex tone, referred to as the voice source, and (3) the filtering of this complex tone in the vocal tract resonator. The main function of the respiratory system is to generate an overpressure of air under the glottis, or a subglottal pressure. Section 16.1 describes different aspects of the respiratory system of significance to speech and singing, including lung volume ranges, subglottal pressures, and how this pressure is affected by the ever-varying recoil forces. The complex tone generated when the air stream from the lungs passes the vibrating vocal folds can be varied in at least three dimensions: fundamental frequency, amplitude and spectrum. Section 16.2 describes how these properties of the voice source are affected by the subglottal pressure, the length and stiffness of the vocal folds and how firmly the vocal folds are adducted. Section 16.3 gives an account of the vocal tract filter, how its form determines the frequencies of its resonances, and Sect. 16.4 gives an account for how these resonance frequencies or formants shape the vocal sounds by imposing spectrum peaks separated by spectrum valleys, and how the frequencies of these peaks determine vowel and voice qualities. The remaining sections of the chapter describe various aspects of the acoustic signals used for vocal communication in speech and singing. The syllable structure is discussed in Sect. 16.5, the closely related aspects of rhythmicity and timing in speech and singing is described in Sect. 16.6, and pitch and rhythm aspects in Sect. 16.7. The impressive control of all these acoustic characteristics of vocal signals is discussed in Sect. 16.8, while Sect. 16.9 considers expressive aspects of vocal communication.

16.1 Breathing

The process of breathing depends both on mechanical and muscular forces (Fig. 16.1).

During *inspiration* the volume of the chest cavity is expanded and air rushes into the lungs. This happens mainly because of the contraction of the *external intercostals* and the *diaphragm*. The external intercostal muscles raise the ribs. The diaphragm which is the dome-shaped muscle located below the lungs, flattens on contraction and thus lowers the floor of the thoracic cavity.

The respiratory structures form an elastic mechanical system that produces expiratory or inspiratory subglottal pressures, depending on the size of the lung volume, Fig. 16.2. Thus, exhalation and inhalation will always produce forces whose effect is to move the ribs and the lungs back to their resting state, often referred

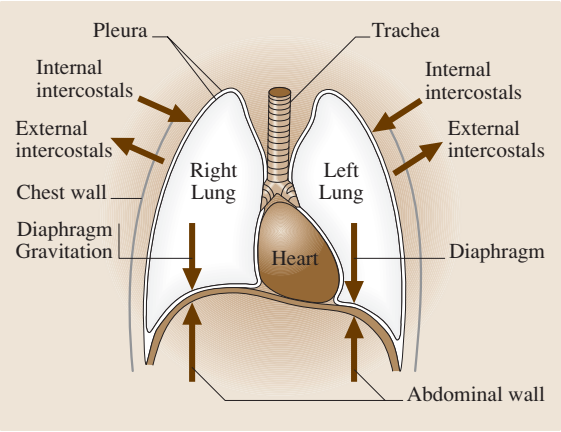

Fig. 16.1 Directions of the muscular and elastic forces. The direction of the gravitation force is applicable to the upright position

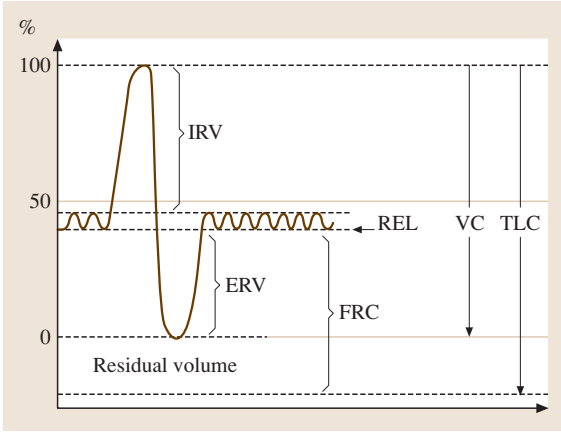

Fig. 16.3 Definition of the various terms for lung volumes. The graph illustrates the lung volume changes during quiet breathing interrupted by a maximal inhalation followed by a maximal exhalation. VC is the vital capacity, TLC is the total lung capacity, IRV and ERV are the inspiratory and expiratory reserve volume, REL is the resting expiratory level, FRC is the functional residual capacity

to as the resting expiratory level (REL). The deeper the breath, the greater this force of *elastic recoil*. This component plays a significant part in pushing air out of the lungs, both in speech and singing, and especially at large lung volumes. The elasticity forces originate both from the rib cage and the lungs. As illustrated in Fig. 16.2, the rib cage produces an exhalatory force at high lung volumes and an inhalatory force at low lung volumes, and the lungs always exert an exhalatory force. As a consequence, activation of inspiratory muscles is needed for producing a low subglottal pressure, e.g. for singing a soft (pianissimo) tone, at high lung volume. Conversely, activation of expiratory muscles is needed for producing a high subglottal pressure, e.g. for singing a loud (fortissimo) tone, at low lung volume.

In addition to mechanical factors, exhaling may involve the activity of the *internal intercostals* and the *abdominal muscles*. Contraction of the former has the effect of lowering the ribs and thus compressing the chest cavity. Activating the abdominal muscles generates upward forces that also contribute towards reducing the volume of the rib cage and the lungs. The function of these muscles is thus expiratory.

Another significant factor is gravity whose role depends on body posture. In an upright position, the diaphragm and adjacent structures tend to be pulled down, increasing the volume of the thoracic cavity. In this sit-

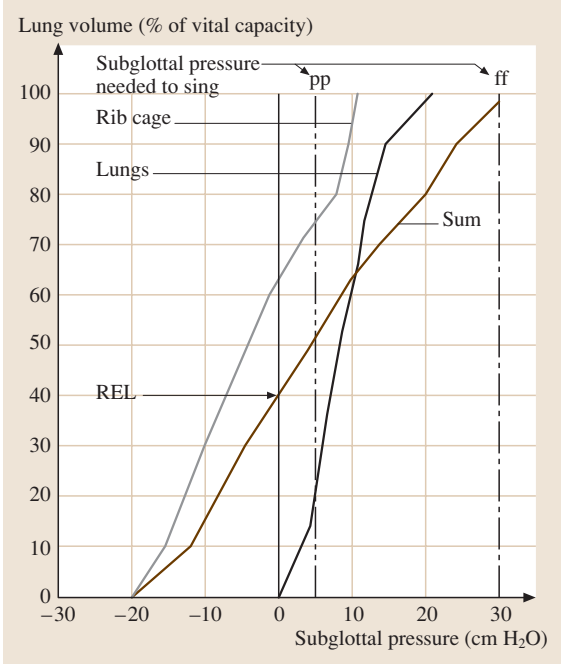

Fig. 16.2 Subglottal pressures produced at different lung volumes in a subject by the recoil forces of rib cage and lungs. The resting expiratory level (REL) is the lung volume at which the inhalatory and the exhalatory recoil forces are equal. The *thin* and *heavy chain-dashed lines* represent subglottal pressures typically needed for soft and very loud phonation

uation, the effect of gravity is inspiratory. In contrast, in the supine position the diaphragm tends to get pushed up into the rib cage, and expiration is promoted [16.1].

The total air volume that is contained in a maximally expanded rib cage is called the *total lung capacity* (TLC in Fig. 16.3). After maximum exhalation a small air volume, the *residual volume*, is still left in the airways. The greatest air volume that can be exhaled after a maximum inhalation is called the *vital capacity* (VC) and thus equals the difference between the TLC and the residual volume. The lung volume at which the exhalatory and inhalatory recoil forces are equal, or REL, is reached after a relaxed sigh, see Figs. 16.2 and 16.3. During tidal breathing inhalation is initiated from REL, so that inhalation is active resulting from an activation of inspiratory muscles, and exhalation is passive, produced by the recoil forces. In tidal breathing only some 10% of VC is inhaled, such that a great portion of the VC, the *inspiratory reserve volume*, is left. The air volume between REL and the residual volume is called the *expiratory reserve volume*.

VC varies depending on age, body height and gender. At the age of about 20 years, an adult female has a vital capacity of 3–3.6 l depending on body height, and for males the corresponding values are about 4–5.5 l.

Experimental data [16.4] show that, during tidal breathing, lung volume variations are characterized by a regular quasi-sinusoidal pattern with alternating segments of inspiration and expiration of roughly equal duration (Fig. 16.4). In speech and singing, the pattern is transformed. Inspirations become more rapid and expiration occurs at a slow and relatively steady rate. Increasing vocal loudness raises the amplitude of the lung volume records but leaves its shape relatively unchanged.

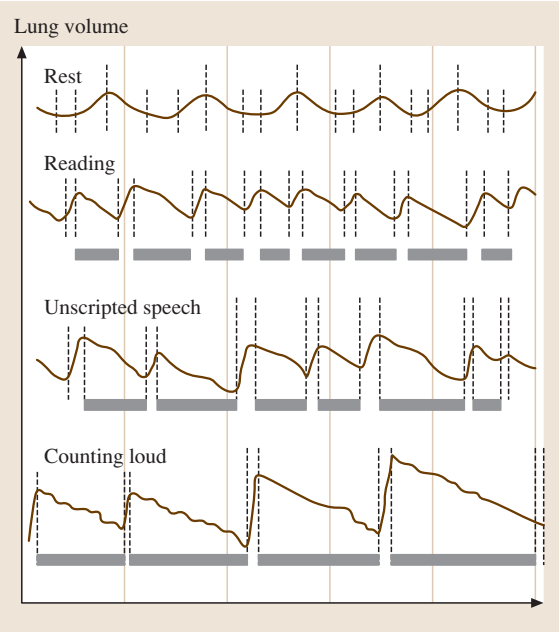

Fig. 16.4 Examples of speech breathing

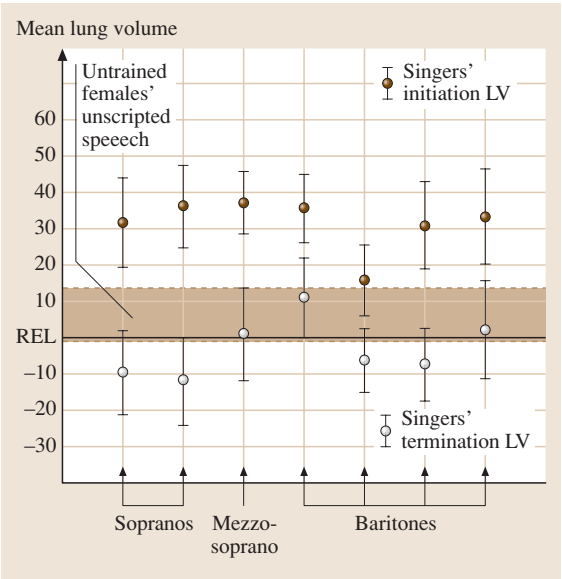

Fig. 16.5 Lung volume averages used in speech and operatic singing expressed as percentages of vital capacity relative to the resting expiratory level REL. The shaded band represents the mean lung volume range observed in untrained female voices' unscripted speech (after [16.2]). The filled and open symbols show the corresponding measures for professional opera singers of the indicated classifications when performing opera arias according to *Thomasson* [16.3]. The bars represent ± one SD

Figure 16.5 shows mean lung volumes used by professional singers when singing well-rehearsed songs. The darker band represents the mean lung volume range observed in spontaneous speech [16.2]. The lung volumes of conversational speech are similar to those in tidal breathing. Loud speech shows greater air consumption and thus higher volumes (Fig. 16.4). Breath groups in speech typically last for 3–5 seconds and are terminated when lung volumes approach the relaxation expiratory level REL, as illustrated in Fig. 16.5. Thus, in phonation, lung volumes below REL are mostly avoided.

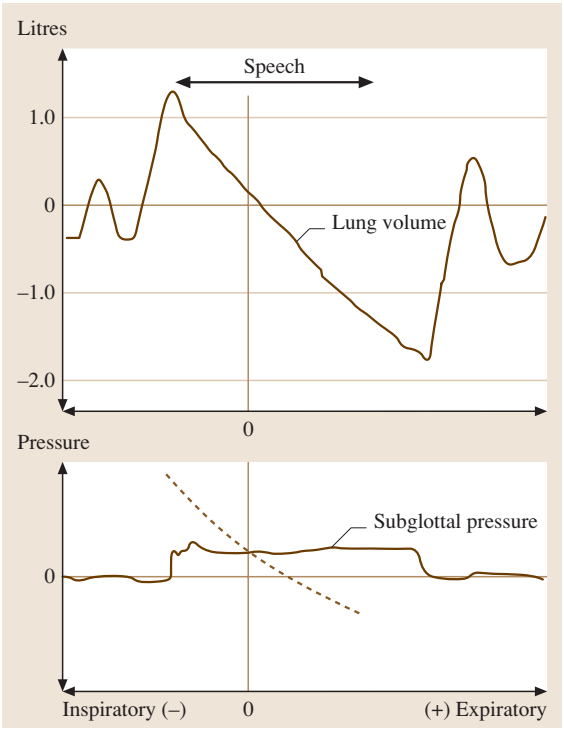

Fig. 16.6 The records show lung volume (relative to the mid-respiratory level), subglottal (esophageal) pressure and stylized muscular activity for speaker counting from one to 32 at a conversational vocal effort. (After Draper et al. [16.5]) . To the *left* of the vertical line recoil forces are strongly expiratory, to the *right* they are inspiratory. *Arrows* have been added to the *x*-axis to summarize the original EMG measurements which indicate that the recoil forces are balanced by muscular activity (EMG = electromyography, measurement of the electrical activity of muscles). To the *left of the vertical line* (as indicating by *left-pointing arrow*) the net muscular force is inspiratory, to the *right* it is expiratory (*right-pointing arrow*). To keep loudness constant the talker maintains a stable subglottal pressure and recruits muscles according to the current value of the lung volume. This behavior exemplifies the phenomenon known as *motor equivalence*

In singing, breath groups tend to be about twice as long or more, and air consumption is typically much greater than in conversational speech. Mostly they are terminated close to the relaxation expiratory level, as in speech, but sometimes extend into the *expiratory reserve volume* as illustrated in Fig. 16.5. This implies that, in singing, breath groups typically start at much higher lung volumes than in speech. This use of high lung volumes implies that singers have to deal with much greater recoil forces than in speech.

Figure 16.6 replots, in stylized form, a diagram published by Ladefoged et al. [16.7]. It summarizes measurements of lung volume and subglottal pressure, henceforth referred to as Ps, recorded from a speaker asked to take a deep breath and then to start counting. The dashed line intersecting the Ps record represents the relaxation pressure. This line tells us that elastic recoil forces are strongly expiratory to the left of the vertical line. To the right they produce a pressure lower than the target pressure and eventually an inspiratory pressure. The Ps curve remains reasonably flat throughout the entire utterance.

The question arises: how can this relative constancy be achieved despite the continually changing contribution of the relaxation forces? In a recent replication of this classical work [16.8], various criticisms of the original study are addressed but basically the original answer to this question is restated: the motor system adapts to the external goal of keeping the Ps fairly constant (Fig. 16.6). Initially, when recoil forces are strong, muscle activity is predominantly in the *inspiratory* muscles such as the *diaphragm* and the *external intercostals*. Gradually, as recoil forces decline, the *ex-*

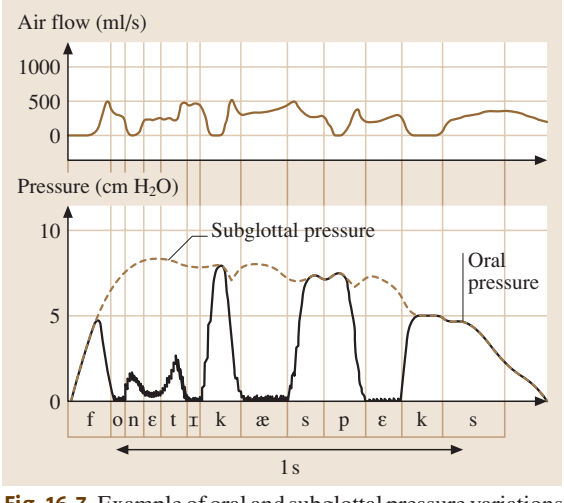

Fig. 16.7 Example of oral and subglottal pressure variations for the phrase "*phonetic aspects*" (after *Netsell* [16.6]). As the glottis opens and closes for voiceless and voiced sounds and the vocal tract opens or closes for vowels and consonants, the expired air is opposed by varying degrees of glottal and supraglottal impedance. Both the oral and the subglottal pressure records reflect the combined effect of these resistance variations

piratory muscles (the *internal intercostals*, the *rectus abdominis* among others) take over increasingly as the other group relaxes (cf. arrows, Fig. 16.6). According to our present understanding [16.8, 9], this adaptation of breathing for speech is achieved by constantly updating the balance between agonist and antagonist muscles in accordance with current conditions (lung volume, body posture, etc.) and in order to meet the goal of constant Ps. Muscle recruitment depends on lung volume.

Figure 16.7 presents a representative example of oral and Ps records. The phrase is "*phonetic aspects*" (After *Netsell* [16.6]). The top panel shows a record of oral air flow. Vertical lines indicate acoustic segment boundaries. The bottom diagram superimposes the curves for oral and subglottal pressure.

The Ps shows a falling pattern which becomes more pronounced towards the end of the phrase and which is reminiscent of the declination pattern of the fundamental frequency contour typical of declarative sentences [16.10, p. 127]. The highest values occur at the midpoints of [ɛ] and [æ], the stressed vowels of the utterance. For vowels, oral pressure is close to atmospheric pressure (near zero on the *y*-axis cm H_2O scale). The [k] of *phonetic* and the [p] of *aspects* show highly similar traces. As the tongue makes the closure for [k], the air flow is blocked and the trace is brought down to zero. This is paralleled by the oral pressure rising until it equals the Ps. As the [k] closure progresses, a slight increase builds up in both curves. The release of the [k] is signaled by a peak in the air flow and a rapid decrease in Ps. An almost identical pattern is seen for [p].

In analyzing Ps records, phoneticians aim at identifying variations based on an *active control* of the respiratory system and phenomena that can be attributed to the system's *passive response* to ongoing activity elsewhere, e.g., in the vocal tract and or at the level of the vocal folds [16.11].

To exemplify passive effects let us consider the events associated with [k] and [p] just discussed. As suggested by the data of Figs 16.4 and 16.6, speech breathing proceeds at a relatively steady lung volume decrement. However, the open or closed state of the glottis, or the presence of a vocal tract constriction/closure, is capable of creating varying degrees of impedance to the expired air. The oral pressure record reflects the combined effect of glottal and articulatory resistance variations. Ps is also affected by such changing conditions. As is evident from Fig. 16.7, the Ps traces during the [k] and [p] segments first increase during the stop closures. Then they decrease rapidly during the release and the aspiration phases. These effects are passive responses to the segment-based changes in supraglottal resistance and are unlikely to be actively programmed [16.12, 13].

Does respiration play an active role in the production of stressed syllables? In "*phonetic aspects*" main stress occurs on [ɛ] and [æ]. In terms of Ps, these vowels exhibit the highest values. Are they due to an active participation of the respiratory system in signaling stress, or are they fortuitous by-products of other factors?

An early contribution to research on breathing and speech is the work by *Stetson* [16.14]. On the basis of aerodynamic, electromyographic and chest movement measurements, Stetson proposed the notion of the *chest pulse*, a chunk of expiratory activity corresponding to the production of an individual syllable.

In the late fifties, Ladefoged and colleagues published an electromyographic study [16.7] which cast doubt on Stetson's interpretations. It reported increased activity in expiratory muscles (*internal intercostals*) for English syllables. However, it failed to support Stetson's chest pulse idea in that the increases were found only on stressed syllables. *Ladefoged* [16.8] reports findings from a replication of the 1958 study in which greater activity in the internal intercostals for stressed syllables was confirmed. Moreover, reduced activity in inspiratory muscles (*external intercostals*) was observed to occur immediately before each stressed syllable.

Ps measurements provide further evidence for a positive role for respiration in the implementation of stress. *Ladefoged* [16.15, p. 143] states: "Accompanying every stressed syllable there is always an increase in the Ps". This claim is based on data on disyllabic English noun–verb pairs differing in the position of the main stress: *TORment* (noun)–*torMENT* (verb), *INsult* (noun)–*inSULT* (verb) etc. A clear distinction between the noun and verb forms was observed, the former having Ps peaks in the first syllable and the latter on the second syllable.

As for segment-related Ps variations (e.g., the stops in Fig. 16.7), there is wide agreement that such local perturbations are induced as automatic consequences of speech production aerodynamics. Explaining short-term ripple on Ps curves in terms of aerodynamics is consistent with the observation that the respiratory system is mechanically sluggish and therefore less suited to implement rapid Ps changes in individual phonetic segments. Basically, its primary task is to produce a Ps contour stable enough to maintain vocal intensity at a fairly constant level (Fig. 16.6).

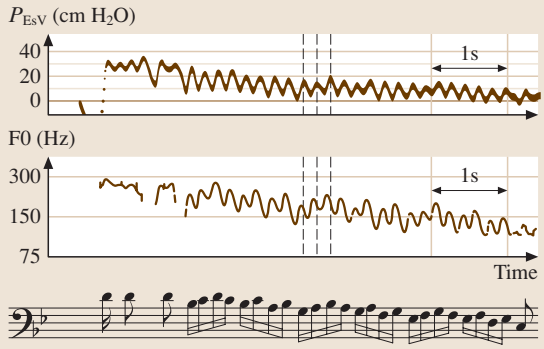

Fig. 16.8 Changes in esophageal pressure, corresponding to subglottic pressure changes (*upper curve*) and phonation frequency (*middle*) in a professional baritone singer performing the coloratura passage shown at the bottom. (After *Leanderson* et al. [16.17])

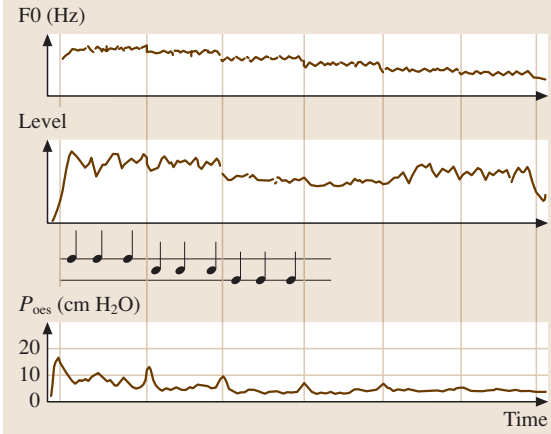

Fig. 16.9 The three *curves* show, from *top to bottom* F0 sound level and esophageal pressure in a professional baritone singer singing the music example shown in the *graph*, i. e. a descending scale with three tones on each scale tone. The singer marked the first beat in each bar by a pressure pulse. (After *Sundberg* et al. [16.18])

Singing provides experimental data on the behavior of the respiratory system that tend to reinforce the view that a time varying respiration activity can be an active participant in the sound generation process. Although there is ample justification for saying that the mechanical response of the respiratory structures is characterized by a long time constant, Ps in singing varies quickly and accurately. One example is presented in Fig. 16.8. It shows variations in pressure, captured as the pressure variations in the esophagus and thus mirroring the Ps variations [16.16], during a professional baritone singer's performance of the music example shown at the bottom. Each tone in the sequence of sixteenth notes is produced with a pressure pulse. Note also that in the fundamental frequency curve each note corresponds to a small rise and fall. The tempo of about six sixteenth notes per second implies that the duration of each rise–fall cycle is about 160 ms. It would take quite special skills to produce such carefully synchronized Ps and fundamental frequency patterns.

Synthesis experiments have demonstrated that this particular fundamental frequency pattern is what produces what is perceived as a sung legato coloratura [16.19, 20]. The voice organ seems incapable of producing a stepwise-changing F0 curve in legato, and such a performance therefore sounds as if it was produced by a music instrument rather than by a voice. In this sense this F0 variation pattern seems to be needed for eliciting the perception of a sung sequence of short legato tones.

The pressure variations are not likely to result from a modulation of glottal adduction. A weak glottal adduction should result in a voice source dominated by the fundamental and with very weak high spectrum partials, and mostly the amplitudes of the high overtones do not vary in coloratura sequences.

It is quite possible that the F0 variations are caused by the Ps variations. In the example shown in Fig. 16.8, the pressure variations amount to about 10 cm H_2O, which should cause a F0 modulation of about 30 Hz or so, corresponding to two semitones in the vicinity of 220 Hz. This means that the coloratura pattern may simply be the result of a ramp produced by the F0 regulating system which is modulated by the pressure pulses produced by the respiratory system.

As another example of the skilled use of Ps in singing, Fig. 16.9 shows the pressure variation in the esophagus in a professional baritone singer performing a descending scale pattern in 3/4 time, with three tones on each scale step as shown by the score fragment in the first three bars. The singer was instructed to mark the first tone in each bar. The pressure record demonstrates quite clearly that the first beat was produced with a marked pulse approximately doubling the pressure. Pulses of this magnitude must be produced by the respiratory apparatus. When the subject was instructed to avoid marking the first tone in each bar, no pressure pulses were observed [16.18].

The effect of Ps on fundamental frequency has been investigated in numerous studies. A commonly used

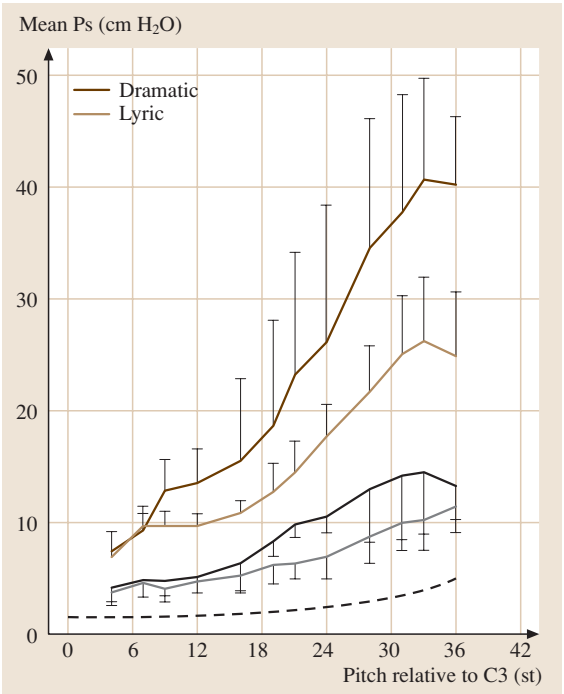

Fig. 16.10 Mean subglottal pressure, captured as the oral pressure during /p/-occlusion, in five lyric and six dramatic professional sopranos (*heavy* and *thin curves*, respectively) who sang as loudly and as softly as possible at different F0. The *dashed curve* shows Titze's prediction of threshold pressure, i. e., the lowest pressure that produces vocal fold vibration. The *bars* represent one SD (After [16.21])

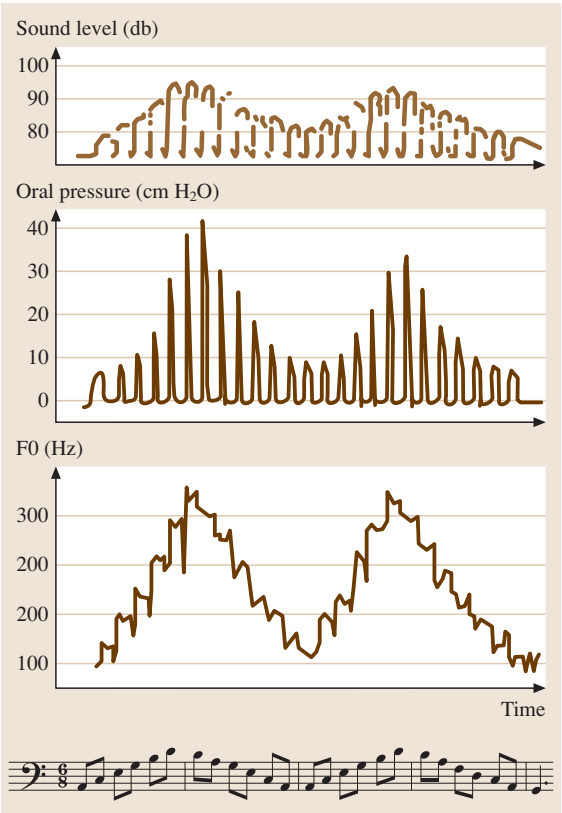

Fig. 16.11 Simultaneous recordings of sound level, subglottic pressure (captured as the oral pressure during /p/ occlusion) and F0 in a professional singer singing the note example shown at the bottom. (After *Leanderson* et al. [16.17])

method is to ask the subject to produce a sustained vowel at a given steady pitch and then, at unpredictable moments, change the Ps by applying a push to the subject's abdomen or chest. The assumption underlying these studies is that there will be an initial interval during which possible reflex action of laryngeal muscles will not yet come into play. Hence the data from this time segment should give a valid picture of the relationship between fundamental frequency and Ps.

In a study using the push method, *Baer* [16.22] established this relationship for a single male subject. Data on steady phonations at 94, 110, 220 Hz and a falsetto condition (240 Hz) were collected. From electromyograms from the *vocalis* and the *interarytenoid* it was possible to tell that the first 30 ms after the push onset were uncontaminated by reflex muscle responses. Fundamental frequency was plotted against Ps for the four F0 conditions. These plots showed linear data clusters with slopes of 4, 4, 3 and 9 Hz/cm H_2O respectively. Other studies which assumed a longer reflex response latency (100 ms) report that F0's dependence on Ps occurs between 2 to 7 Hz/cm H_2O.

From such observations phoneticians have concluded that, in producing the F0 variations of natural speech, Ps plays only a secondary role. F0 control is primarily based on laryngeal muscle activity. Nonetheless, the falling F0 contours of phrases with statement intonation tend to have Ps contours that are also falling. Moreover, in questions with final F0 increases, the Ps records are higher towards the end of the phrase than in the corresponding statements [16.15].

There is clear evidence that the Ps needs to be carefully adapted to the target fundamental frequency in singing, especially in the high range ([16.23], [16.24,

p. 36]). An example is presented in Fig. 16.10 which shows mean Ps for five dramatic and five lyric sopranos when they sang pp and ff tones throughout their range [16.21]. The bars represent one standard deviation. The dashed curve is the threshold pressure, i. e. the lowest pressure that produced vocal fold vibration, for a female voice according to *Titze* [16.25, 26].

The singers mostly used more than twice as high Ps values when they were singing their loudest as compared with their softest tones. The subjects reach up to 40 cm H_2O. This can be compared with values typical of normal adult conversational speech, which tend to occur in the range of 5–10 cm H_2O. However, loud speech and particularly stage speech may occasionally approach the higher values for singing.

Also interesting is the fact that lyric sopranos use significantly lower Ps values than dramatic sopranos both in soft and loud phonation. It seems likely that this difference reflects difference in the mechanical properties of the vocal folds.

Figure 16.11 presents another example of the close link between Ps and fundamental frequency. From top to bottom it plots the time functions for sound level (SPL in dB), pressure (cm H_2O), fundamental frequency (Hz) for a professional baritone singing an ascending triad followed by a descending dominant-seventh triad (according to the score at the bottom). The pressure record was obtained by recording the oral pressure as the subject repeated the syllable [pæ] on each note. During the occlusion of the stop the peak oral pressure becomes equal to the Ps (as discussed in connection with Fig. 16.7). That means that the peak values shown in the middle diagram are good estimates of the actual Ps. It is important to note that, when the trace repeatedly returns to a value near zero, what we see is not the Ps, but the oral pressure reading for the [æ] vowel, which is approximately zero.

The exercise in Fig. 16.11 is often sung staccato, i. e. with short pauses rather than with a /p/ consonant between the tones. During these pauses the singer has to get ready for the next fundamental frequency value and must therefore avoid exhaling the pressurized air being built up in the lungs. Singers do this by opening the glottis between the tones and simultaneously reducing their Ps to zero, so that no air will be exhaled during the silent intervals between the tones. A remarkable fact demonstrated here is that, particularly when sung loudly – so that high Ps values are used – this exercise requires nothing less than a virtuoso mastering of both the timing and tuning of the breathing apparatus and the pitch control process. A failure to reach a target pressure is likely to result in a failure to reach the target F0.

16.2 The Glottal Sound Source

In speech and singing the general method to generate sound is to make a constriction and to let a strong flow of air pass through it. The respiratory component serves as the power source providing the energy necessary for sound production. At the glottis the steady flow of air generated by the respiratory component is transformed into a quasiperiodic series of glottal pulses. In the vocal tract, the glottally modified breath stream undergoes further modifications by the resonance characteristics of the oral, pharyngeal and nasal cavities.

Constrictions are formed at the glottis - by adjusting the separation of the vocal folds – and above the glottis – by positioning the articulators of the vocal tract. As the folds are brought close together, they respond to the air rushing through by initiating an open–close vibration and thereby imposing a quasiperiodic modulation of airflow. Thus, in a manner of speaking, the glottal structures operate as a device that imposes an AC modulation on a DC flow. This is basically the way that *voicing*, the sound source of voiced vowels and consonants and the carrier of intonation and melody, gets made.

A second mechanism is found in the production of *noise*, the acoustic raw materials for voiceless sounds (e.g., [f], [s], [p], [k]). The term refers to irregular turbulent fluctuations in airflow which arise when air comes out from a constriction at a high speed. This process can occur at the glottis – e.g., in [h] sounds, whisper and breathy voice qualities – or at various places of articulation in the vocal tract.

The framework for describing both singing and speech is that of the *source-filter theory of speech production* [16.27, 28]. The goal of this section is to put speech and singing side by side within that framework and to describe how the speaker and the singer coordinate respiration, phonation and articulation to shape the final product: the acoustic wave to be perceived by the listener.

Figure 16.12 [16.29] is an attempt to capture a few key aspects of vocal fold vibrations. At the center a sin-

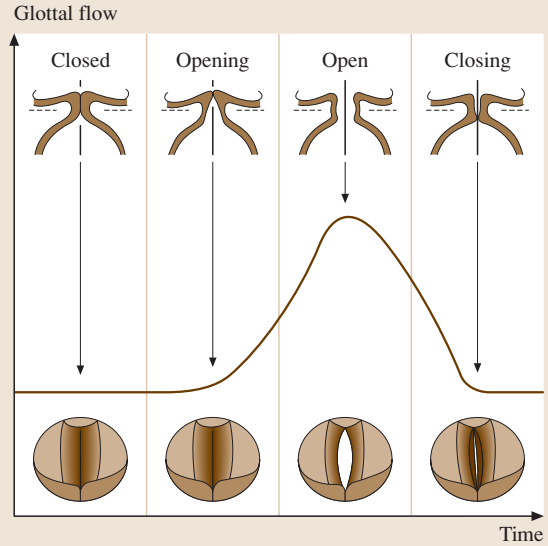

Fig. 16.12 Relationship between the flow glottogram, showing transglottal airflow versus time, and glottal configurations in a coronal plane (*upper series* of images) and from above, as through a laryngeal mirror (*lower series* of images). The airflow increases when the folds open the glottis and allow air to pass and decreases quickly when they close the glottis and arrest the airflow. (After [16.29])

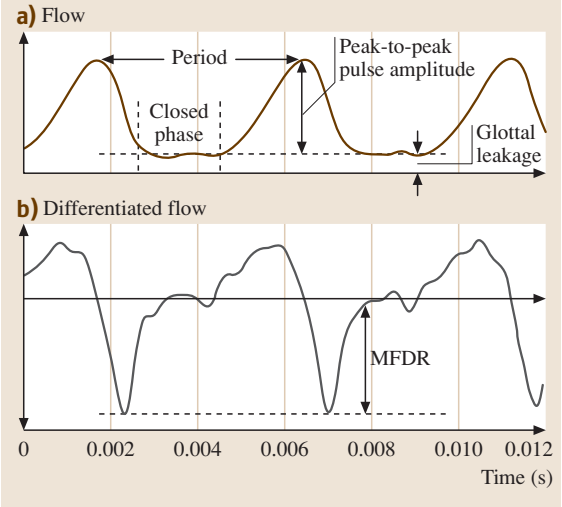

Fig. 16.13a,b Illustration of measures commonly used to characterize a flow glottogram **(a)** and its time derivative, the differentiated flow glottogram **(b)**. (The *wiggles* in the latter are artifacts caused by an imperfect inverse filtering)

gle cycle of a glottal waveform is seen. It plots airflow through the glottis as a function of time. Alternatively, the graph can be used to picture the time variations of the glottal area which present a pattern very similar to that for airflow. The top row shows stylized cross sections of the vocal folds at selected time points during the glottal cycle. From left to right they refer to the opening of the folds, the point of maximum area and the point of closure. Below is a view of the vocal folds from above corresponding to the profiles at the top of the diagram.

There are a number of different methods for visualizing vocal fold vibrations. By placing an electrode on each side of the thyroid cartilage, a minute current can be transferred across the glottis. This current increases substantially when the folds make contact. The resulting *electroglottogram*, also called a *laryngogram*, thus shows how the contact area varies with time. It is quite efficient in measurement of F0 and closed phase. *Optical glottograms* are obtained by illuminating the trachea from outside by means of a strong light source and capturing the light traveling through the glottis by means of an optical sensor in the pharynx. The signal therefore reflects the glottal area, but only as long as the light successfully finds it way to the sensor. A posterior tilting of the epiglottis may easily disturb or eliminate the signal.

Flow glottograms show transglottal airflow versus time and are derived by inverse filtering the audio signal, often picked up as a flow signal by means of a pneumotachograph mask [16.30]. Inverse filtering implies that the signal is passed through a filter with a transfer function equalling the inverted transfer function of the vocal tract. Therefore correct inverse filtering requires that the inverted resonance peaks of the inverse filter are tuned to the formant frequencies of the vowel being filtered.

As transglottal airflow is zero when the glottis is closed and nonzero when it is open, the flow glottogram is physiologically relevant. At the same time it is a representation of the sound of the voice source.

A typical example of a flow glottogram is given in the upper graph of Fig. 16.13. The classical parameters derived from flow glottograms are the durations of the period and of the closed phase, pulse peak-to-peak amplitude, and glottal leakage. The lower graph shows the differentiated glottogram. The negative peak amplitude is often referred to as the *maximum flow declination rate* (MFDR). As we shall see it has special status in the process of voice production.

In the study of both speech and singing, the acoustic parameter of main relevance is the time variations in sound pressure produced by the vocal system and received by the listener's ears. Theoretically, this signal

is roughly proportional to the derivative of the output airflow at the lips [16.28, 31] and it is related to the derivative of the glottal waveform via the transfer function of the vocal tract. Formally, the excitation signal for voiced sounds is defined in terms of this differentiated signal. Accordingly, in source-filter theory, it is the derivative of glottal flow that represents the source and is applied to the filter or resonance system of the vocal tract. The amplitude of the vocal tract excitation, generally referred to as the *excitation strength*, is quantified by the maximum velocity of flow decrease during vocal-fold closing movement (the MFDR, Fig. 16.13) which is a determinant of the level of the radiated sound. At the moment of glottal closure a drastic modification of the air flow takes place. This change is what generates voicing for both spoken and sung sounds and produces a sound with energy across a wide range of frequencies.

The Liljencrants–Fant (LF) model [16.32, 33] is an attempt to model glottal waveforms using parameters such as fundamental frequency, excitation strength, dynamic leakage, open quotient and glottal frequency (defined by the time period of glottal opening phase). Other proposals based on waveform parameters have been made by Klatt and Klatt [16.34], Ljungqvist and Fujisaki [16.35], Rosenberg [16.36] and Rothenberg et al. [16.37]. A second line of research starts out from assumptions about vocal fold mechanics and applies aerodynamics to simulate glottal vibrations [16.38, 39]. Insights from such work indicate the importance of parameters such as Ps, the adducted/abducted position of vocal folds and their stiffness [16.28].

During the early days of speech synthesis it became clear that the simplifying assumption of a constant voice source was not sufficient to produce high-quality natural-sounding copy synthesis. Experimental work on the voice source and on speech synthesis has shown that, in the course of an utterance, source parameters undergo a great deal of variation. The determinants of this dynamics are in part prosodic, in part segmental. Figure 16.14 [16.32] presents a plot of the time variations of the *excitation strength* parameter (i. e. MFDR) during the Swedish utterance: *Inte i DETta århundrade* [ɪntɪ ˣdɛtːaoːrhəndradɛ]. The upper-case letters indicate that the greatest prominence was on the first syllable of *detta*. Vertical lines represent acoustic segment boundaries.

Gobl collected flow data using the mask developed by Rothenberg [16.30] and applied inverse filtering to obtain records of glottal flow which, after differentiation, enabled him to make measurements of excitation strength and other LF parameters.

Figure 16.14 makes clear that excitation strength is in no way constant. It varies depending on both prosodic and segmental factors. The effect of the segments is seen near the consonant boundaries. As the vocal tract is constricted, e.g., in [d] and [tː], and as transglottal pressure therefore decreases (cf. the pressure records of Fig. 16.7), excitation strength is reduced. In part these variations also occur to accommodate the voicing and the voicelessness of the consonant [16.28]. This influence of consonants on the voice source has been documented in greater detail by *Ni Chasaide* and *Gobl* [16.40] for German, English, Swedish, French, and Italian. Particularly striking effects were observed in the context of voiceless consonants.

Prosodically, we note in Fig. 16.14 that excitation strength exhibits a peak on the contrastively stressed syllable in *detta* and that the overall pattern of the phrase is similar to the falling declination contour earlier mentioned for declarative statements.

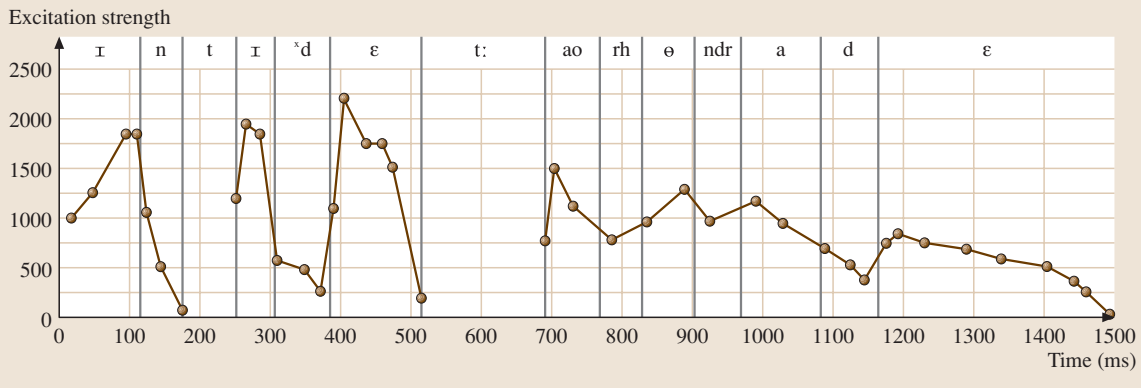

Fig. 16.14 Running speech data on the 'excitation strength parameter' of the LF model. (After [16.32])

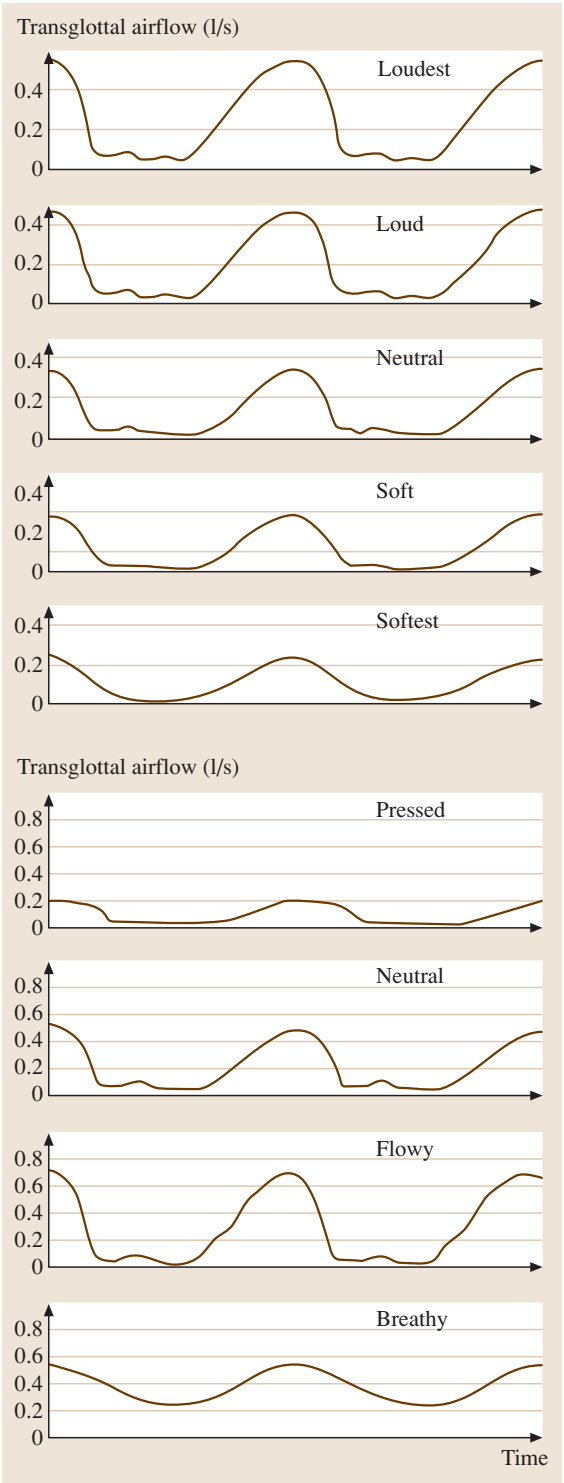

Table 16.1 Measurements of subglottal pressure (cm H_2O) and SPL at 0.3 m (dB) measured for the glottograms shown in Fig. 16.15

Table A	Ps (cm H_2O)	SPL at 0.3 m(dB)
loudest	14.3	84
loud	7.3	82
neutral	5.7	78
soft	3.9	75
softest	3	70
Table B	**Ps (cm H_2O)**	**SPL at 0.3 m(dB)**
pressed	11.4	83
neutral	5.1	79
flowy	8	88
breathy	6.6	84

Examples of the fact that the Ps has a strong influence on the flow glottogram are given in the upper set of graphs of Fig. 16.15, which shows a set of flow glottograms for phonations produced at the same pitch but with varying degrees of *vocal loudness*. As we examine the series of patterns from loudest to softest we note that both the peak flow and the maximum steepness of the trailing end of the pulse, i. e., MFDR, increase significantly with increasing Ps. These shape changes are lawfully related to the variations in Ps and are directly reflected in sound pressure levels as indicated by the numbers in Table 16.1.

Holmberg and colleagues [16.41] made acoustic and airflow recordings of 25 male and 20 female speakers producing repetitions of [pæ] at soft, normal and loud vocal efforts [16.42, p. 136]. Estimates of Ps and glottal airflow were made from recordings of oral pressure and oral airflow. Ps was derived by interpolating between peak oral pressures for successive [p] segments and then averaging across repetitions. A measure of average flow was obtained by low-pass filtering the airflow signal and averaging values sampled at the vowel midpoints.

Fig. 16.15 Typical flow glottogram changes associated with changes of loudness or mode of phonation (*top* and *bottom series*). As loudness of phonation is raised, the closing part of the curve becomes more steep. When phonation is pressed, the glottogram amplitude is low and the closed phase is long. As the adduction force is reduced, pulse amplitude grows and the closed phase becomes shorter (note that the flow scales are different for the *top* and *bottom series* of glottograms. Flowy phonation is the least adducted, and yet not leaky phonation

A copy of the airflow signal was low-pass filtered and inverse filtered to remove the effect of F1 and other formants. The output was then differentiated for the purpose of determining the MFDR (Fig. 16.13).

Figure 16.15 also illustrates how the voice source can be continuously varied between different *modes of phonation*. These modes range from hyperfunctional, or pressed, over neutral and flowy to hypofunctional, or breathy. The corresponding physiological control parameter can be postulated to be glottal adduction, i. e. the force by which the folds press against each other. It varies from minimum in hypofunctional to extreme in hyperfunctional. Flowy phonation is produced with the weakest degree of glottal adduction compatible with a full glottal closure. The physiologically relevant property that is affected is the vibration amplitude of the vocal folds, which is small in hyperfunctional/pressed phonation and wide in breathy phonation.

As illustrated in Fig. 16.15 the flow glottogram is strongly affected by these variations in phonation mode [16.43]. In pressed phonation, the pulse amplitude is small and the closed phase is long. It is larger in neutral and even more so in flowy. In breathy phonation, typically showing a waveform similar to a sine wave, airflow is considerable, mainly because of a large leakage, so there is no glottal closure.

Phonation mode affects the relation between Ps and the SPL of the sound produced. As shown in Table 16.1B pressed phonation is less economical from an acoustic point of view: a Ps of 11.4 cm H_2O produces an SPL at 0.3 m of only 83 dB, while in flowy phonation a lower Ps produces a higher SPL.

Pitch, loudness and phonation mode are voice qualities that we can vary continuously. By contrast, vocal registers, also controlled by glottal parameters, appear more like toggles, at least in untrained voices. The voice is operating either in one or another register. There are at least three vocal registers, *vocal fry*, *modal* and *falsetto*. When shifting between the modal and falsetto registers, F0 discontinuities are often observed [16.44].

The definition of vocal registers is quite vague, a set of tones along the F0 continuum that sound similar and are felt to be produced in a similar way. As registers depend on glottal function, they produce different flow glottogram characteristics. Figure 16.16 shows typical examples of flow glottograms for the falsetto and modal registers as produced by professional baritone, tenor and countertenor singers. The pulses are more rounded, the closed phase is shorter, and the glottal leakage is greater in the falsetto than in the modal register. However, the waveform of a given register often varies substantially between individuals. Classically trained sopranos, altos, and tenors learn to make continuous transitions between the modal and the falsetto registers, avoiding abrupt changes in voice timbre.

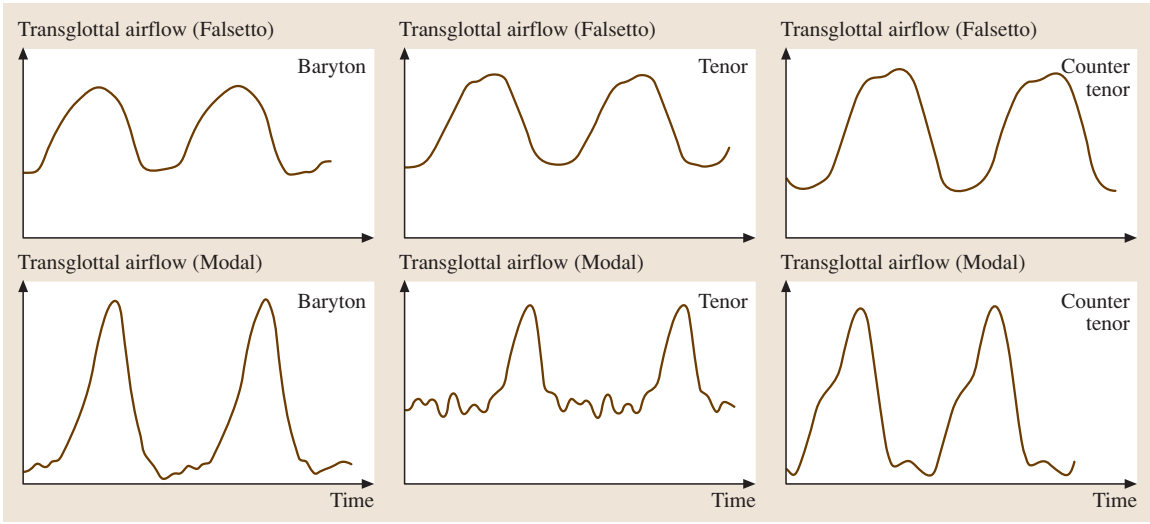

Fig. 16.16 Typical flow glottograms for falsetto and modal register in a baritone, tenor and countertenor singer, all professional. The flow scale is the same within subjects. The ripple during the closed phase in the lower middle glottogram is an artifact. In spite of the great inter-subject variability, it can be seen that the glottogram pulses are wider and more rounded in the falsetto register

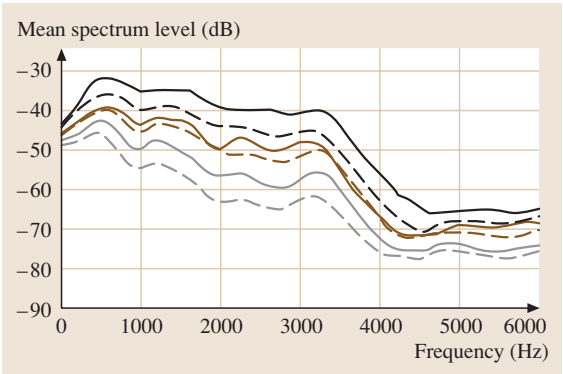

Fig. 16.17 Long-term-average spectra curves obtained from an untrained male speaker reading the same text at 6 different degrees of vocal loudness. From *top to bottom* the corresponding L_{eq} values at 0.3 m were 93 dB, 88 dB, 85 dB, 84 dB, 80 dB, and 76 dB

Variation of vocal loudness affects the spectrum slope as illustrated in Fig. 16.17, which shows long-term-average spectra (LTAS) from a male untrained voice. In the figure loudness is specified in terms of the so-called equivalent sound level L_{eq}. This is a commonly used time average of sound level, defined as

$$L_{eq} = 10 \log \frac{1}{T} \int_0^T \frac{p^2}{p_0^2} \, dt \, ,$$

where t is time and T the size of the time window. p and p_0 are the sound pressure and the reference pressure, respectively.

When vocal loudness is changed, the higher overtones change much more in sound level than the lower overtones. In the figure, a 14 dB change of the level near 600 Hz is associated with a 22 dB change near 3000 Hz, i. e., about 1.5 times the level change near 600 Hz. Similar relationships have been observed for professional singers [16.47]. In other words, the slope of the voice source spectrum decreases with increasing vocal loudness.

The physiological variable used for variation of vocal loudness is Ps. This is illustrated in the upper graph of Fig. 16.18, comparing averaged data observed in untrained female and male subjects and data obtained from professional operatic baritone singers [16.45, 46]. The relationship between the Ps and MFDR is approximately linear. It can be observed that the pressure range used by the singer is considerably wider than that used by the untrained voices. The MFDR produced with a given Ps by the untrained female and male subjects is mostly

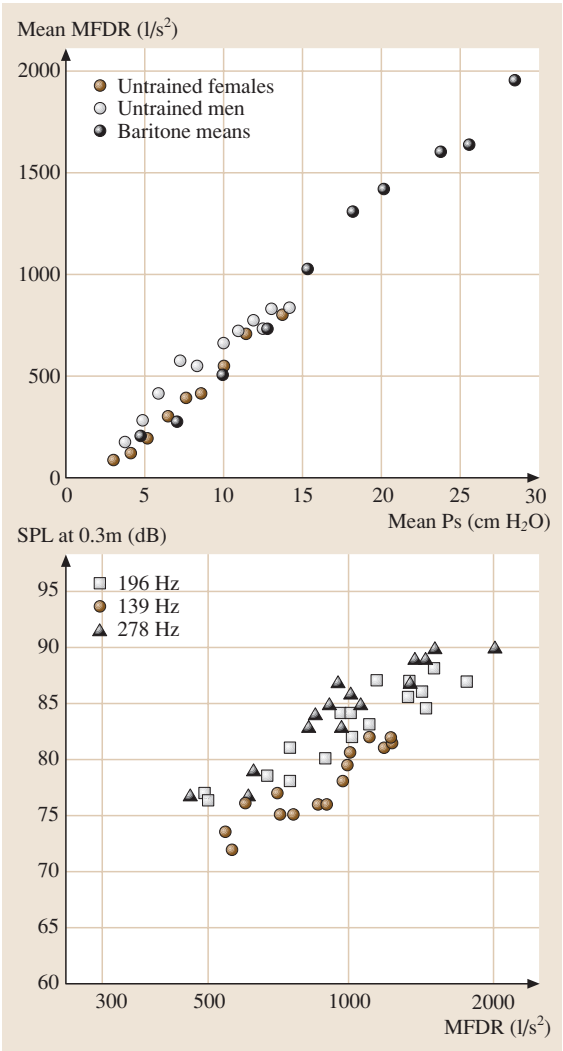

Fig. 16.18 The *top* graph shows the relationship between the mean subglottal pressure and the mean MFDR for the indicated subject groups. The *bottom* graph shows the relationship between MFDR and the SPL at 0.3 m for a professional baritone singing the vowels /a/ and /æ/ at different F0s. (After [16.45] p. 183, [16.46] p. 184)

higher than that produced by the baritones with the same pressure.This may depend on different mechanical characteristics of the vocal folds.

As we will see later, SPL depends on the strength of the excitation of the vocal tract, i. e. on MFDR. This variable, in turn, depends on Ps and F0; the higher the pressure, the greater the MFDR value and the higher the F0, the greater the MFDR. The top graph of Fig. 16.18

shows how accurately MFDR could be predicted from Ps and F0 for previously published data for untrained male and female singers and for professional baritone singers [16.45, 46]. Both Ps and F0 are linearly related to MFDR. However, the singers showed a much greater variation with F0 than the untrained voices. This difference reflected the fact that unlike the untrained subjects the singers could sing a high F0 much more softly than the untrained voices. The ability to sing high notes also softly would belong to the essential expressive skills of a singer. Recalling that an increase of Ps increases F0 by a few Hz/cm H_2O, we realize that singing high tones softly requires more forceful contraction of the pitch-raising laryngeal muscles than singing such tones loudly.

16.3 The Vocal Tract Filter

The source-filter theory, schematically illustrated in Fig. 16.19, describes vocal sound production as a three-step process: (1) generation of a steady flow of air from the lungs (DC component); (2) conversion of this airflow into a pseudo-periodically pulsating transglottal airflow (DC-to-AC conversion), referred to as the voice source; and (3) response of the vocal tract to this excitation signal (modulation of AC signal) which is characterized by the frequency curve or transfer function of the vocal tract. So far the first two stages, respiration and phonation, have been considered.

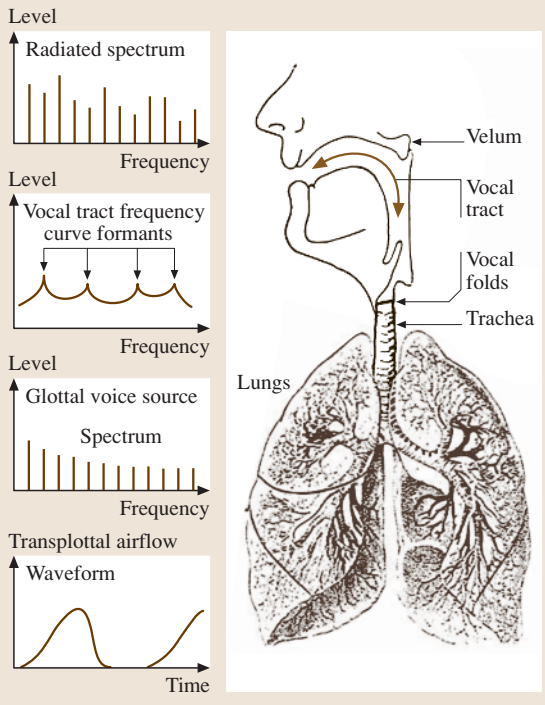

Fig. 16.19 Schematic illustration of the generation of voice sounds. The vocal fold vibrations result in a sequence of voice pulses (*bottom*) corresponding to a series of harmonic overtones, the amplitudes of which decrease monotonically with frequency (*second from bottom*). This spectrum is filtered according to the sound transfer characteristics of the vocal tract with its peaks, the formants, and the valleys between them. In the spectrum radiated from the lip opening, the formants are depicted in terms of peaks, because the partials closest to a formant frequency reach higher amplitudes than neighboring partials

In this section we will discuss the third step, viz. how the vocal tract filter, i. e. the resonance characteristics of the vocal tract, modifies, and to some extent interacts with, the glottal source and shapes the final sound output radiated from the talker's/singer's lips.

Resonance is a key feature of the filter response. The oral, pharyngeal and nasal cavities of the vocal tract form a system of resonators. During each glottal cycle the air enclosed by these cavities is set in motion by the glottal pulse, the main moment of excitation occurring during the closing of the vocal folds, more precisely at the time of the MFDR, the maximum flow declination rate (cf. the previous section on source).

The behavior of a vocal tract resonance, or *formant*, is specified both in the time and the frequency domains. For any transient excitation, the time response is an exponentially decaying cosine [16.27, p. 46]. The frequency response is a continuous amplitude-frequency spectrum with a single peak. The shape of either function is uniquely determined by two numbers (in Hz): the formant frequency F and the bandwidth B. The bandwidth quantifies the degree of damping, i. e., how fast the formant oscillation decays. Expressed as sound pressure variations, the time response is

$$p(t) = A\,\mathrm{e}^{-\pi Bt}\cos(2\pi Ft)\ . \tag{16.1}$$

For a single formant curve, the amplitude variations as a function of frequency f is given (in dB) by

$$L(f) = 20\log\frac{\left[F^2+\left(\frac{B}{2}\right)^2\right]}{\sqrt{(f-F)^2+\left(\frac{B}{2}\right)^2}\sqrt{(f+F)^2+\left(\frac{B}{2}\right)^2}}\ . \tag{16.2}$$

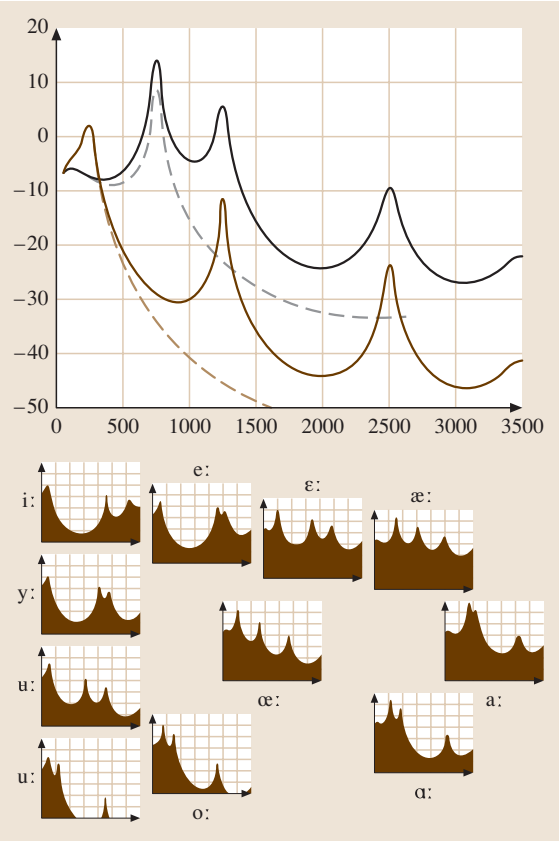

Fig. 16.20 Predictability of formant levels. Calculated spectral envelopes for a set of Swedish vowels. The *upper graph* illustrates how varying only the frequency of F1 affects the amplitudes of other formants

In the frequency domain the bandwidth is defined as the width of the formant 3 dB down from the peak, that is, at the half-power points. A large bandwidth produces a flat peak whereas a small value (less damping, reduced acoustic losses) makes the peak higher and sharper.

Figure 16.20 shows spectral envelopes for a set of Swedish vowels. They were calculated from formant frequency data [16.48, 49] in accordance with source-filter theory which assumes that a vowel spectrum can be decomposed into formants (all specified with respect to frequency and bandwidth), the spectrum of the voice source, the contribution of formants above F4 and radiation [16.27].

The individual panels are amplitude versus frequency plots. Vowels are portrayed in terms of their envelopes rather than as line spectra with harmonics. The panels are arranged in a *formant space*. From top to bottom the F2 in the panels decreases. From left to right F1 increases. The frequency scale is limited to showing the first three formant peaks.

We note that all envelopes have falling overall slopes and that the amplitudes of the formant peaks vary a great deal from vowel to vowel and depending on the relative configuration of formant frequency positions.

In acoustic phonetic specifications of vowels, it is customary to report no more than the frequencies of the first two or three formants ([16.50, p. 59], [16.51]). Experiments in speech synthesis [16.52] have indicated that this compact description is sufficient to capture the quality of steady-state vowels reasonably well. Its relative success is explained by the fact that most of the building blocks of a vowel spectrum are either predictable (bandwidths and formant amplitudes), or show only limited spectral variations (source, radiation, higher-formant correction) [16.27].

Formant bandwidths [16.28, 53] reflect acoustic losses. They depend on factors such as radiation, sound transmission through the vocal tract walls, viscosity, heat conduction, constriction size as well as the state of the glottis. For example, a more open glottis, as in a breathy voice, will markedly increase the bandwidth of the first formant.

Despite the complex interrelations among these factors, bandwidths pattern in regular ways as a function of formant frequency. Empirical formulas have been proposed [16.54] that summarize bandwidth measurements made using transient and sweep-tone excitation of the vocal tract for closed-glottis conditions [16.55, 56].

To better understand how formant levels vary let us consider the top diagram of Fig. 16.20. It compares envelopes for two vowel spectra differing only in terms of F1. It is evident that the lowering of F1 (from 750 to 250 Hz) reduces the amplitudes of F2 and F3 by about 15 dB. This effect is predicted by acoustic theory which derives the spectral envelope of an arbitrary vowel as a summation of individual formant curves on a dB scale [16.27]. Figure 16.20 makes clear that, as F1 is shifted, its contribution to the envelope is drastically changed. In this case the shift moves the entire F1 curve down in frequency and, as a result, its upper skirt (dashed line) provides less of a *lift* to the upper formants. This interplay between formant frequency and formant levels is the major determinant of the various envelope shapes seen in Fig. 16.20.

One of the main lessons of Fig. 16.20 is accordingly that, under normal conditions of a stable voice source, formant amplitudes are predictable. Another

important consequence of the source-filter theory is that, since knowing the formants will enable us to reconstruct the vowel's envelope, it should also make it possible to derive estimates about a vowel's overall intensity.

A vowel's intensity can be calculated from its power spectrum as

$$I = 10 \log \left[\sum (A_i)^2 \right] , \qquad (16.3)$$

where A_i is the sound pressure of the i-th harmonic. This measure tends to be dominated by the strongest partial or partials. In very soft phonation the strongest partial is generally the fundamental while in neutral and louder phonation it is the partial that lies closest to the first formant. Thus, typically all partials which are more than a few dB weaker than the strongest partial in the spectrum do not contribute appreciably to the vowel's intensity.

As suggested by the envelopes in Fig. 16.20, the strongest harmonics in vowels tend to be found in the F1 region. This implies that a vowel's intensity is primarily determined by its F1. Accordingly, in the set shown in Fig. 16.20, we would expect [iːyːɯːuː] to be least intense and [æːɑɑː] the most intense vowels. This is in good agreement with experimental observations [16.57].

The intrinsic intensity of vowels and other speech sounds has been a topic of interest to phoneticians studying the acoustic correlates of stress [16.58]. It is related to sonority, a perceptual attribute of speech sounds that tends to vary in a regular way in syllables [16.59]. We will return below to that topic in Sect. 16.5.

Let us continue to illustrate how formant levels depend on formant frequencies with an example from singing: the singer's formant. This striking spectral phenomenon is characteristic of classically trained male singers. It is illustrated in Fig. 16.21. The figure shows a spectrogram of a commercial recording of an operatic tenor voice (Jussi Björling) performing a *recitativo* from Verdi's opera Aida. Apart from the vibrato undulation of the partials, the high levels of partials in the frequency range 2200–3200 Hz are apparent. They represent the singer's formant.

The singer's formant is present in all voiced sounds as sung by operatic male singers. It was first discovered by *Bartholomew* [16.60]. It manifests itself as a high, marked peak in the long-term-average spectrum (LTAS).

A second example is presented in Fig. 16.22 showing an LTAS of the same tenor voice as in Fig. 16.21, accompanied by an orchestra of the traditional western

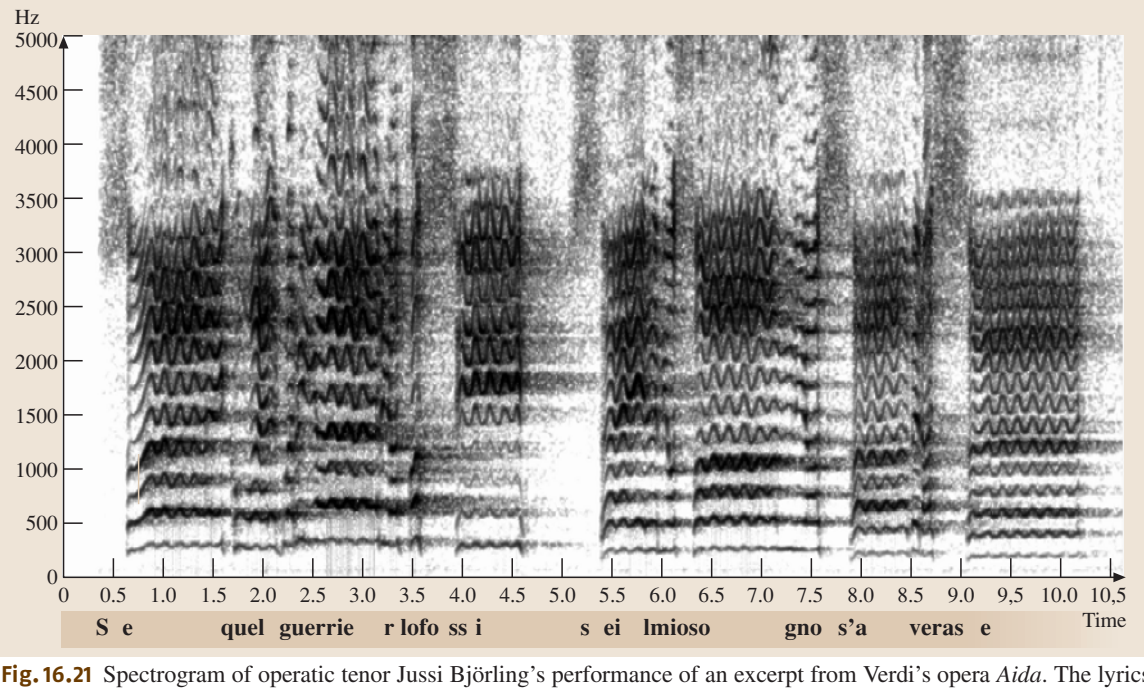

Fig. 16.21 Spectrogram of operatic tenor Jussi Björling's performance of an excerpt from Verdi's opera *Aida*. The lyrics are given below the graph

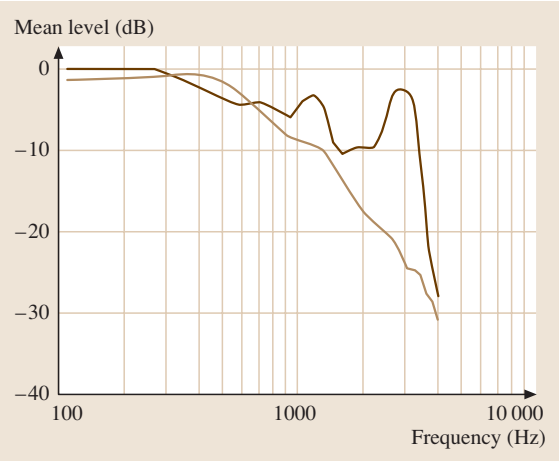

Fig. 16.22 LTAS of the sound of a symphonic orchestra with and without a singer soloist (*dark* and *light curves*). The singer's formant constitutes a major difference between the orchestra with and without the singer soloist. (After [16.61])

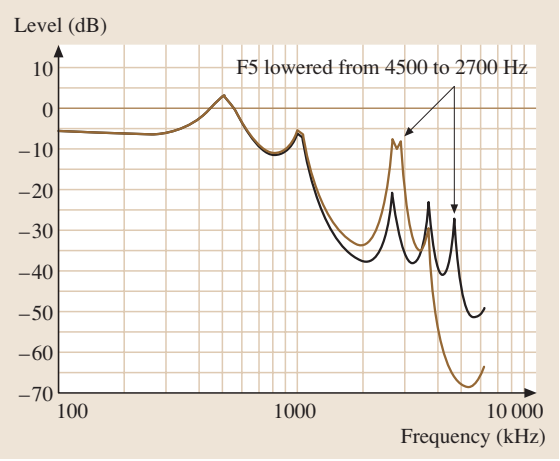

Fig. 16.23 Effect on the spectrum envelope of lowering formants F4 and F5 from 3500 Hz and 4500 Hz to 2700 Hz and 3500 Hz, respectively. The resulting gain in level at F3 is more than 10 dB

opera type. The LTAS peak caused by the singer's formant is just a few dB lower than the main peak in the low-frequency range.

In the same figure an LTAS of a classical symphony orchestra is also shown. On the average, the partials in the low-frequency range are quite strong but those above 600 Hz decrease by about 10 dB/octave with rising frequency. Even though this decrease varies depending on how loudly the orchestra is playing, this implies that the level of the orchestral accompaniment is much lower at 3000 Hz than at about 600 Hz. In other words, the orchestral sound offers the singer a rather reasonable competition in the frequency range of the singer's formant. The fact of the matter is that a voice possessing a singer's formant is much easier to hear when the orchestral accompaniment is loud than a voice lacking a singer's formant. Thus, it helps the singer's voice to be heard when the orchestral accompaniment is loud.

The singer's formant can be explained as a resonance phenomenon [16.62]. It is a product of the same rules that we invoked above to account for the formant amplitudes of vowels and for intrinsic vowel intensities. The strategy of a classically trained male singer is to shape his vocal tract so as to make F3, F4, and F5 form a tight cluster in frequency. As the frequency separations among these formants are decreased, their individual levels increase, and hence a high spectral peak is obtained between 2500 and 3000 Hz.

Figure 16.23 shows the effects on the spectrum envelope resulting from lowering F5 and F4 from 3500 Hz and 4500 Hz to 2700 Hz and 3500 Hz, respectively. The resulting increase of the level of F3 amounts to 12 dB, approximately. This means that male operatic singers produce a sound that can be heard more easily through a loud orchestral accompaniment by tuning vocal tract resonances rather than by means of producing an excessive Ps.

The acoustical situation producing the clustering of F3, F4, and F5 is obtained by acoustically mismatching the aperture of the larynx tube, also referred to as the epilaryngeal tube, with the pharynx [16.62]. This can be achieved by narrowing this aperture. Then, the larynx tube acts as a resonator with a resonance that is not much affected by the rest of the vocal tract but rather by the shape of the larynx tube. Apart from the size of the aperture, the size of the laryngeal ventricle would be influential: the larger the ventricle, the lower the larynx tube resonance. Presumably singers tune the larynx tube resonance to a frequency close to F3. The articulatory means used to establish this cavity condition seems mainly to be a lowering of the larynx, since this tends to widen both the pharynx and the laryngeal ventricle. Many singing teachers recommend students to sing with a comfortably low larynx position.

The level of the singer's formant is influenced also by the slope of the source spectrum which, in turn, depends on vocal loudness, i. e. on Ps, as mentioned. Thus, the singer's formant tends to increase by about 15 dB for a 10 dB change of the overall SPL [16.47].

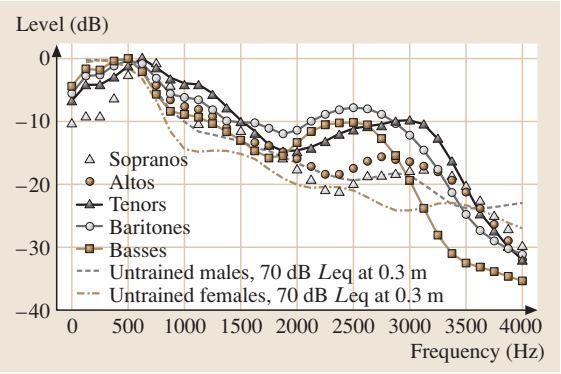

Fig. 16.24 Mean LTAS derived from commercial recordings of four representatives of each of the indicated voice classifications. The *dashed curves* show LTAS of untrained female (red) and male (blue) voices' speech with an L_{eq} of 70 dB at 0.3 m. Note that the singer's formant produces a prominent peak with a level well above the level of the untrained voices' LTAS only for the male singers

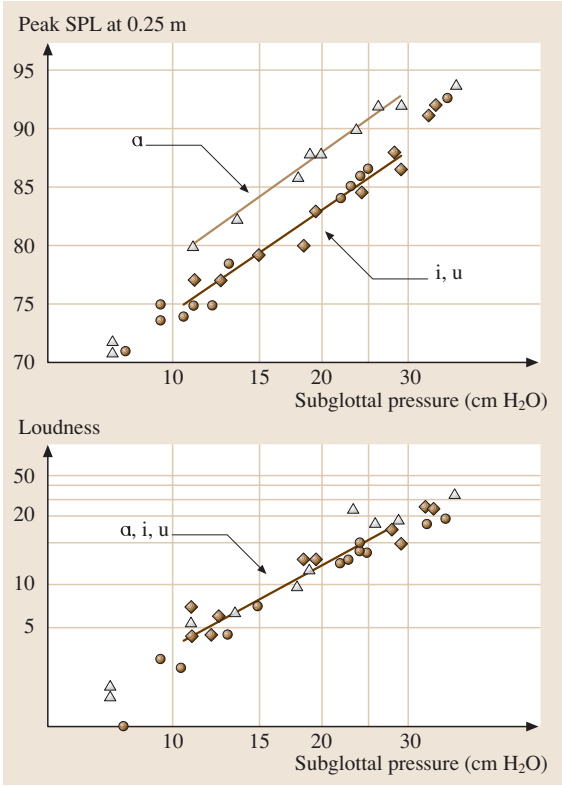

Fig. 16.25 Perceived loudness of a vowel (*bottom panel*) correlates better with subglottal pressure Ps than does SPL (*top panel*) which depends on the intrinsic intensity of the vowel. This, in turn, is determined by its formant pattern. (After [16.66])

The center frequency of the singer's formant varies slightly between voice classifications, as illustrated in the mean LTAS in Fig. 16.24 which shows mean LTAS derived from commercial recordings of singers classified as soprano, alto, tenor, baritone and bass [16.63]. Each group has four singers. The center frequency of the singer's formant for basses, baritones and tenors are about 2.4, 2.6, and 2.8 kHz, respectively. These small differences are quite relevant to the typical voice timbres of these classifications [16.64]. Their origin is likely to be vocal tract length differences, basses tending to have longer vocal tracts than baritones who in turn have longer vocal tracts than tenors [16.65]. On the other hand, substantial variation in the center frequency of the singer's formant occurs also within classifications.

Also shown for comparison in Fig. 16.24 is a mean LTAS of untrained voices reading at an L_{eq} of 70 dB at 0.3 m distance. The singer's formant produces a marked peak some 15 dB above the LTAS of the untrained voices for the male singers. The female singers, on the other hand, do not show any comparable peak in this frequency range. This implies that female singers do not have a singer's formant [16.67–69].

Female operatic singers' lack of a singer's formant is not surprising, given the fact that: (1) they sing at high F0, i. e. have widely spaced spectrum partials, and (2) the F3, F4, F5 cluster that produces a singer's formant is rather narrow in frequency. The latter means that a narrow formant cluster will be hit by a partial only in some tones of a scale while in other scale tones there will be no partial in the formant cluster. This would lead to salient and quasi-random variation of voice timbre between scale tones.

The singer's formant is a characteristic of classically trained male singers. It is not found in nonclassical singing, e.g. in pop or musical theater singing, where audibility is the responsibility of the sound engineer rather than of the singer. Likewise, choir singers generally do not possess a singer's formant.

From the evidence in Fig. 16.18 we concluded that the logarithm of Ps does a good job of predicting the SPL over a large range of vocal intensities. This was shown for untrained male and female speakers as well as a professional baritone singer. In the next few paragraphs we will look at how vowel identity affects that prediction and how listeners perceive loudness in the presence of vowel-dependent variations in SPL at constant Ps.

This topic was addressed by Ladefoged [16.66] who measured the peak SPL and the peak Ps in 12 repetitions of *bee*, *bay*, *bar*, *bore* and *boo* spoken by a British talker at varying degrees of loudness. SPL values were plotted against log(Ps) for all tokens. Straight lines could readily be fitted to the vowels individually. The vowels of *bay* and *bore* tended to have SPL values in between those of *bar* and *bee*/*boo*. The left half of Fig. 16.25 replots the original data for [ɑ] and [i]/[u] pooled. The [ɑ] line is higher by 5–6 dB as it should be in view of F1 being higher in [ɑ] than in [i] and [u] (cf. preceding discussion).

In a second experiment listeners were asked to judge the loudness of the test words. Each item was presented after a carrier phrase: *Compare the words: bar and* __. The word bar served as reference. The subjects were instructed to compare the test syllable and the reference in terms of loudness, to give the value of 10 to the reference and another relative number to the test word. The analysis of the responses in terms of SPL indicated that, for a given SPL, it was consistently the case that *bee* and *boo* tended to be judged as louder than *bar*. On the other hand, there were instances of *bee* and *bar* with similar Ps that were judged to be equally loud. This effect stood out clearly when loudness judgements were plotted against log(Ps) as in the bottom half of Fig. 16.25.

Ladefoged concludes that " ... in the case of speech sounds, loudness is directly related to the physiological effort" – in other words, Ps – rather than the SPL as for many other sounds.

Speech Sounds with Noise and Transient Excitation

A comprehensive quantitative treatment of the mechanisms of noise production in speech sounds is found in *Stevens* [16.28, pp. 100–121]. This work also provides detailed analyses of how the noise source and vocal tract filtering interact in shaping the bursts of stops and the spectral characteristics of voiced and voiceless fricatives. While the normal source mechanism for vowels always involves the glottis, noise generation may take place not only at the glottis but at a wide range of locations along the vocal tract. A useful rule of thumb for articulations excited by noise is that the output will spectrally be dominated by the cavity in front of the noise source. This also holds true for sounds with transient excitation such as stop releases and click sounds [16.28, Chapts. 7,8].

While most of the preceding remarks were devoted to the acoustics of vowels, we should stress that the source-filter theory applies with equal force to the production of consonants.

16.4 Articulatory Processes, Vowels and Consonants

X-ray films of normal speech movements reveal a highly dynamic process. Articulatory activity comes across as a complex flow of rapid, parallel lip, tongue and other movements which shows few, if any, steady states. Although the speaker may be saying no more than a few simple syllables, one nonetheless has the impression of a virtuoso performance of a *polyphonic* motor score. As these events unfold, they are instantly reflected in the acoustic output. The articulatory movements modify the geometry of the vocal tract. Accordingly, the filter (transfer function) undergoes continual change and, as a result, so do the output formant frequencies.

Quantitative modeling is a powerful tool for investigating the speech production process. It has been successfully applied to study the relations between formant and cavities. Significant insights into this mapping have been obtained by representing a given vocal tract configuration as an *area function* – i. e., a series of cylindrical cross sections of variable lengths and cross-sectional areas – and by simulating the effects of changes in the area function on the formant frequencies [16.27].

In the past, pursuing this approach, investigators have used lateral X-ray images – similar to the magnetic

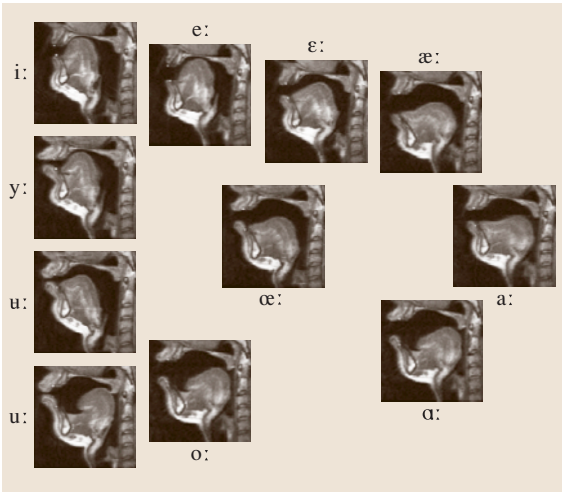

Fig. 16.26 Magnetic resonance images of the Swedish vowels of Fig. 16.20

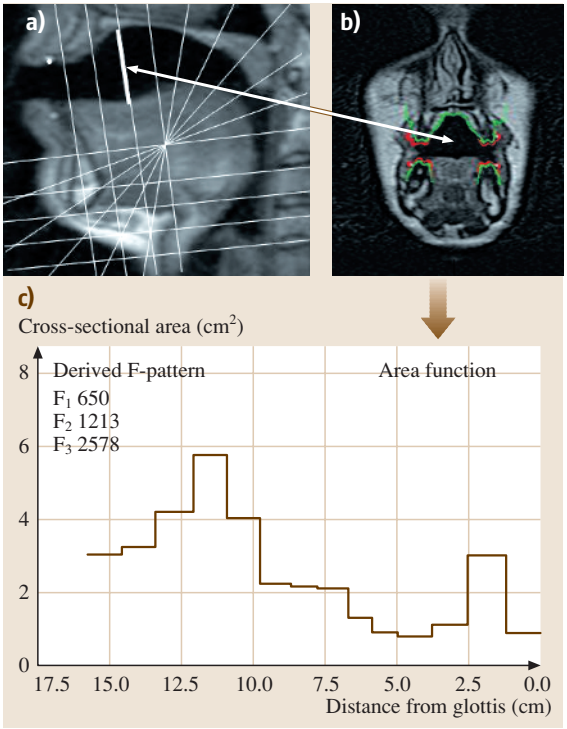

Fig. 16.27a–c From articulatory profile to acoustic output. Deriving the formant patterns of a given vowel articulation involves **(a)** representing the vocal tract profile in terms of the cross-distances from the glottis to the lips; **(b)** converting this specification into a cross-sectional area function; and **(c)** calculating the formant pattern from this area function. To go from **(a)** to **(b)**, profile data (*top left*) need to be supplemented by area measurements obtained from transversal (coronal and axial) images of the cross sections

resonance imaging (MRI) pictures in Fig. 16.26 – to trace the outlines of the acoustically relevant articulatory structures. To make estimates of cross-sectional areas along the vocal tract such lateral profiles need to be supplemented with information on the transverse geometry of the cross sections, e.g., from casts of the vocal tract [16.70] and tomographic pictures [16.27, 71].

More currently, magnetic resonance imaging methods have become available, making it possible to obtain three-dimensional (3-D) data on sustained steady articulations [16.72–76]. Figure 16.26 presents MRI images taken in the mid-sagittal plane of a male subject during steady-state productions of a set of Swedish vowels. These data were collected in connection with work on APEX, an articulatory model developed for studying the acoustic consequences of articulatory movements, e.g., the lips, the tongue and the jaw [16.77–81].

Figure 16.27 highlights some of the steps involved in deriving the formant pattern from lateral articulatory profiles such those of Fig. 16.26. The top-left picture is the profile for the subject's [a]. The white lines indicate the coronal, coronal oblique and axial planes where ad-

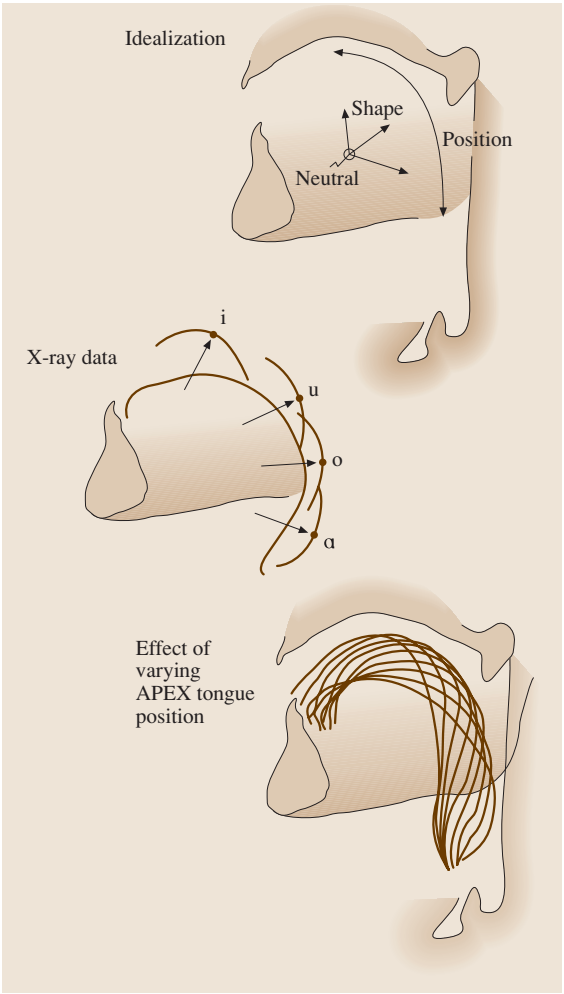

Fig. 16.28 The top drawing shows superimposed observed tongue shapes for [i], [u], [o] and [ɑ] differing in the place of the main constriction. Further analyses of similar data suggest that, for vowels, two main dimensions are used: The anterior–posterior location of the tongue body and its displacement from neutral which controls the degree of vocal tract constriction (*lower right*). These parameters are implemented numerically to produce vowels in the APEX articulatory model

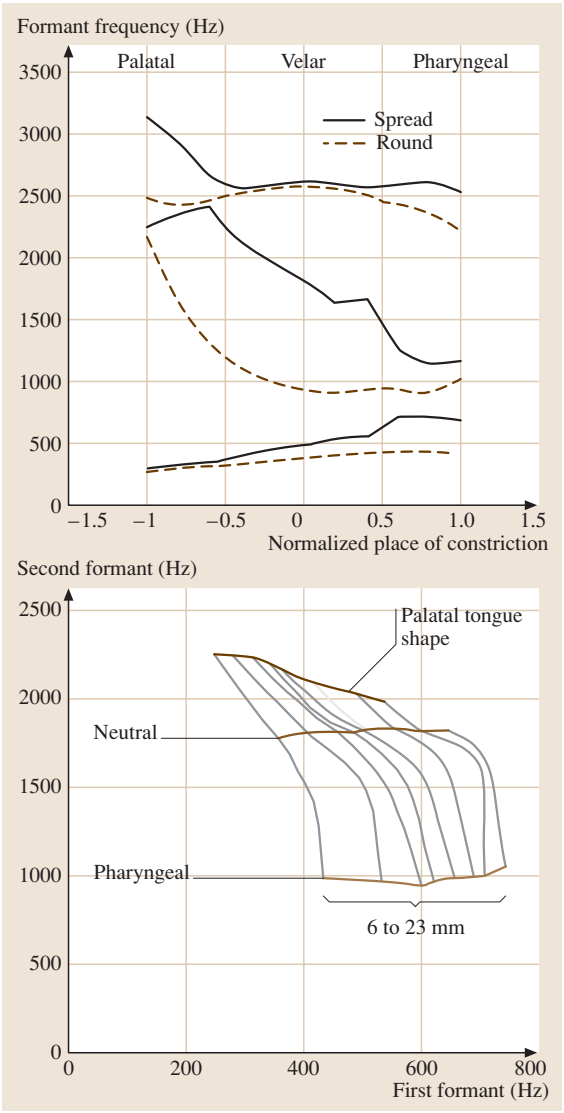

Fig. 16.29 Acoustic consequences of tongue body, jaw and lip movements as examined in APEX simulations. The tongue moves continuously between palatal, velar and pharyngeal locations while maintaining a minimum constriction area A_{min} of 0.25 cm, a 7 mm jaw opening and a fixed larynx height. The *dashed* and *solid lines* refer to rounded and spread conditions. The *lower diagram* illustrates the effect of varying the jaw. The *thin lines* show tongue movement at various degrees of jaw opening (6, 7, 8, 9, 12, 15, 19 and 23 mm). The *bold lines* pertain to fixed palatal, neutral or pharyngeal tongue contours

ditional MR images were taken to get information on the cross sections in the transverse dimension.

The location of the transverse cut is indicated by the bold white line segment in the lateral profile. In the coronal section to the right the airway of the vocal tract is the dark area at the center. Below it, the wavy surface of the tongue is evident. Since the teeth lack the density of hydrogen nuclei needed to show up on MR images, their intersection contours (green) were reconstructed from casts and added computationally to the image. The red lines indicate the outline of a dental plate custom-fitted to the subject's upper and lower teeth and designed to carry a contrast agent used to provide reference landmarks (tiny white dots) [16.82].

At any given point in the vocal tract, it would appear that changes in cross-sectional area depend primarily on the midsagittal distance from the tongue surface to the vocal tract wall. An empirically adequate approach to capturing distance-to-area relations is to use power functions of the form $A_x = \alpha d_x^{\beta}$, where d_x is the mid-sagittal cross distance at a location x in the vocal tract and α and β are constants whose values tend to be different in different regions and depend on the geometry of the vocal tract walls [16.24,70,72]. The cross-sectional area function can be obtained by applying a set of d-to-A rules of this type to cross distances measured along the vocal tract perpendicular to the midline. The final step consists in calculating the acoustic resonance frequencies of the derived area function.

In producing a vowel and the vocal tract shape appropriate for it, what are the main articulatory parameters that the speaker has to control? The APEX model [16.80] assumes that the significant information about the vowel articulations in Fig. 16.26 is the following. All the vowels exhibit tongue shapes with a single constriction. They differ with respect to how narrow this constriction is and where it is located. In other words, the vowels appear to be produced with control of two degrees of freedom: the palatal–pharyngeal dimension (*position*) and tongue height, or, in APEX terminology, *displacement* from neutral. This interpretation is presented in stylized form in Fig. 16.28 together with APEX tongue shapes sampled along the palatal–pharyngeal continuum. The choice of tongue parameters parallels the degree and place of constriction in the area-function models of *Stevens* and *House* [16.83] and *Fant* [16.27].

A useful application of articulatory models is that they can be set up to change a certain variable while keeping others constant [16.84]. Clearly, asking a human subject to follow such an instruction does not create an easily controlled task.

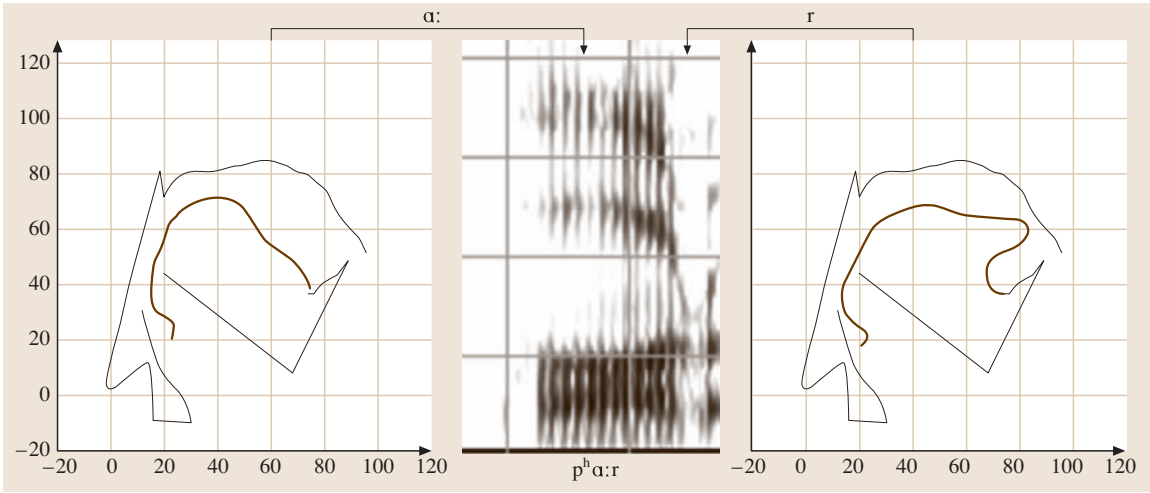

Fig. 16.30 Role of the tongue tip and blade in tuning F3. As the tongue is raised to produce the retroflex configuration required for the Swedish [r] sound an acoustically significant volume is created under the tongue blade. Simulations using APEX indicate that this cavity is responsible for the considerable lowering of the third formant associated with the [r]

Figure 16.29 plots results from APEX simulation experiments. It demonstrates how F1, F2 and F3 vary in response to the tongue moving continuously between palatal, velar and pharyngeal locations and maintaining a minimum constriction area A_{min} of 0.25 cm^2, a 7 mm jaw opening and keeping the larynx height constant. The dashed and solid lines refer to rounded and spread conditions. The calculations included a correction for the impedance of the vocal tract walls ([16.54], [16.28, p. 158]).

It is seen that the tongue movement has strong effect on F2. As the constriction becomes more posterior F2 decreases and F1 rises. In general, rounding lowers formants by varying degrees that depends on the formant-cavity relations that apply to the articulation in question. However, for the most palatal position in Fig. 16.29 – an articulation similar to an [i] or an [y], it has little effect on F2 whereas F3 is affected strongly.

The F2 versus F1 plot of the lower diagram of Fig. 16.29 was drawn to illustrate the effect of varying the jaw. The thin lines show how, at various degrees of jaw opening (6, 7, 8, 9, 12, 15, 19 and 23 mm), the tongue moves between palatal and pharyngeal constrictions by way of the neutral tongue shape. The bold lines pertain to fixed palatal, neutral or pharyngeal tongue contours. We note that increasing the jaw opening while keeping the tongue configuration constant shifts F1 upward.

Figure 16.30 exemplifies the role of the tongue tip and blade in tuning F3. The data come from an X-ray study with synchronized sound [16.79,85]. At the center a spectrogram of the syllable "*par*" [pʰɑ:r]. Perhaps the most striking feature of the formant pattern is the extensive lowering of F3 and F4 into the [r].

The articulatory correlates are illustrated in the tracings. As we compare the profiles for [ɑ:] and [r], we see that the major change is the raising of the tongue blade for [r] and the emergence of a significant cavity in front of, and below, the tongue blade. Simulations us-

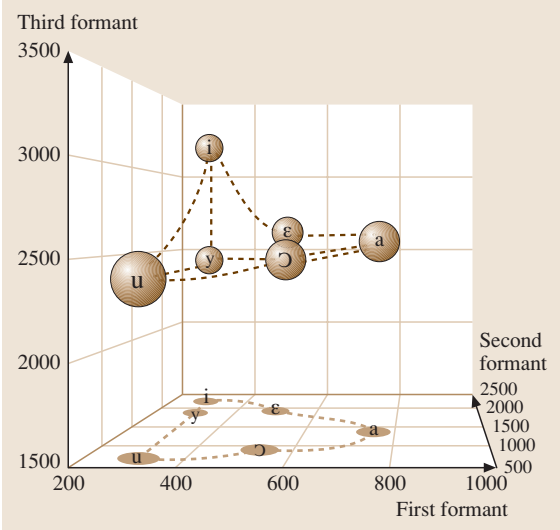

Fig. 16.31 The acoustic vowel space: *possible* vowel formant combinations according to the APEX model

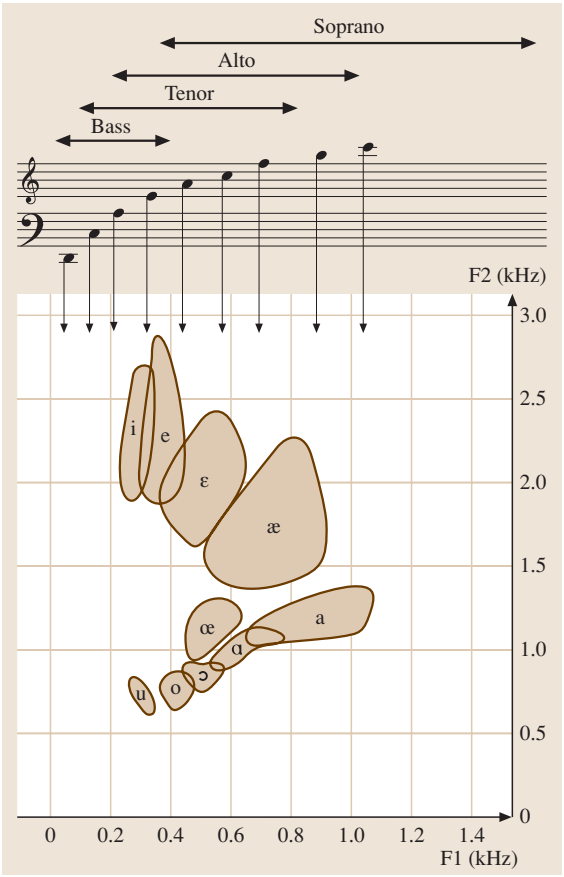

Fig. 16.32 Schematic illustration of the variation ranges of F1 and F2 for various vowels as pronounced by female and male adults. Above the graph F1 is given in musical score notation together with typical F0 ranges for the indicated voice classifications

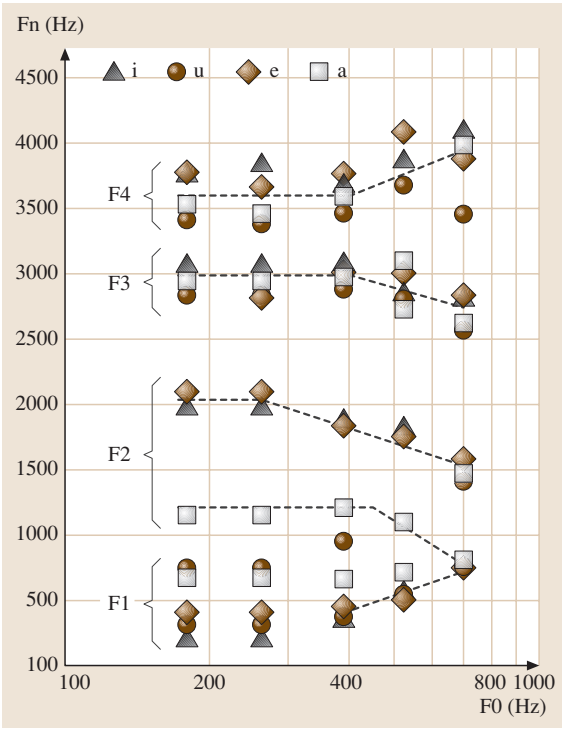

Fig. 16.33 Values of F1, F2, F3, and F4 for the indicated vowels in a professional soprano as function of F0. F1 and F3 are represented by open symbols and F2 and F4 by filled symbols. The values for F0 ≈180 Hz were obtained when the singer sustained the vowels in a speech mode. (After *Sundberg* [16.87])

ing APEX and other models [16.28,86] indicate that this subapical cavity is responsible for the drastic lowering of F3.

The curves of Fig. 16.30 are qualitatively consistent with the nomograms published for three-parameter area-function models, e.g., Fant [16.27]. An advantage of more-realistic physiological models is that the relationships between articulatory parameters and formant patterns become more transparent. We can summarize observations about APEX and the articulation-to-formants mapping as:

- F1 is controlled by the jaw in a direct way;
- F2 is significantly changed by front–back movement of the tongue;
- F3 is influenced by the action of the tongue blade.

Another way of summarizing the acoustic properties of a speech production model is to translate all of its articulatory capabilities into a formant space. By definition that contains all the formant patterns that the model is capable of producing (and specifying articulatorily). Suppose that a model uses n independent parameters and each parameter is quantized into a certain number of steps. By forming all *legal* combinations of those parametric values and deriving their vocal tract shapes and formant patterns, we obtain the data needed to represent that space graphically.

The APEX vowel space is shown in 3-D in Fig. 16.31. The x-axis is F1. The depth dimension is F2 and the vertical axis is F3. Smooth lines were drawn to enclose individual data points (omitted for clarity).

A cloud-like structure emerges. Its projection on the F2/F1 floor plane takes the form of the familiar trian-

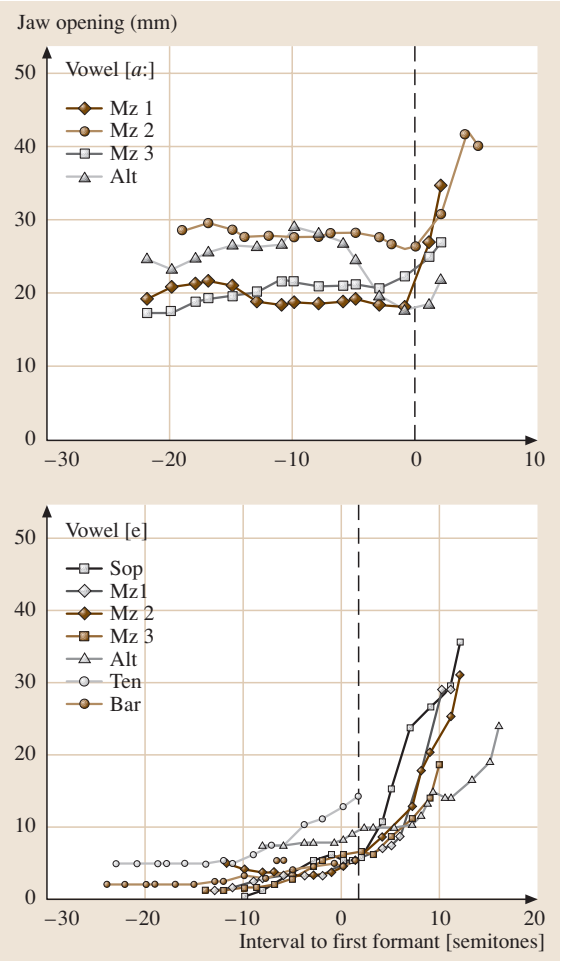

Fig. 16.34 Singers' jaw opening in the vowels [ɑ:] and [e] plotted as functions of the distance in semitones between F0 and the individual singer's normal F1 value for these vowels. The singers belonged to different classifications: Sop = soprano, Mz = mezzosoprano, Alt = alto, Ten = tenor, Bar = baritone. Symbols refer to subjects. (After *Sundberg* and *Skoog* [16.88])

gular pattern appears with [i], [a] and [u] at the corners. Adding F3 along the vertical axis offers additional room for vowel timber variations especially in the [i/y] region.

Figure 16.32 shows typical values for F1 and F2 for various vowels. At the top of the diagram, F1 is given also on the musical staff. The graph thus demonstrates that the fundamental frequency in singing is often higher than the normal value of F1. For example, the first formant of [i:] and [u:] is about 250 Hz (close to the pitch of C4), which certainly is a quite low note for a soprano.

Theory predicts that the vocal fold vibration will be greatly disturbed when F0 = F1 [16.89], and singers seem to avoid allowing F0 to pass F1 [16.87, 90]. Instead they raise F1 to a frequency slightly higher than F0. Some formant frequency values for a soprano singer are shown in Fig. 16.33. The results can be idealized in terms of lines, also shown in the figure, relating F1, F2, F3, and F4 to F0. Also shown are the subject's formant frequencies in speech. The main principle seems to be as follows. As long as fundamental frequency is lower than the normal value of the vowel's first formant frequency, this formant frequency is used. At higher pitches, the first formant is raised to a value somewhat higher than the fundamental frequency. In this way, the singer avoids having the fundamental frequency exceed the first formant frequency. With rising fundamental frequency, F2 of front vowels is lowered, while F2 of back vowels is raised to a frequency just above the second spectrum partial; F3 is lowered, and F4 is raised.

A commonly used articulatory trick for achieving the increase of F1 with F0 is a widening of the jaw opening [16.85]. The graphs of Fig. 16.34 show the jaw opening of professional singers as a function of the frequency separation in semitones between F0 and the F1 value that the singer used at low pitches. The graph referring to the vowel [ɑ:] shows that most singers began to widen their jaw opening when the F0 was about four semitones below the normal F1 value. The lower graph of Fig. 16.34 shows the corresponding data for the vowel [e]. For this vowel, most subjects started to widen the jaw opening when the fundamental was about four semitones above the normal value of F1. It is likely that below this pitch singers increase the first formant by other articulatory means than the jaw opening. A plausible candidate in front vowels is the degree of vocal tract constriction; a reduced constriction increases the first formant. Many singing teachers recommend their students to *release the jaw* or to *give space to the tone*; it seems likely that the acoustic target of these recommendations is to raise F1.

There are also other articulators that can be recruited for the purpose of raising F1. One is the lip opening. By retracting the mouth corners, the vocal tract is shortened; and hence the frequencies of the formants will increase. The vocal tract can also be shortened by raising the larynx, and some professional female singers take advantage of this tool when singing at high pitches.

This principle of tuning formant frequencies depending on F0 has been found in all singers who encounter the situation that the normal value of the first formant is lower than their highest pitches. In fact, all singers except basses encounter this situation at least for some

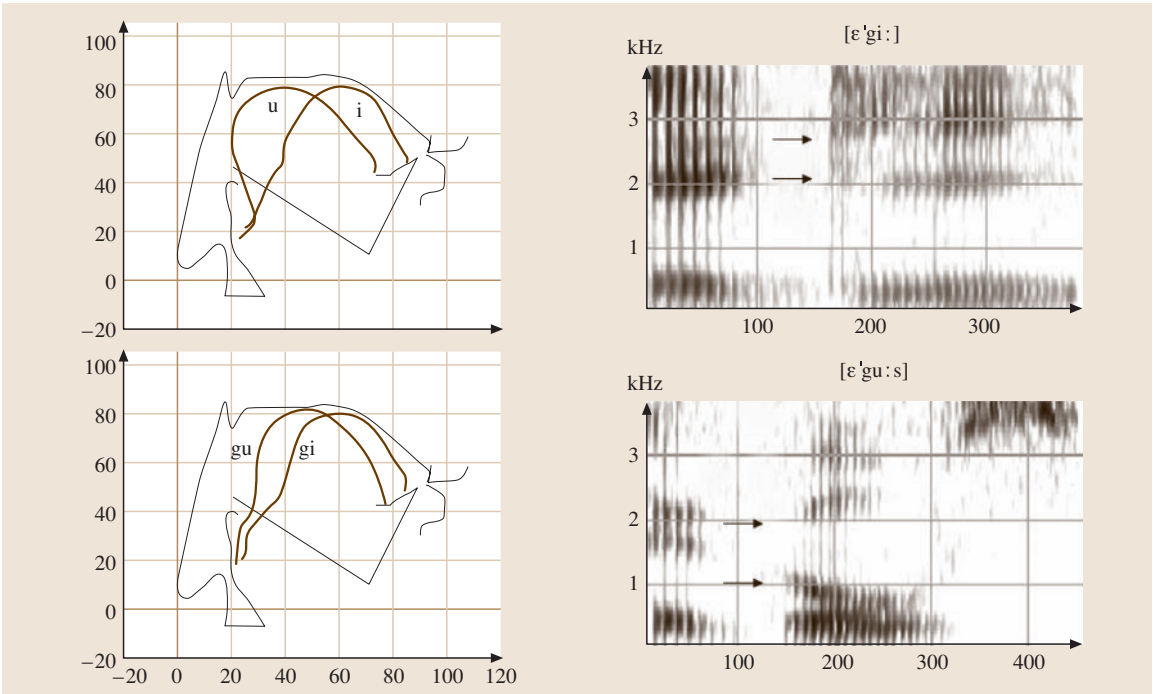

Fig. 16.35 The universal phenomenon of coarticulation illustrated with articulatory and acoustic data on Swedish [g]. The influence of the following vowel extends throughout the articulation of the stop closure. There is no point in time at which a *pure* (context-free) sample of the [g] could be obtained

vowels sung at high pitches. The benefit of these arrangements of the formant frequencies is an enormous increase of sound level, gained by sheer resonance.

Vowel quality is determined mainly by F1 and F2, as mentioned. Therefore, one would expect drastic consequences regarding vowel intelligibility in these cases. However, the vowel quality of sustained vowels seems to survive these pitch-dependent formant frequency changes surprisingly well, except when F0 exceeds the pitch of F5 (about 700 Hz). Above that frequency no formant frequency combination seems to help, and below it, the vowel quality would not be better if normal formant frequencies were chosen. The amount of text intelligibility which occurs at very high pitches relies almost exclusively on the consonants surrounding the vowel. Thus, facing the choice between inaudible tones with normal formant frequencies or audible tones with strange vowel quality, singers probably make a wise choice.

One of the major challenges both for applied and theoretical speech research is the great variability of the speech signal. Consider a given utterance as pronounced by speakers of the same dialect. A few moments' reflection will convince us that this utterance is certain to come in a large variety of physical shapes. Some variations reflect the speaker's age, gender, vocal anatomy as well as emotional and physiological state. Others are stylistic and situational as exemplified by speaking clearly in noisy environments, speaking formally in public, addressing large audiences, chatting with a close friend or talking to oneself while trying to solve a problem. Style and situation make significant contributions to the variety of acoustic waveforms that instantiate what we linguistically judge as the *same utterance*.

In phonetic experimentation investigators aim at keeping all of these stylistic and situational factors constant. This goal tends to limit the scope of the research to a style labeled *laboratory speech*, which consists of test items that are chosen by the experimenter and are typically read from a list. Despite the marked focus on this type of material, laboratory speech nonetheless presents several variability puzzles. We will mention two: segmental interaction, or *coarticulation*, and prosodic modulation. Both exemplify the ubiquitous context dependence of phonetic segments. First a few remarks on coarticulation.

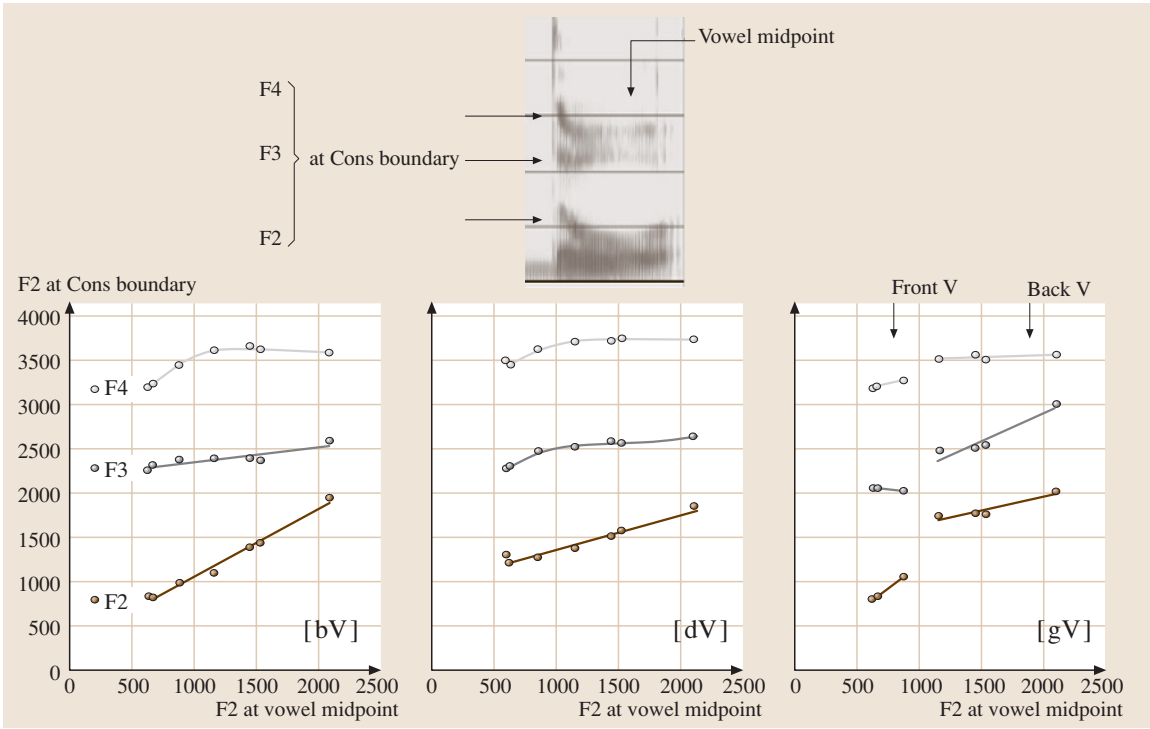

Fig. 16.36 Coarticulation in [bV], [dV] and [gV] as reflected by measurements of F2, F3 and F4 onsets plotted against an acoustic correlate of front back tongue movement, viz., F2 at the vowel midpoint

In Fig. 16.35 the phoneme [g] occurs in two words taken from an X-ray film [16.79, 85] of a Swedish speaker: [ɛˈgiː] and [ɛˈguːs]. In these words the first three phonemes correspond to the first three segments on the spectrogram: an initial [ɛ] segment, the [g] stop gap and then the final vowel. So far, so good. However, if we were to try to draw a vertical line on the spectrogram to mark the point in time where [ɛ] ends and [g] begins, or where [g] ends and the final vowel begins, we would soon realize that we have an impossible task. We would perhaps be able to detect formant movements in the [ɛ] segment indicating that articulatory activity towards the [g] had been initiated. However, that point in time occurs during the acoustic [ɛ] segment. Similarly, we could identify the endpoint of the formant transitions following [g] but that event occurs when the next segment, the final vowel, is already under way.

What we arrive at here is the classical conclusion that strings of phonemes are not organized as *beads on a necklace* [16.91–93]. The acoustic correlates of phonemes, the acoustic segments, are produced according to a motor schema that requires parallel activity in several articulatory channels and weaves the sequence of phonemes into a smooth fabric of overlapping movements. We are talking about *coarticulation*, the overlap of articulatory gestures in space and time.

Not only does this universal of motor organization give rise to the segmentation problem, i. e., make it impossible to chop up the time scale of the speech wave into phoneme-sized chunks, it creates another dilemma known as the *invariance issue*. We can exemplify it by referring to Fig. 16.35 again and the arrows indicating the frequencies of F2 and F3 at the moment of [g] release. With [iː] following they are high in frequency. Next to [uː] they are lower. What acoustic attributes define the [g] phoneme? How do we specify the [g] phoneme acoustically in a context-independent way?

The answer is that, because of coarticulation, we cannot provide an acoustic definition that is context independent. There is no such thing as an acoustic *pure sample* of a phoneme.

Articulatory observations confirm this account. There is no point in time where the shape of the tongue shows zero influence from the surrounding vowels. The tracings at the top left of Fig. 16.35 show the tongue con-

tours for [iː] and [uː] sampled at the vowel midpoints. The bottom left profiles show the tongue shapes at the [g] release. The effect of the following vowel is readily apparent.

So where do we look for the invariant phonetic correlates for [g]? Work on articulatory modeling [16.49, 94] indicates that, if there is such a thing as a single context-free target underlying the surface variants of [g], it is likely to occur at a deeper level of speech production located before the motor commands for [g] closure and for the surrounding vowels blend.

This picture of speech production raises a number of questions about speech perception that have been addressed by a large body of experimental work [16.95, 96] but which will not be reviewed here.

It should be remarked though that, the segmentation and invariance issues notwithstanding, the context sensitivity of phonetic segments is systematic. As an illustration of that point Fig. 16.36 is presented. It shows average data on formant transitions that come from the Swedish speaker of Fig. 16.36 and Figs. 16.26–16.28. The measurements are from repetitions of CV test words in which the consonants were [b], [d] or [g] and were combined with [iː] [ɛː] [æː] [a] [ɑː] [oː] and [uː]. Formant transition onsets for F2, F3 and F4 are plotted against F2 midpoints for the vowels.

If the consonants are coarticulated with the vowels following, we would expect consonant onset patterns to co-vary with the vowel formant patterns. As shown by Fig. 16.36 that is also what we find. Recall that, in the section on articulatory modeling, we demonstrated that F2 correlates strongly with the front–back movement of the tongue. This implies that, in an indirect way, the *x*-axis labeled 'F2 at vowel midpoint' can be said to range from back to front. The same reasoning applies to F2 onsets.

Figure 16.36 shows that the relationship between F2 onsets and F2 at vowel midpoint is linear for bV and dV. For gV, the data points break up into back (low F2) and front (high F2) groups. These straight lines – known as *locus equations* [16.97] – have received considerable attention since they provide a compact way of quantifying coarticulation. Data are available for several languages showing robustly that slopes and intercepts vary in systematic ways with places of articulation.

Furthermore, we see from Fig. 16.36 that lawful patterns are obtained also for F3 and F4 onsets. This makes sense if we assume that vocal tract cavities are not completely uncoupled and that hence, all formants – not only F2 – are to some extent influenced by where along the front–back dimension the vowel is articulated.

16.5 The Syllable

A central unit in both speech and singing is the syllable. It resembles the phoneme in that it is hard to define but it can be described in a number of ways.

Linguists characterize it in terms of how vowels and consonants pattern within it. The central portion, the nucleus, is normally a vowel. One or more consonants can precede and/or follow forming the onset and the coda respectively. The vowel/nucleus is always there; the onset and coda are optional.

Languages vary with respect to the way they combine consonants and vowels into syllables. Most of them favor a frame with only two slots: the CV syllable. Others allow more-elaborated syllable structures with up to three consonants initially and the mirror image in syllable final position. If there is also a length distinction in the vowel and/or consonant system, syllables frames can become quite complex. A rich pattern with consonant clusters and phonological length usually implies that the language has a strong contrast between stressed and unstressed syllables.

In languages that allow consonant sequences, there is a universal tendency for the segments to be serially ordered on an articulatory continuum with the consonants compatible with the vowel's greater jaw opening occurring next to the vowel, e.g., [l] and [r], while those less compatible, e.g. [s], are recruited at the syllable margins [16.98, 99]. In keeping with this observation, English and other languages use [spr] as an initial, but not final, cluster. The reverse sequence [rps] occurs in final, but not initial, position, cf. *sprawl* and *harps*. Traditionally and currently, this trend is explained in terms of an auditory attribute of speech sounds, sonority. The *sonority principle* [16.59] states that, as the most *sonorous* segments, vowels take the central nucleus position of the syllable and that the *sonority* of the surrounding consonants must decrease to the left and to the right starting from the vowel. Recalling that the degree of articulatory opening affects F1 which in turn affects sound intensity, we realize that these articulatory and auditory accounts are not incompatible. However,

the reason for the syllabic variations in sonority is articulatory: the tendency for syllables to alternate close and open articulations in a cyclical manner.

The syllable is also illuminated by a developmental perspective. An important milestone of normal speech acquisition is *canonical babbling*. This type of vocalization makes its appearance sometime between 6–10 months. It consists of sequences of CV-like events, e.g., [dædæ], [baba] [16.101–105]. The phonetic output of deaf infants differs from canonical babbling both quantitatively and quantitatively [16.106–108], suggesting that auditory input from the ambient language is a prerequisite for canonical babbling [16.109–112]. What babbling shares with adult speech is its "syllabic" organization, that is, the alternation of open and close articulations in which jaw movement is a major component [16.113].

As mentioned, the regular repetition of open-close vocal tract states gives rise to an amplitude modulation of the speech waveform. Vowels tend to show the highest amplitudes contrasting with the surrounding consonants which have various degrees of constriction and hence more reduced amplitudes. At the acoustic boundary between a consonant and a vowel, there is often an abrupt rise in the amplitude envelope of the waveform.

When a Fourier analysis is performed on the waveform envelope, a spectrum with primarily low, sub-audio frequency components is obtained. This is to be expected given the fact that amplitude envelopes vary slowly as a function of time. This representation is known as the modulation spectrum [16.114]. It reflects recurring events such as the amplitude changes at consonant-vowel boundaries. It provides an approximate record of the rhythmic pulsating stream of stressed and unstressed syllables.

The time envelope may at first appear to be a rather crude attribute of the signal. However, its perceptual importance should not be underestimated. Room acoustics and noise distort speech by modifying and destroying its modulation spectrum. The modulation transfer function was proposed by Houtgast and Steeneken as a measure of the effect of the auditorium on the speech signal and as a basis for an index, the speech transmission index (STI), used to predict speech intelligibility under different types of reverberation and noise. The success of this approach tells us that the modulation spectrum, and hence the waveform envelope, contains information that is crucial for robust speech perception [16.115]. Experimental manipulation of the temporal envelope has been performed by Drullman et al. [16.116, 117] whose work reinforces the conclusions reached by Houtgast and Steeneken.

There seems to be something special about the front ends of syllables. First, languages prefer CVs to VCs. Second, what children begin with is strings of CV-like pseudosyllables that emulate the syllable onsets of adult speech. Third there is perceptually significant information for the listener in the initial dynamics of the syllable. Let us add another phenomenon to this list: the *syllable beat*, or the syllable's P-center [16.118, 119].

In reading poetry, or in singing, we have a very strong sense that the syllables are spoken/sung in accordance with the rhythmic pattern of the meter. Native speakers agree more or less on how many syllables there are in a word or a phrase. In making such judgments, they seem to experience syllables as unitary events. Although it may take several hundred milliseconds to pronounce, subjectively a syllable appears to occur at a specific moment in time. It is this impression to which the phonetic term *syllable beat* refers and that has been studied experimentally in a sizable number of publications [16.120–126].

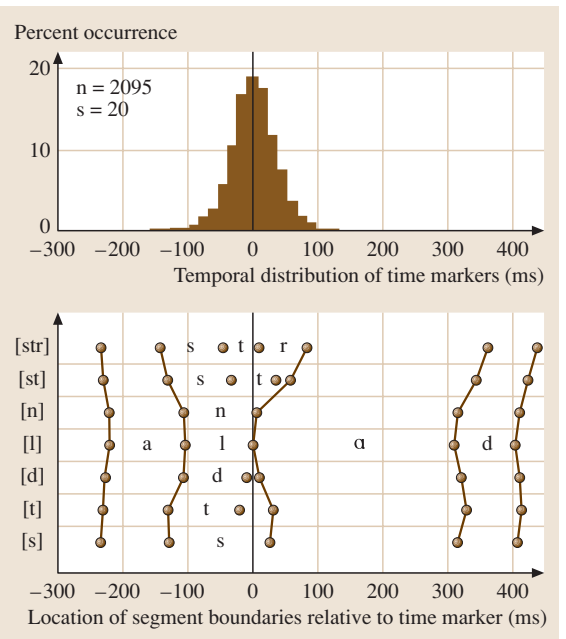

Fig. 16.37 In *Rapp* [16.100] three male Swedish subjects were asked to produce test words based on an [a_ˈɑːd] frame with [s, t, d, l, n, st, str] as intervocalic segment(s). These items were pronounced in synchrony with metronome beats presented over headphones. The figure shows the temporal distribution of the metronome beats (*upper panel*) in relation to the average time location of acoustic segment boundaries (*lower panel*)

Rapp [16.100] asked three native speakers of Swedish to produce random sets of test words built from [aC^{l}ɑːd] were the consonant C was selected from [s, t, d, l, n, st, str]. The instruction was to synchronize the stressed syllable with a metronome beat presented over earphones.

The results are summarized in Fig. 16.37. The *x*-axis represents distance in milliseconds from the point of reference, the metronome beat. The top diagram shows the total distribution of about 2000 time markers around the mean. The lower graph indicates the relative location of the major acoustic segment boundaries.

Several phonetic correlates have been proposed for the syllable beat: some acoustic/auditory, others articulatory. They all hover around the vowel onset, e.g., the amplitude envelope of a signal [16.127], rapid increase in energy in spectral bands [16.128, 129] or the onset of articulatory movement towards the vowel [16.130].

Rapp's data in Fig. 16.37 indicate that the mean beat time tends to fall close to the release or articulatory opening in [t, d, l, n] but that it significantly precedes the acoustic vowel onsets of [str-], [st-] and [s-]. However, when segment boundaries were arranged in relation to a fixed landmark on the F0 contour and vowel onsets were measured relative to that landmark, the range of vowel onsets was reduced. It is possible that the syllable beat may have its origin, not at the acoustic surface, nor at some kinematic level, but in a deeper motor control process that coordinates and imposes coherence on respiratory, phonatory and articulatory activity needed to produce a syllable.

Whatever the definitive explanation for the syllable's psychological *moment of occurrence* will be, the syllable beat provides a useful point of entry for attempts to understand how the control of rhythm and pitch works in speech and singing. Figure 16.38 compares spectrograms of the first few bars of *Over the rainbow*, spoken (left) and sung (right).

Vertical lines have been drawn at vowel onsets and points where articulators begin to move towards a more open configuration. The lines form an isochronous temporal pattern in the sung version which was performed at a regular rhythm. In the spoken example, they occur at intervals that seem more determined by the syllable's degree of prominence.

The subject reaches F0 targets at points near the beats (the vertical lines). From there target frequencies are maintained at stationary values until shortly before it is time to go to the next pitch. Thus the F0 curve resembles a step function with some smoothing applied to the steps.

On the other hand, the F0 contour for speech shows no such steady states. It makes few dramatic moves as it

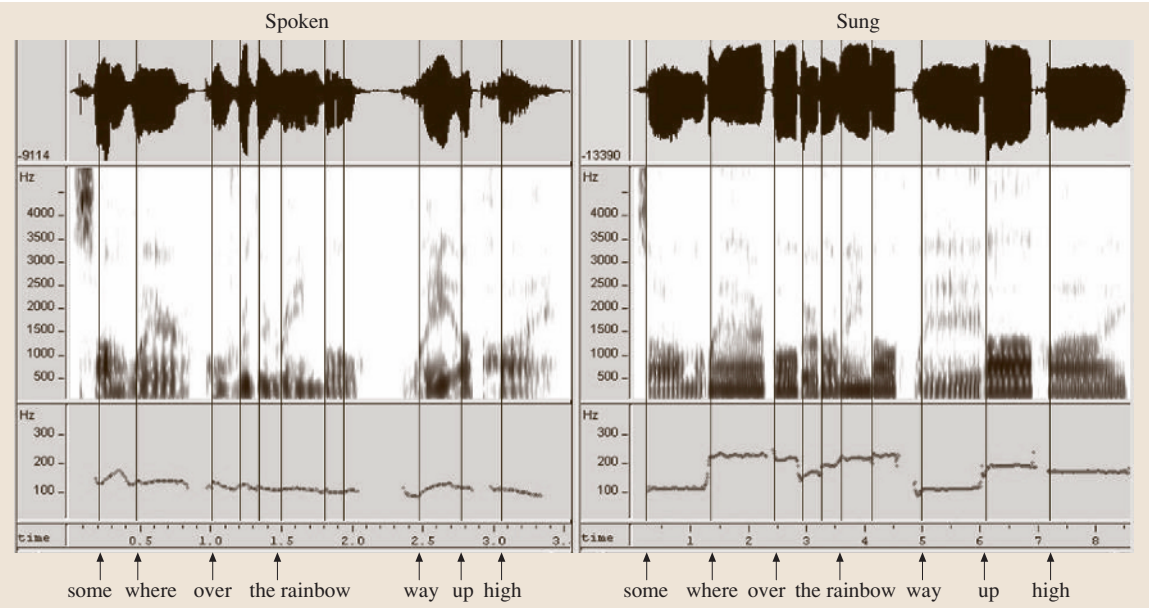

Fig. 16.38 Waveforms and spectrograms of the first few bars of Over the rainbow, spoken (*left*) and sung (*right*). Below: F0 traces in Hz. Vertical lines were drawn at time points corresponding to "vowel onsets" defined in terms of onset of voicing after a voiceless consonant, or the abrupt vocal tract opening following a consonant

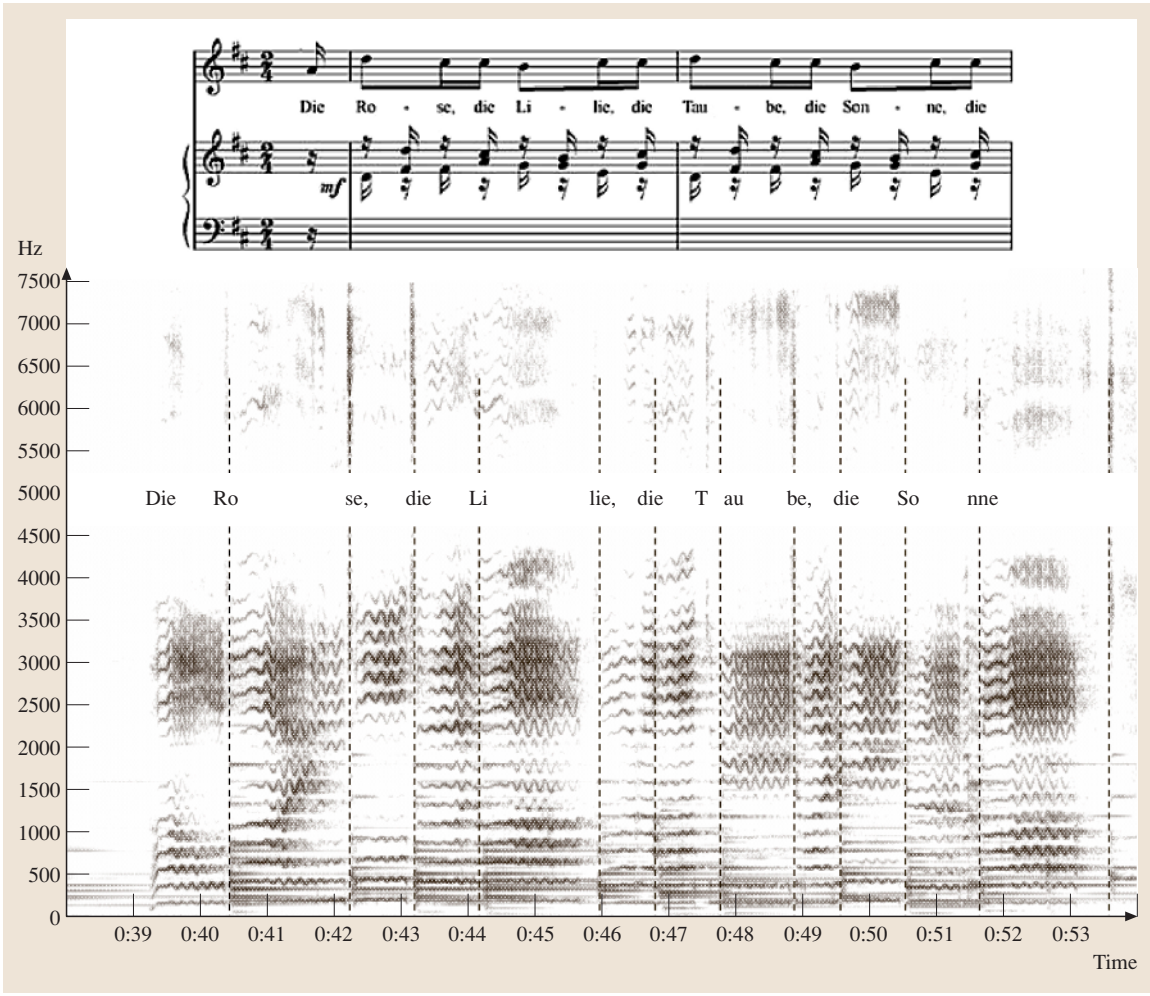

Fig. 16.39 Spectrogram of Dietrich Fischer-Dieskau's recording of song number 4 from Robert Schumann's Dichterliebe, op. 48. The *vertical bars* mark the onset of the tones of the piano accompaniment. Note the almost perfect synchrony between vowel onset and the piano. Note also the simultaneous appearance of high and low partials after the consonants also after the unvoiced /t/ in Taube

gradually drifts downward in frequency (the declination effect).

Figure 16.39 shows a typical example of classical singing, a spectrogram of a commercial recording of Dietrich Fischer-Dieskau's rendering of the song "*Die Rose, die Lilie*" from Robert Schumann's *Dichterliebe*, op. 48. The vertical dashed lines show the onsets of the piano accompaniment. The wavy patterns, often occurring somewhat after the vowel onset, reflect the vibrato. Apart from the vibrato undulation of the partials the gaps in the pattern of harmonics are quite apparent. At these points we see the effect of the more constricted articulations for the consonants. Note the rapid and crisply synchronous amplitude rise in all partials at the end of the consonant segments, also after unvoiced consonants, e.g. the /t/ in *Taube*. These rises are synchronized with the onset of the piano accompaniment tones, thus demonstrating that in singing a beat is marked by the vowel. The simultaneous appearing of low and high partials seems to belong to the characteristics of classical singing as opposed to speech, where the higher partials often arrive with a slight delay after unvoiced consonants. As mentioned before, such consonants are produced with abduction of the vocal folds. To generate

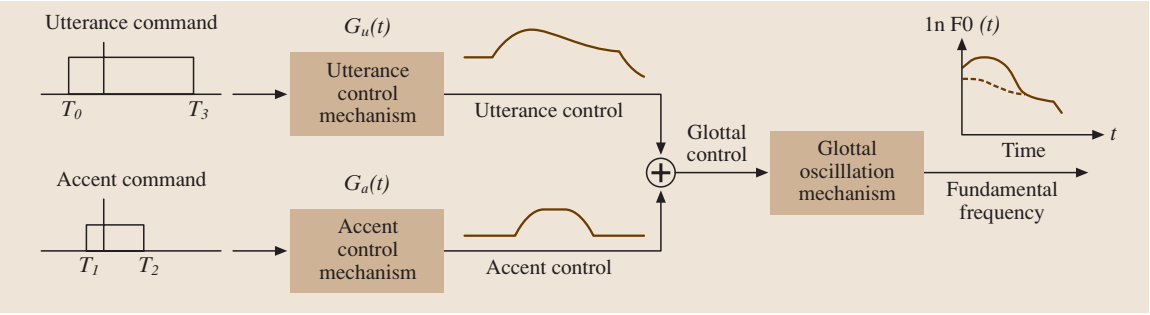

Fig. 16.40 A quantitative approach to synthesizing F0 contours by rule (After *Fujisaki* [16.131]). The framework shown here is an example of the so-called *superposition models* which generate F0 patterns by adding strongly damped responses to input step function target commands for phrase and accents

also high partials in this context, the vocal folds need to close the glottis at the very first vibratory cycle at the vowel onset. A potential benefit of this might be that this enhances the higher formants which are important to text intelligibility.

Several quantitative frameworks have been proposed for generating speech F0 contours by rule (for overview see Frid [16.132]). A subgroup of these has been called *superposition models* [16.131, 133]. A characteristic of such models is that they decompose the F0 curve into separate phrase and accent components. The accent commands are F0 step or impulse functions temporally linked to the syllables carrying stress. Accent pulses are superimposed on the phrase component. For a declarative sentence, the phrase component is a falling F0 contour produced from rectangular step function *hat patterns* [16.134]. The phrase and accent commands are passed through critically damped filters to convert their sum into the smoothly varying output F0 contour. An example of this approach is shown in Fig. 16.40 [16.131].

The last few figures suggest that the F0 patterns of singing and speech may be rather different if compared in terms of the raw F0 curves. However, the superposition models suggest that speech F0 contours have characteristics in part attributable to the passive response characteristics of the neuro-mechanical system that produces them, and in part due to active control signals. These control commands take the form of stepwise changes, some of short, others of longer duration. This representation is not unlike the sequence of F0 targets of a melody.

The implication of this analysis is that F0 is controlled in similar ways in speech and singing in the sense that both are based on sequences of underlying steady-state targets. On the other hand, a significant difference is that in singing high accuracy in attainment of acoustic target frequencies is required whereas in speech such demands are relaxed and smoothing is stronger.

16.6 Rhythm and Timing

A striking characteristic of a foreign language is its rhythm. Phoneticians distinguish between stress-timed and syllable-timed languages. English, Russian, Arabic and Thai are placed in the first group [16.135]. French, Spanish, Greek, Italian, Yoruba and Telugu are examples of the second.

Stress timing means that stressed syllables recur at approximately equal intervals. For instance, it is possible to say "*in ENGlish STRESses reCUR at EQual INtervals*", spacing the stressed syllables evenly in time and without sounding too unnatural. Stress increases syllable duration. In case several unstressed syllables occur in between stresses they are expected to undergo compensatory shortening. Syllable, or *machine-gun*, timing [16.136] implies that syllables recur at approximately equal intervals. In the first case it is the stresses that are isochronous, in the second the syllables.

To what extent are speech rhythms explicitly controlled by the speaker? To what extent are they fortuitous by-products of other factors? Does acquiring a new language involve learning new rhythmic target patterns?

The answer depends on how speech is defined. In the reading of poetry and nursery rhymes it is

clear that an external target, viz. the metric pattern, is a key determinant of how syllables are produced. Similarly, the groupings and the pausing in narrative prose, as read by a professional actor, can exhibit extrinsic rhythmic stylization compared with unscripted speech [16.137].

Dauer [16.138] points the way towards addressing such issues presenting statistics on the most frequently occurring syllable structures in English, Spanish and French. In Spanish and French, syllables were found to be predominantly open, i. e. ending with a vowel, whereas English showed a preference for closed syllables, i. e. ending with a consonant, especially in stressed syllables. Duration measurements made for English and Thai (stress-timed) and Greek, Spanish and Italian (syllable-timed) indicated that the duration of the interstress intervals grew at the same constant rate as a function of the number of syllables between the interstress intervals. In other words, no durational evidence was found to support the distinction between stress timing and syllable timing. How do we square that finding with the widely shared impression that some languages do indeed sound *stress-timed* and others *syllable-timed*?

Dauer [16.138, p. 55] concludes her study by stating that: " . . . *the rhythmic differences we feel to exist between languages such as English and Spanish are more a result of phonological, phonetic, lexical and syntactic facts about that language than any attempt on the part of the speaker to equalize interstress or intersyllable intervals*."

It is clear that speakers are certainly capable of imposing a rhythmic template in the serial read-out of syllables. But do they put such templates in play also when they speak spontaneously? According to *Dauer* and others [16.139, 140] rhythm is not normally an explicitly controlled variable. It is better seen as an emergent product of interaction among the choices languages make (and do not make) in building their syllables: e.g., from open versus closed syllable structures, heavy versus weak clusters, length distinctions and strong stressed/unstressed contrast. We thus learn that the distinction between syllable timing and stress timing may be a useful descriptive term but should primarily be applied to the phonetic output, more seldom to its input control.

Do speech rhythms carry over into music? In other words, would music composed by speakers of syllable-timed or stress-timed languages also be syllable timed or stress timed? The question has been addressed [16.141, 142] and some preliminary evidence has been reported.

There is more to speech timing than what happens inside the syllable. Processes at the word and phrase levels also influence the time intervals between syllables. Consider the English words *digest*, *insult* and *pervert*. As verbs their stress pattern can be described as a sequence of *weak–strong* syllables. As nouns the order is reversed: *strong–weak*. The syllables are the same but changing the word's stress contour (the lexical stress) affects their timing.

The *word length effect* has been reported for several languages [16.58, 143]. It refers to the tendency for the stressed vowel of the word to shorten as more syllables are appended, cf. English *speed*, *speedy*, *speedily*. In Lehiste's formulation: "It appears that in some languages the word as a whole has a certain duration that tends to remain relatively constant, and if the word contains a greater number of segmental sounds, the duration of the segmental sounds decreases as their number in the word increases."

At the phrase level we find that rhythmic patterns can be used to signal differences in syntactic structure. Compare:

1. *The 2000-year-old skeletons*,
2. *The two 1000-year-old skeletons*.

The phrases contain syllables with identical phonetic content but are clearly timed differently. The difference is linked to that fact that in (1) *2000-year-old* forms a single unit, whereas in (2) *two 1000-year-old* consists of two constituents.

A further example is:

3. *Johan greeted the girl with the flowers*.

Hearing this utterance in a specific context a listener might not find it ambiguous. But it has two interpretations: (a) either Johan greeted the girl who was carrying flowers, or (b) he used the flowers to greet her. If spoken clearly to disambiguate, the speaker is likely to provide temporal cues in the form of shortening the syllables within each syntactic group to signal the coherence of the constituent syllables (cf. the word length effect above) and to introduce a short pause between the two groups. There would be a lengthening of the last syllable before the pause and of the utterance-final syllable.

Version (a) can be represented as

4. [*Johan greeted*] # [*the girl with the flowers*].

In (b) the grouping is

5. [*Johan greeted the girl*] # [*with the flowers*].

The # symbol indicates the possibility of a short juncture or pause and a lengthening of the segments preceding the pause. This boundary cue is known as *pre-pausal lengthening*, or more generally as *final lengthening* [16.143, 144], a process that has been observed for a great many languages. It is not known whether this process is a language universal. It is fair to say that is typologically widespread.

Curiously it resembles a phenomenon found in poetry, folk melodies and nursery tunes, called *catalexis*. It consists in omitting the last syllable(s) in a line or other metrical unit of poetry. Instead of four mechanically repeated trochaic feet as in:

|– ⏑ |– ⏑ |– ⏑ |– ⏑ |
|– ⏑ |– ⏑ |– ⏑ |– ⏑ |

we typically find catalectic lines with durationally implied but deleted final syllables as in:

|– ⏑ |– ⏑ |– ⏑ |– |

Old McDonald had a farm

|– ⏑ |– ⏑ |– | |

ee-i ee-i oh

Final lengthening is an essential feature of speech prosody. Speech synthesized without it sounds both highly unnatural and is harder to perceive. In music performance frameworks, instrumental as well as sung, final lengthening serves the purpose of grouping and constituent marking [16.145, 146].

16.7 Prosody and Speech Dynamics

Degree of prominence is an important determinant of segment and syllable duration in English. Figure 16.41 shows spectrograms of the word '*squeal*' spoken with four degrees of stress in sentences read in response to a list of questions (source: [16.147]). The idea behind this method was to elicit tokens having emphatic, strong (as in focus position), neutral and weak (unfocused) stress on the test syllable. The lower row compares the strong and the emphatic pronunciations (left and right respectively). The top row presents syllables with weaker degrees of stress.

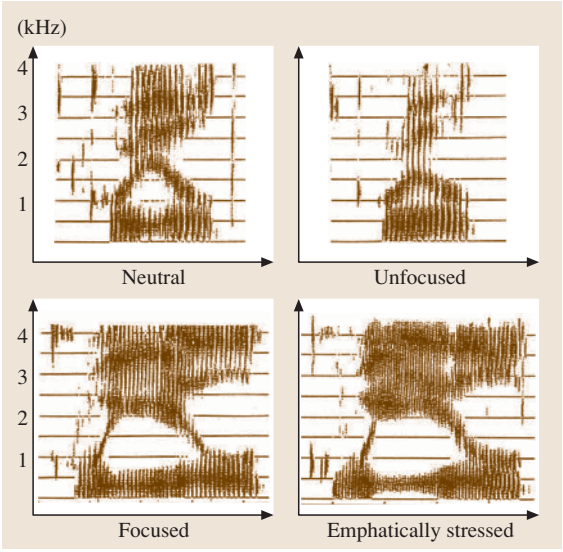

Fig. 16.41 Spectrograms of the word 'squeal' spoken with four degrees of stress in sentences read in response to a list of questions. (After *Brownlee* [16.147])

The representations demonstrate that the differences in stress have a marked effect on the word's spectrographic pattern. Greater prominence makes it longer. Formants, notably F2, show more extensive and more rapid transitions.

A similar experimental protocol was used in two studies of the dynamics of vowel production [16.147, 148]. Both report formant data on the English front vowels [i], [ɪ], [ɛ] and [eɪ] occurring in syllables selected to maximize and minimize contextual assimilation to the place of articulation of the surrounding consonants. To obtain a maximally assimilatory frame, words containing the sequence [w_ɬ] were chosen, e.g., as in *wheel*, *will*, *well* and *wail*. (The [w] is a labio-velar and English [ɬ] is velarized. Both are thus made with the tongue in a retracted position). The minimally assimilatory syllable was [h_d].

Moon's primary concern was *speaking style* (clear versus casual speech). Brownlee investigated changes due to *stress variations*. In both studies measurements were made of vowel duration and extent of formant transitions. Vowel segment boundaries were defined in terms of fixed transition onset and endpoint values for [w] and [ɬ] respectively. The [h_d] frame served as a *null context* reference. The [w_ɬ] environment produced formant transitions large enough to provide a sensitive and robust index of articulatory movement (basically the front–back movement of the tongue).

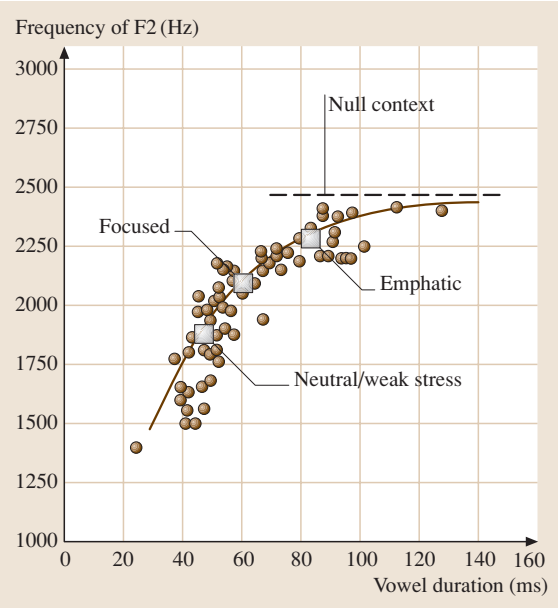

Fig. 16.42 Measurements of F2 as function of vowel duration from tokens of [i] in the word 'squeal' spoken by a male subject under four conditions of stress: emphatically stressed, focused, neutral stress and weak stress (out of focus). *Filled circles* pertain to the individual measurements from all the four stress conditions. The *dashed horizontal line*, is the average F2 value of citation form productions of [h_d]. The averages for the four stress conditions are indicated by the *open squares*. As duration increases – which is approximately correlated with increasing stress – the *square symbols* are seen to move up along the smooth curve implying decreasing undershoot. (After *Brownlee* [16.147])

Figure 16.42 presents measurements of F2 and vowel duration from tokens of [i] in the word '*squeal*' spoken by a male subject under four conditions of stress: emphatically stressed, focused, neutral stress and weak stress (out of focus). The subject was helped to produce the appropriate stress in the right place by reading a question before each test sentence.

Filled circles pertain to individual measurements pooled for all stress conditions. The points form a coherent cluster which is fairly well captured by an exponential curve. The asymptote is the dashed horizontal line, i. e., the average F2 value of the [h_d] data for [i]. The averages for the stress conditions are indicated by the unfilled squares. They are seen to move up along the curve with increasing stress (equivalently, vowel duration).

To understand the articulatory processes underlying the dynamics of vowel production it is useful to adopt a biomechanical perspective. In the [w_ɬ] test words the tongue body starts from a posterior position in [w] and moves towards a vowel target located in the front region of the vocal tract, e.g., [i], [ɪ], [ɛ] or [e]. Then it returns to a back configuration for the dark [ɬ]. At short vowel durations there is not enough time for the vowel gesture to be completed. As a result, the F2 movement misses the reference target by several hundred Hz. Note that in unstressed tokens the approach to target is off by almost an octave. As the syllable is spoken with more stress, and the vowel accordingly gets longer, the F2 transition falls short to a lesser degree. What is illustrated here is the phenomenon known as *formant undershoot* [16.149, 150].

Formant undershoot has received a fair amount of attention in the literature [16.151–155]. It is generally seen as an expression of the sluggish response characteristics of the speech production system. The term *sluggish* here describes both neural delays and mechanical time constants. The response to the neural commands for a given movement is not instantaneous. It takes time for neural impulses to be transformed into muscular contractions and it takes time for the tongue, the jaw and other articulatory structures to respond mechanically to the forces generated by those contractions. In other words, several stages of filtering take place between control signals and the unfolding of the movements. It is this filtering that makes the articulators sluggish. When commands for successive phonemes arrive faster than the articulatory systems are able to respond, the output is a set of incomplete movements. There is undershoot and failure to reach spatial, and thus also acoustic, targets.

However, biomechanics tells us that, in principle, it should be possible to compensate for system characteristics by applying greater force to the articulators thereby speeding up movements and improving target attainment [16.156]. Such considerations make us expect that, in speech movement dynamics, a given trajectory is shaped by

1. the distance between successive articulatory goals (the extent of movement);
2. articulatory effort (input force); and
3. duration (the time available to execute the movement).

The data of Brownlee are compatible with such a mechanism in that the stress-dependent undershoot observed is likely to be a joint consequence of: (1) the large distances between back and front targets; (2) the stress differences corresponding to variations in articulatory effort; and (3) durational limitations.

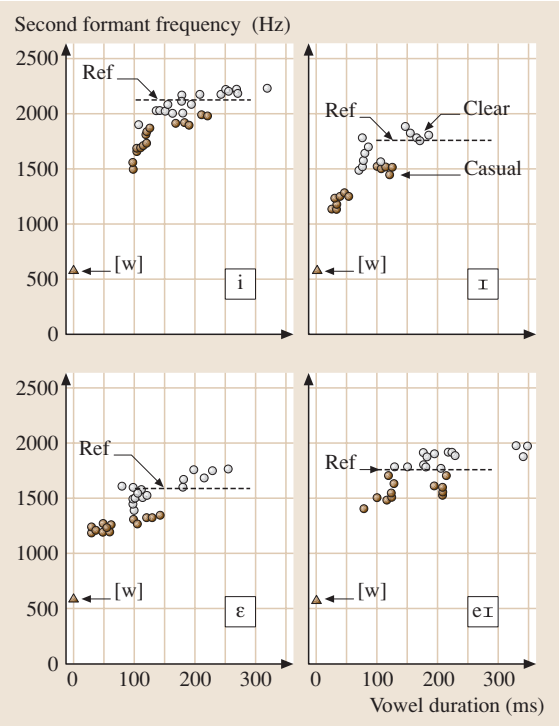

Fig. 16.43 Measurements F2 as function of vowel duration for five male subjects' casual citation-form style pronunciation of the words *wheel*, *wheeling*, *Wheelingham*, *will*, *willing*, *willingly* and for their pronunciation of the same words when asked to say them clearly in an overarticulated way. Casual style is represented by the *solid circles*, clear speech by *open circles*. The frequency of F2 in [w] is indicated by the arrows at 600 Hz. The mean F2 for the [h_d] data is entered as dashed horizontal lines. (After [16.157], Fig. 4)

The three factors – movement extent, effort and duration – are clearly demonstrated in Moon's research on clear and casual speech. As mentioned, Moon's test syllables were the same as Brownlee's. He varied speaking style while keeping stress constant. This was achieved by taking advantage of the word-length effect.The segment strings [wiɬ-], [wɪɬ-], [wɛɬ-] and [weɬ-] were used as the first main-stressed syllable in mono-, bi- and trisyllabic words to produce items such as *wheel*, *wheeling*, *Wheelingham, will*, *willing*, *willingly*, etc.

In the first part of his experiment, five male subjects were asked to produce these words in casual citation-form style. In the second section the instruction was to say them clearly in an *overarticulated* way. For the citation-form pronunciation the results replicated previously observed facts in that: (a) all subjects exhibited duration-dependent formant undershoot, and (b) the magnitude of this effect varied in proportion to the extent of the [w]–[vowel] transition. In the clear style, undershoot effects were reduced. The mechanism by which this was achieved varied somewhat from subject to subject. It involved combinations of increased vowel duration and more rapid and more extensive formant transitions. Figure 16.43 shows the response of one of the subjects to the two tasks [16.157].

Casual style is represented by the solid circles, clear speech by open circles. The frequency of F2 in [w] is indicated by the arrows at 600 Hz. The mean F2 for the [h_d] data is entered as dashed horizontal lines.

The solid points show duration-dependent undershoot effects in relation to the reference values for the [h_d] environment. The open circles, on the other hand, overshoot those values suggesting that this talker used more extreme F2 targets for the clear condition. The center of gravity of the open circles is shifted to the right for all four vowels showing that clear forms were longer than citation forms. Accordingly, this is a subject who used all three methods to decrease undershoot: clear pronunciations exhibited consistently more extreme targets (F2 higher), longer durations and more rapid formant movements.

These results demonstrate that duration and context are not the only determinants of vowel reduction since, for any given duration, the talker is free to vary the degree of undershoot by choosing to articulate more forcefully (as in overarticulated *hyperspeech*) or in a more relaxed manner (as in casual hypospeech). Taken together, the studies by Moon and Brownlee suggest that the dynamics of articulatory movement, and thus of formant transitions, are significantly constrained by three factors: extent of movement, articulatory effort and duration.

16.8 Control of Sound in Speech and Singing

When we perform an action – walk, run, or reach for and manipulate objects – our motor systems are faced with the fact that the contexts under which movements are made are never exactly the same. They change significantly from one situation to the next. Nonetheless, motor systems adapt effortlessly to the continually changing

conditions presumably because, during evolution, they were shaped by the need to cope with unforeseen events and obstacles [16.158]. Their default mode of operation is compensatory [16.159].

Handwriting provides a good illustration of this ability. A familiar fact is that it does not matter if something is written on the blackboard or on a piece of paper. The characteristic features of someone's handwriting are nevertheless easily recognized. Different sets of muscles are recruited and the size of the letters is different but their shapes remain basically similar. What this observation tells us is that movements are not specified in terms of fixed set of muscles and constant contraction patterns. They are recruited in functionally defined groups, *coordinative structures* [16.160]. They are planned and executed in an external coordinate space, in other words in relation to the 3-D world in which they occur so as to attain goals defined outside the motor system itself. The literature on motor mechanisms teaches us that voluntary movements are prospectively organized or future-oriented.

Speech and singing provide numerous examples of this output-oriented mode of motor control [16.161–163]. Earlier in the chapter we pointed out that, in upright position, the diaphragm and adjacent structures are influenced by gravity and tend to be pulled down, thereby causing the volume of the thoracic cavity to increase. In this position, gravity contributes to the inspiratory forces. By contrast, in the supine position, the diaphragm tends to get pushed up into the rib cage, which promotes expiration [16.1, p. 40].

Sundberg et al. [16.164] investigated the effect of upright and supine positions in two baritone singers using synchronous records of esophageal and gastric pressure, EMG from inspiratory and expiratory muscles, lung volume and sound. Reorganization of respiratory activity was found and was interpreted as compensation for the different mechanical situations arising from the upright and supine conditions.

This finding is closely related to what we know about breathing during speech. As mentioned above (Fig. 16.6), the Ps tends to stay fairly constant for any given vocal effort. It is known that this result is achieved by tuning the balance between inspiratory and expiratory muscles. When the lungs are expanded so that the effect of elastic recoil creates a significant expiration force, inspiratory muscles predominate to put a brake on that force. For reduced lung volumes the situation is the opposite. The effect of elastic recoil is rather to increase lung volume. Accordingly, the muscle recruitment needed to maintain Ps is expected to be primarily expiratory. That is indeed what the data show [16.8, 66, 165].

The bite-block paradigm offers another speech example. In one set of experiments [16.166] subjects were instructed to pronounce syllables consisting only of a long vowel under two conditions: first normally, then with a bite block (BB) between the teeth. The subjects all non-phoneticians were told to try to sound as normal as possible despite the BB. No practice was allowed. The purpose of the BB was to create an abnormally large jaw opening for close vowels such as [i] and [u]. It was argued that, if no tongue compensation occurred, this large opening would drastically change the area function of the close vowels and disrupt their formant patterns. In other words, the question investigated was whether the subjects were able to sound normal despite the BB.

Acoustic recordings were made and formant pattern data were collected for comparisons between conditions, vowels and subjects. The analyses demonstrated clearly that subjects were indeed able to produce normal sounding vowels in spite of the BB. At the moment of the first glottal pulse formant patters were well within the normal ranges of variation.

In a follow-up X-ray investigation [16.167] it was found that compensatory productions of [i] and [u] were made with super-palatal and super-velar tongue shapes. In other words, tongue bodies were raised high above normal positions so as to approximate the normal cross-sectional area functions for the test vowels.

Speech and singing share a lot of features but significant differences are brought to light when we consider what the goals of the two behaviors are.

It is more than 50 years since the sound spectrograph became commercially available [16.91]. In this time we have learned from visible speech displays and other records that the relation between phonetic units and the acoustic signal is extremely complex [16.168]. The lesson taught is that invariant correlates of linguistic categories are not readily apparent in the speech wave [16.169]. In the preceding discussion we touched on some of the sources of this variability: coarticulation, reduction and elaboration, prosodic modulation, stylistic, situational and speaker-specific characteristics. From this vantage point it is astonishing to note that, even under noisy conditions, speech communication is a reliable and robust process. How is this remarkable fact to be accounted for?

In response to this problem a number of ideas and explanatory frameworks have been proposed. For instance, it has been suggested that acoustic invariants are rela-

tional rather than absolute (à la tone intervals defined as frequency ratios).

Speech is rather a set of movements made audible than a set of sounds produced by movements [16.14, p. 33].

In keeping with this classical statement, many investigators have argued that speech entities are to be found at some upstream speech production level and should be defined as *gestures* [16.170–173].

At the opposite end, we find researchers (e.g., Perkell [16.163]) who endorse Roman Jakobson's view which, in the search for units gives primacy to the perceptual representation of speech sounds. To Jakobson the stages of the speech chain form an " . . . *operational hierarchy of levels of decreasing pertinence: perceptual, aural, acoustical and articulatory (the latter carrying no direct information to the receiver)*." [16.174]

These viewpoints appear to conflict and do indeed divide the field into those who see speech as a motoric code (the *gesturalist* camp) and those who maintain that it is primarily shaped by perceptual processes (the *listener-oriented* school).

There is a great deal of experimental evidence for both sides, suggesting that this dilemma is not an either–or issue but that both parties offer valuable complementary perspectives.

A different approach is taken by the H & H (Hyper and Hypo) theory [16.175, 176]. This account is developed from the following key observations about speaker–listener interaction:

1. Speech perception is always a product of signal information and listener knowledge;
2. Speech production is adaptively organized.

Here is an experiment that illustrates the first claim about perceptual processing. Two groups of subjects listen to a sequence of two phrases: a question followed by an answer. The subject groups hear different questions but a single physically identical reply. The subjects' task is to say how many words the reply contains.

The point made here is that Group 1 subjects hear [lɛsn̩ faɪv] as "*less than five*". Those in Group 2 interpret it as "*lesson five*". The first group's response is "*three words*", and the answer of the second is "*two words*". This is *despite the fact* that physically the [lɛsn̩ faɪv] stimulus is exactly the same. The syllabic [n̩] signals the word *than* in one case and the syllable *–on* in the other. Looking for the invariant correlates of the initial consonant *than* is doomed to failure because of the severe degree of reduction.

To proponents of H & H theory this is not an isolated case. This is the way that perception in general works. The speech percepts can never be raw records of the signal because listener knowledge will inevitably interact with the stimulus and will contribute to shaping the percept.

Furthermore, H & H theory highlights the fact that spoken messages show a non-uniform distribution of information in that predictability varies from situation to situation, from word to word and from syllable to syllable. Compare (a) and (b) below. What word is likely to be represented by the gap?

1. "*The next word is* __."
2. "*A bird in the hand is worth two in the* __."

Any word can be expected in (1) whereas in (2) the predictability of the word "*bush*" is high.

H & H theory assumes that, while learning and using their native language, speakers develop a sense of this informational dynamics. Introducing the abstraction of an *ideal speaker*, it proposes that the talker estimates the running contribution that signal-complementary information (listener knowledge) will make during the course of the utterance and then tunes his articulatory performance to the presumed short-term listener needs. Statistically, this type of behavior has the long-term consequence of distributing phonetic output forms along a continuum with clear and elaborated forms (hyperspeech) at one end and casual and reduced pronunciations (hypospeech) at the other.

The implication for the invariance issue is that the task of the speaker is not to encode linguistic units as physical invariants, but to make sure that the signal attributes (of phonemes, syllables, words and phrases) carry discriminative power *sufficient* for successful lexical access. To make the same point in slightly different terms, the task of the signal is not to embody phonetic patterns of constancy but to provide *missing information*.

It is interesting to put this account of speech processes next to what we know about singing. Experimental observations indicate that singing in tune is not simply about invariant F0 targets. F0 is affected by the singer's expressive performance [16.177] and by the tonal context in which a given note is embedded [16.178]. Certain deviations from nominal target frequencies are not unlike the coarticulation and undershoot effects ubiquitously present in speech. Accordingly, with respect to frequency control, speaking and singing are qualitatively similar. However, quantitatively, they differ drastically. Recall that in describing the dynamics of vowel reduction we noted that formant fre-

quencies can be displaced by as much as 50% from target values. Clearly, a musician or singer with a comparable under-/overshoot record would be recommended to embark on an alternative career, the margins of perceptual tolerance being much narrower for singing.

What accounts for this discrepancy in target attainment? Our short answer is that singing or playing *out of tune* is a bad thing. Where does this taboo come from? From consonance and harmony constraints. In simplified terms, an arbitrary sample of tonal music can be analyzed into a sequence of chords. Its melodic line is a rhythmic and tonal elaboration of this harmonic structure. Statistically, long prominent melody tones tend to attract chord notes. Notes of less prominence typically *interpolate* along scales in smaller intervals between the metrically heavier chord notes. The notion of *consonance* goes a long way towards explaining why singing or playing out of tune is tabooed in music: hitting the right pitch is required by the consonance and harmony constraints expected by the listener and historically presumably linked with a combination of the polyphony in our Western tradition of music composition and the fact that most of our music instruments produce tones with harmonic spectra. This implies that intervals departing too much from just intonation will generate beats between simultaneously sounding and only nearly coinciding partials, particularly if the fundamental frequencies are constant, lacking vibrato and flutter.

16.9 The Expressive Power of the Human Voice

The human voice is an extremely expressive instrument both when used for speech and for singing. By means of subtle variations of timing and pitch contour speakers and singers add a substantial amount of expressiveness to the linguistic or musical content and we are quite skilled in deciphering this information. Indeed a good deal of vocal artistry seems to lie in the artist's skill in making nothing but such changes of pitch, timbre, loudness and timing that a listener can perceive as carrying some meaning.

We perceive the extra-linguistic or expressive information in speech and singing in various shapes. For example, we can interpret certain combinations of acoustic characteristics in speech in terms of a smile or a particular forming of the lips on the face of the speaker [16.179]. For example *Fónagy* [16.180] found that listeners were able to replicate rather accurately the facial expression of speakers only by listening to their voices.

The emotive transforms in speech and singing seem partly similar and sometimes identical [16.181, 182]. Final lengthening, mentioned above, uses the same code for marking the end of a structural element, such as a sentence in speech or a phrase in sung and played music performance. Emphasis by delayed arrival is another example, i. e., delaying an emphasized stressed syllable or note by lengthening the unstressed syllable/note preceding it [16.183].

The expressive potential of the human voice is indeed enormous, and would transpire from the ubiquitous use of the voice for the purpose of communication. Correct interpretation of the extra-linguistic content of a spoken utterance is certainly important in our daily life, so we are skilled in deciphering vocal signals also along those dimensions. The importance of correct encoding of the extra-linguistic implies that speakers acquire a great skill in this respect. This skill would be the basic requirement for vocal art, in singing as well as in acting.

References

16.1 T.J. Hixon: *Respiratory Function in Speech and Song* (Singular, San Diego 1991) pp. 1–54

16.2 A. L. Winkworth, P. J. Davis, E. Ellis, R. D. Adams: Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors, JSHR **37**, 535–556 (1994)

16.3 M. Thomasson: *From Air to Aria*, Ph.D. Thesis (Music Acoustics, KTH 2003)

16.4 B. Conrad, P. Schönle: Speech and respiration, Arch. Psychiat. Nervenkr. **226**, 251–268 (1979)

16.5 M.H. Draper, P. Ladefoged, D. Whitteridge: Respiratory muscles in speech, J. Speech Hear. Disord. **2**, 16–27 (1959)

16.6 R. Netsell: Speech physiology. In: *Normal aspects of speech, hearing, and language*, ed. by P.D. Minifie, T.J. Hixon, P. Hixon, P. Williams (Prentice-Hall, Englewood Cliffs 1973) pp. 211–234

16.7 P. Ladefoged, M.H. Draper, D. Whitteridge: Syllables and stress, Misc. Phonetica **3**, 1–14 (1958)

16.8 P. Ladefoged: Speculations on the control of speech. In: *A Figure of Speech: A Festschrift for John Laver*, ed. by W.J. Hardcastle, J. Mackenzie Beck (Lawrence Erlbaum, Mahwah 2005) pp. 3–21

16.9 T.J. Hixon, G. Weismer: Perspectives on the Edinburgh study of speech breathing, J. Speech Hear. Disord. **38**, 42–60 (1995)

16.10 S. Nooteboom: The prosody of speech: melody and rhythm. In: *The Handbook of Phonetic Sciences*, ed. by W.J. Hardcastle, J. Laver (Blackwell, Oxford 1997) pp. 640–673

16.11 M. Rothenberg: *The breath-stream dynamics of simple-released-plosive production Bibliotheca Phonetica 6* (Karger, Basel 1968)

16.12 D.H. Klatt, K.N. Stevens, J. Mead: Studies of articulatory activity and airflow during speech, Ann. NY Acad. Sci. **155**, 42–55 (1968)

16.13 J.J. Ohala: Respiratory activity in speech. In: *Speech production and speech modeling*, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 23–53

16.14 R.H. Stetson: *Motor Phonetics: A Study of Movements in Action* (North Holland, Amsterdam 1951)

16.15 P. Ladefoged: Linguistic aspects of respiratory phenomena, Ann. NY Acad. Sci. **155**, 141–151 (1968)

16.16 L.H. Kunze: Evaluation of methods of estimating sub-glottal air pressure muscles, J. Speech Hear. Disord. **7**, 151–164 (1964)

16.17 R. Leanderson, J. Sundberg, C. von Euler: Role of the diaphragmatic activity during singing: a study of transdiaphragmatic pressures, J. Appl. Physiol. **62**, 259–270 (1987)

16.18 J. Sundberg, N. Elliot, P. Gramming, L. Nord: Short-term variation of subglottal pressure for expressive purposes in singing and stage speech. A preliminary investigation, J. Voice **7**, 227–234 (1993)

16.19 J. Sundberg: Synthesis of singing, in Musica e Technologia: Industria e Cultura per lo Sviluppo del Mezzagiorno. In: *Proceedings of a symposium in Venice*, ed. by R. Favaro (Unicopli, Milan 1987) pp. 145–162

16.20 J. Sundberg: Synthesis of singing by rule. In: *Current Directions in Computer Music Research, System Development Foundation Benchmark Series*, ed. by M. Mathews, J. Pierce (MIT, Cambridge 1989), 45–55 & 401–403

16.21 J. Molinder: *Akustiska och perceptuella skillnader mellan röstfacken lyrisk och dramatisk sopran, unpublished thesis work* (Lund Univ. Hospital, Dept of Logopedics, Lund 1997)

16.22 T. Baer: Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes, J. Acoust. Soc. Am. **65**, 1271–1275 (1979)

16.23 T. Cleveland, J. Sundberg: Acoustic analyses of three male voices of different quality. In: *SMAC 83. Proceedings of the Stockholm Internat Music Acoustics Conf*, Vol. 1, ed. by A. Askenfelt, S. Felicetti, E. Jansson, J. Sundberg (Roy. Sw. Acad. Music, Stockholm 1985) pp. 143–156, No. 46:1

16.24 J. Sundberg, C. Johansson, H. Willbrand, C. Ytterbergh: From sagittal distance to area, Phonetica **44**, 76–90 (1987)

16.25 I.R. Titze: Phonation threshold pressure: A missing link in glottal aerodynamics, J. Acoust. Soc. Am. **91**, 2926–2935 (1992)

16.26 I.R. Titze: *Principles of Voice Production* (Prentice-Hall, Englewood Cliffs 1994)

16.27 G. Fant: *Acoustic theory of speech production* (Mouton, The Hague 1960)

16.28 K.N. Stevens: *Acoustic Phonetics* (MIT, Cambridge 1998)

16.29 M. Hirano: *Clinical Examination of Voice* (Springer, New York 1981)

16.30 M. Rothenberg: A new inversefiltering technique for deriving the glottal air flow waveform during voicing, J. Acoust. Soc. Am. **53**, 1632–1645 (1973)

16.31 G. Fant: The voice source – Acoustic modeling. In: *STL/Quart. Prog. Status Rep. 4* (Royal Inst. of Technology, Stockholm 1982) pp. 28–48

16.32 C. Gobl: *The voice source in speech communication production and perception experiments involving inverse filtering and synthesis. D.Sc. thesis* (Royal Inst. of Technology (KTH), Stockholm 2003)

16.33 G. Fant, J. Liljencrants, Q. Lin: A four-parameter model of glottal flow. In: *STL/Quart. Prog. Status Rep. 4, Speech, Music and Hearing* (Royal Inst. of Technology, Stockholm 1985) pp. 1–13

16.34 D.H. Klatt, L.C. Klatt: Analysis, synthesis and perception of voice quality variations among female and male talkers, J. Acoust. Soc. Am. **87**(2), 820–857 (1990)

16.35 M. Ljungqvist, H. Fujisaki: A comparative study of glottal waveform models. In: *Technical Report of the Institute of Electronics and Communications Engineers*, Vol. EA85-58 (Institute of Electronics and Communications Engineers, Tokyo 1985) pp. 23–29

16.36 A.E. Rosenberg: Effect of glottal pulse shape on the quality of natural vowels, J. Acoust. Soc. Am. **49**, 583–598 (1971)

16.37 M. Rothenberg, R. Carlson, B. Granström, J. Lindqvist-Gauffin: A three- parameter voice source for speech synthesis. In: *Proceedings of the Speech Communication Seminar 2*, ed. by G. Fant (Almqvist & Wiksell, Stockholm 1975) pp. 235–243

16.38 K. Ishizaka, J.L. Flanagan: Synthesis of voiced sounds from a two-mass model of the vocal cords, The Bell Syst. Tech. J. **52**, 1233–1268 (1972)

16.39 Liljencrants: Chapter A translating and rotating mass model of the vocal folds. In: *STL/Quart. Prog. Status Rep. 1, Speech, Music and Hearing* (Royal Inst. of Technology, Stockholm 1991) pp. 1–18

16.40 A. Ní Chasaide, C. Gobl: Voice source variation. In: *The Handbook of Phonetic Sciences*, ed. by

W.J. Hardcastle, J. Laver (Blackwell, Oxford 1997) pp. 427–462

16.41 E.B. Holmberg, R.E. Hillman, J.S. Perkell: Glottal air flow and pressure measurements for loudness variation by male and female speakers, J. Acoust. Soc. Am. **84**, 511–529 (1988)

16.42 J.S. Perkell, R.E. Hillman, E.B. Holmberg: Group differences in measures of voice production and revised values of maximum airflow declination rate, J. Acoust. Soc. Am. **96**, 695–698 (1994)

16.43 J. Gauffin, J. Sundberg: Spectral correlates of glottal voice source waveform characteristics, J. Speech Hear. Res. **32**, 556–565 (1989)

16.44 J. Svec, H. Schutte, D. Miller: On pitch jumps between chest and falsetto registers in voice: Data on living and excised human larynges, J. Acoust. Soc. Am. **106**, 1523–1531 (1999)

16.45 J. Sundberg, M. Andersson, C. Hultqvist: Effects of subglottal pressure variation on professional baritone singers voice sources, J. Acoust. Soc. Am. **105**, 1965–1971 (1999)

16.46 J. Sundberg, E. Fahlstedt, A. Morell: Effects on the glottal voice source of vocal loudness variation in untrained female and male subjects, J. Acoust. Soc. Am. **117**, 879–885 (2005)

16.47 P. Sjölander, J. Sundberg: Spectrum effects of subglottal pressure variation in professional baritone singers, J. Acoust. Soc. Am. **115**, 1270–1273 (2004)

16.48 P. Branderud, H. Lundberg, J. Lander, H. Djamshidpey, I. Wäneland, D. Krull, B. Lindblom: *X-ray analyses of speech: Methodological aspects, Proc. of 11th Swedish Phonetics Conference* (Stockholm Univ., Stockholm 1996) pp. 168–171

16.49 B. Lindblom: A numerical model of coarticulation based on a Principal Components analysis of tongue shapes. In: *Proc. 15th Int. Congress of the Phonetic Sciences*, ed. by D. Recasens, M. Josep Solé, J. Romero (Universitat Autònoma de Barcelona, Barcelona 2003), CD-ROM

16.50 G.E. Peterson, H. Barney: Control methods used in a study of the vowels, J. Acoust. Soc. Am. **24**, 175–184 (1952)

16.51 Hillenbrand et al.: Acoustic characteristics of American English vowels, J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995)

16.52 G. Fant: Analysis and synthesis of speech processes. In: *Manual of Phonetics*, ed. by B. Malmberg (North-Holland, Amsterdam 1968) pp. 173–277

16.53 G. Fant: Formant bandwidth data. In: *STL/Quart. Prog. Status Rep. 7* (Royal Inst. of Technology, Stockholm 1962) pp. 1–3

16.54 G. Fant: Vocal tract wall effects, losses, and resonance bandwidths. In: *STL/Quart. Prog. Status Rep. 2–3* (Royal Inst. of Technology, Stockholm 1972) pp. 173–277

16.55 A.S. House, K.N. Stevens: Estimation of formant bandwidths from measurements of transient response of the vocal tract, J. Speech Hear. Disord. **1**, 309–315 (1958)

16.56 O. Fujimura, J. Lindqvist: Sweep-tone measurements of vocal-tract characteristics, J. Acoust. Soc. Am. **49**, 541–558 (1971)

16.57 I. Lehiste, G.E. Peterson: Vowel amplitude and phonemic stress in American English, J. Acoust. Soc. Am. **3**, 428–435 (1959)

16.58 I. Lehiste: *Suprasegmentals* (MIT Press, Cambridge 1970)

16.59 O. Jespersen: *Lehrbuch der Phonetik* (Teubner, Leipzig 1926)

16.60 T. Bartholomew: A physical definition of good voice quality in the male voice, J. Acoust. Soc. Am. **6**, 25–33 (1934)

16.61 J. Sundberg: Production and function of the singing formant. In: *Report of the eleventh congress Copenhagen 1972 (Proceedings of the 11th international congress of musicology)*, ed. by H. Glahn, S. Sörensen, P. Ryom (Wilhelm Hansen, Copenhagen 1972) pp. 679–686

16.62 J. Sundberg: Articulatory interpretation of the 'singing formant', J. Acoust. Soc. Am. **55**, 838–844 (1974)

16.63 J. Sundberg: Level and center frequency of the singer´s formant, J. Voice. **15**, 176–186 (2001)

16.64 G. Berndtsson, J. Sundberg: Perceptual significance of the center frequency of the singers formant, Scand. J. Logopedics Phoniatrics **20**, 35–41 (1995)

16.65 L. Dmitriev, A. Kiselev: Relationship between the formant structure of different types of singing voices and the dimension of supraglottal cavities, Fol. Phoniat. **31**, 238–41 (1979)

16.66 P. Ladefoged: *Three areas of experimental phonetics* (Oxford Univ. Press, London 1967)

16.67 J. Barnes, P. Davis, J. Oates, J. Chapman: The relationship between professional operatic soprano voice and high range spectral energy, J. Acoust. Soc. Am. **116**, 530–538 (2004)

16.68 M. Nordenberg, J. Sundberg: Effect on LTAS on vocal loudness variation, Logopedics Phoniatrics Vocology **29**, 183–191 (2004)

16.69 R. Weiss, W.S. Brown, J. Morris: Singer's formant in sopranos: Fact or fiction, J. Voice **15**, 457–468 (2001)

16.70 J.M. Heinz, K.N. Stevens: On the relations between lateral cineradiographs, area functions, and acoustics of speech. In: *Proc. Fourth Int. Congress on Acoustics*, Vol. 1a (1965), paper A44

16.71 C. Johansson, J. Sundberg, H. Willbrand: X-ray study of articulation and formant frequencies in two female singers. In: *Proc. of Stockholm Music Acoustics Conference 1983 (SMAC 83)*, Vol. 46(1), ed. by A. Askenfelt, S. Felicetti, E. Jansson, J Sundberg (Kgl. Musikaliska Akad., Stockholm 1985) pp. 203–218

16.72 T. Baer, J.C. Gore, L.C. Gracco, P. Nye: Analysis of vocal tract shape and dimensions using magnetic

resonance imaging: Vowels, J. Acoust. Soc. Am. **90**(2), 799–828 (1991)
16.73 D. Demolin, M. George, V. Lecuit, T. Metens, A. Soquet: Détermination par IRM de l'ouverture du velum des voyelles nasales du français. In: *Actes des XXièmes Journées d'Études sur la Parole* (1996)
16.74 A. Foldvik, K. Kristiansen, J. Kvaerness, A. Torp, H. Torp: *Three-dimensional ultrasound and magnetic resonance imaging: a new dimension in phonetic research* (Proc. Fut. Congress Phonetic Science Stockholm Univ., Stockholm 1995), Vol. 4, 46–49
16.75 B.H. Story, I.R. Titze, E.A. Hoffman: Vocal tract area functions from magnetic resonance imaging, J. Acoust. Soc. Am. **100**, 537–554 (1996)
16.76 O. Engwall: *Talking tongues, D.Sc. thesis* (Royal Institute of Technology (KTH), Stockholm 2002)
16.77 B. Lindblom, J. Sundberg: Acoustical consequences of lip, tongue, jaw and larynx movement, J. Acoust. Soc. Am. **50**, 1166–1179 (1971), also in Papers in Speech Communication: Speech Production, ed. by R.D. Kent, B.S. Atal, J.L. Miller (Acoust. Soc. Am., New York 1991) pp.329–342
16.78 J. Stark, B. Lindblom, J. Sundberg: APEX – an articulatory synthesis model for experimental and computational studies of speech production. In: *Fonetik 96: Papers presented at the Swedish Phonetics Conference TMH-QPSR 2/1996* (Royal Institute of Technology, Stockholm 1996) pp. 45–48
16.79 J. Stark, C. Ericsdotter, B. Lindblom, J. Sundberg: Using X-ray data to calibrate the APEX the synthesis. In: *Fonetik 98: Papers presented at the Swedish Phonetics Conference* (Stockholm Univ., Stockholm 1998)
16.80 J. Stark, C. Ericsdotter, P. Branderud, J. Sundberg, H.-J. Lundberg, J. Lander: The APEX model as a tool in the specification of speaker-specific articulatory behavior. In: *Proc. 14th Int. Congress of the Phonetic Sciences*, ed. by J.J. Ohala (1999)
16.81 C. Ericsdotter: Articulatory copy synthesis: Acoustic performane of an MRI and X-ray-based framework. In: *Proc. 15th Int. Congress of the Phonetic Sciences*, ed. by D. Recasens, M. Josep Solé, J. Romero (Universitat Autònoma de Barcelona, Barcelona 2003), CD-ROM
16.82 C. Ericsdotter: *Articulatory-acoustic relationships in Swedish vowel sounds, PhD thesis* (Stockholm University, Stockholm 2005)
16.83 K.N. Stevens, A.S. House: Development of a quantitative description of vowel articulation, J. Acoust. Soc. Am. **27**, 484–493 (1955)
16.84 S. Maeda: Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: *Speech Production and Speech Modeling*, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 131–150
16.85 P. Branderud, H. Lundberg, J. Lander, H. Djamshidpey, I. Wäneland, D. Krull, B. Lindblom: X-ray analyses of speech: methodological aspects. In: *Proc. XIIIth Swedish Phonetics Conf. (FONETIK 1998)* (KTH, Stockholm 1998)
16.86 C.Y. Espy-Wilson: Articulatory strategies, speech acoustics and variability. In: *From sound to Sense: 50+ Years of Discoveries in Speech Communication*, ed. by J. Slifka, S. Manuel, M. Mathies (MIT, Cambridge 2004)
16.87 J. Sundberg: Formant technique in a professional female singer, Acustica **32**, 89–96 (1975)
16.88 J. Sundberg, J. Skoog: Dependence of jaw opening on pitch and vowel in singers, J. Voice **11**, 301–306 (1997)
16.89 G. Fant: Glottal flow, models and interaction, J. Phon. **14**, 393–399 (1986)
16.90 E. Joliveau, J. Smith, J. Wolfe: Vocal tract resonances in singing: The soprano voice, J. Acoust. Soc. Am. **116**, 2434–2439 (2004)
16.91 R.K. Potter, A.G. Kopp, H.C. Green: *Visible Speech* (Van Norstrand, New York 1947)
16.92 M. Joosg: Acoustic phonetics, Language **24**, 447–460 (2003), supplement 2
16.93 C.F. Hockett: *A Manual of Phonology* (Indiana Univ. Publ., Bloomington 1955)
16.94 F.H. Guenther: Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production, Psychol. Rev. **102**, 594–621 (1995)
16.95 R.D. Kent, B.S. Atal, J.L. Miller: *Papers in Speech Communication: Speech Perception* (Acoust. Soc. Am., New York 1991)
16.96 S.D. Goldinger, D.B. Pisoni, P. Luce: Speech perception and spoken word recognition. In: *Principles of experimental phonetics*, ed. by N.J. Lass (Mosby, St Louis 1996) pp. 277–327
16.97 H.M. Sussman, D. Fruchter, J. Hilbert, J. Sirosh: Linear correlates in the speech signal: The orderly output constraint, Behav. Brain Sci. **21**, 241–299 (1998)
16.98 B. Lindblom: Economy of speech gestures. In: *The Production of Speech*, ed. by P.F. MacNeilage (Springer, New York 1983) pp. 217–245
16.99 P.A. Keating, B. Lindblom, J. Lubker, J. Kreiman: Variability in jaw height for segments in English and Swedish VCVs, J. Phonetics **22**, 407–422 (1994)
16.100 K. Rapp: A study of syllable timing. In: *STL/Quart. Prog. Status Rep. 1* (Royal Inst. of Technology, Stockholm(1971) pp. 14–19
16.101 F. Koopmans-van Beinum, J. Van der Stelt (Eds.): *Early stages in the development of speech movements* (Stockton, New York 1986)
16.102 K. Oller: Metaphonology and infant vocalizations. In: *Precursors of early speech*, ed. by B. Lindblom, R. Zetterström (Stockton, New York 1986) pp. 21–36

16.103 L. Roug, L. Landberg, L. Lundberg: Phonetic development in early infancy, J. Child Language **16**, 19–40 (1989)

16.104 R. Stark: Stages of speech development in the first year of life. In: *Child Phonology: Volume 1: Production*, ed. by G. Yeni-Komshian, J. Kavanagh, C. Ferguson (Academic, New York 1980) pp. 73–90

16.105 R. Stark: Prespeech segmental feature development. In: *Language Acquisition*, ed. by P. Fletcher, M. Garman (Cambridge UP, New York 1986) pp. 149–173

16.106 D.K. Oller, R.E. Eilers: The role of audition in infant babbling, Child Devel. **59**(2), 441–449 (1988)

16.107 C. Stoel-Gammon, D. Otomo: Babbling development of hearing-impaired and normally hearing subjects, J. Speech Hear. Dis. **51**, 33–41 (1986)

16.108 R.E. Eilers, D.K. Oller: Infant vocalizations and the early diagnosis of severe hearing impairment, J. Pediatr. **124**(2), 99–203 (1994)

16.109 D. Ertmer, J. Mellon: Beginning to talk at 20 months: Early vocal development in a young cochlear implant recipient, J. Speech Lang. Hear. Res. **44**, 192–206 (2001)

16.110 R.D. Kent, M.J. Osberger, R. Netsell, C.G. Hustedde: Phonetic development in identical twins who differ in auditory function, J. Speech Hear. Dis. **52**, 64–75 (1991)

16.111 M. Lynch, D. Oller, M. Steffens: Development of speech-like vocalizations in a child with congenital absence of cochleas: The case of total deafness, Appl. Psychol. **10**, 315–333 (1989)

16.112 C. Stoel-Gammon: Prelinguistic vocalizations of hearing-impaired and normally hearing subjects: A comparison of consonantal inventories, J. Speech Hear. Dis. **53**, 302–315 (1988)

16.113 P.F. MacNeilage, B.L. Davis: Acquisition of speech production: The achievement of segmental independence. In: *Speech production and speech modeling*, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 55–68

16.114 T. Houtgast, H.J.M. Steeneken: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria, J. Acoust. Soc. Am. **77**, 1069–1077 (1985)

16.115 T. Houtgast, H.J.M. Steeneken: *Past, Present and Future of the Speech Transmission Index*, ed. by S.J. van Wijngaarden (NTO Human Factors, Soesterberg 2002)

16.116 R. Drullman, J.M. Festen, R. Plomp: Effect of temporal envelope smearing on speech reception, J. Acoust. Soc. Am. **95**, 1053–1064 (1994)

16.117 R. Drullman, J.M. Festen, R. Plomp: Effect of reducing slow temporal modulations on speech reception, J. Acoust. Soc. Am. **95**, 2670–2680 (1994)

16.118 J. Morton, S. Marcus, C. Frankish: Perceptual centers (P-centers), Psych. Rev. **83**, 405–408 (1976)

16.119 S. Marcus: Acoustic determinants of perceptual center (P-center) location, Perception & Psychophysics **30**, 247–256 (1981)

16.120 G. Allen: The place of rhythm in a theory of language, UCLA Working Papers **10**, 60–84 (1968)

16.121 G. Allen: The location of rhythmic stress beats in English: An experimental study, UCLA Working Papers **14**, 80–132 (1970)

16.122 J. Eggermont: Location of the syllable beat in routine scansion recitations of a Dutch poem, IPO Annu. Prog. Rep. **4**, 60–69 (1969)

16.123 V.A. Kozhevnikov, L.A. Chistovich: Speech Articulation and Perception, JPRS **30**, 543 (1965)

16.124 C.E. Hoequist: The perceptual center and rhythm categories, Lang. Speech **26**, 367–376 (1983)

16.125 K.J. deJong: The correlation of P-center adjustments with articulatory and acoustic events, Perception Psychophys. **56**, 447–460 (1994)

16.126 A.D. Patel, A. Löfqvist, W. Naito: The acoustics and kinematics of regularly timed speech: a database and method for the study of the P-center problem. In: *Proc. 14th Int. Congress of the Phonetic Sciences*, ed. by J.J. Ohala (1999)

16.127 P. Howell: Prediction of P-centre location from the distribution of energy in the amplitude envelope: I & II, Perception Psychophys. **43**, 90–93, 99 (1988)

16.128 B. Pompino-Marschall: On the psychoacoustic nature of the Pcenter phenomenon, J. Phonetics **17**, 175–192 (1989)

16.129 C.A. Harsin: Perceptual-center modeling is affected by including acoustic rate-of-change modulations, Perception Psychophys. **59**, 243–251 (1997)

16.130 C.A. Fowler: Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet, J. Exp. Psychol. Gen. **112**, 386–412 (1983)

16.131 H. Fujisaki: Dynamic characteristics of voice fundamental frequency in speech and singing. In: *The Production of Speech*, ed. by P.F. MacNeilage (Springer, New York 1983) pp. 39–55

16.132 J. Frid: *Lexical and acoustic modelling of Swedish prosody, Dissertation* (Lund University, Lund 2003)

16.133 S. Öhman: Numerical model of coarticulation, J. Acoust. Soc. Am. **41**, 310–320 (1967)

16.134 J. t'Hart: F0 stylization in speech: Straight lines versus parabolas, J. Acoust. Soc. Am. **90**(6), 3368–3370 (1991)

16.135 D. Abercrombie: *Elements of General Phonetics* (Edinburgh Univ. Press, Edinburgh 1967)

16.136 K.L. Pike: *The intonation of America English* (Univ. of Michigan Press, Ann Arbor 1945)

16.137 G. Fant, A. Kruckenberg: Notes on stress and word accent in Swedish. In: *STL/Quart. Prog. Status Rep. 2-3* (Royal Inst. of Technology, Stockholm 1994) pp. 125–144

16.138 R. Dauer: Stress timing and syllable-timing reanalyzed, J. Phonetics **11**, 51–62 (1983)

16.139 A. Eriksson: *Aspects of Swedish rhythm, PhD thesis, Gothenburg Monographs in Linguistics* (Gothenburg University, Gothenburg 1991)
16.140 O. Engstrand, D. Krull: Duration of syllable-sized units in casual and elaborated speech: cross-language observations on Swedish and Spanish, TMH-QPSR **44**, 69–72 (2002)
16.141 A.D. Patel, J.R. Daniele: An empirical comparison of rhythm in language and music, Cognition **87**, B35–B45 (2003)
16.142 D. Huron, J. Ollen: Agogic contrast in French and English themes: Further support for Patel and Daniele (2003), Music Perception **21**, 267–271 (2003)
16.143 D.H. Klatt: Synthesis by rule of segmental durations in English sentences. In: *Frontiers of speech communication research*, ed. by B. Lindblom, S. Öhman (Academic, London 1979) pp. 287–299
16.144 B. Lindblom: Final lengthening in speech and music. In: *Nordic Prosody*, ed. by E. Gårding, R. Bannert (Department of Linguistics Lund University, Lund 1978) pp. 85–102
16.145 A. Friberg, U Battel: Structural communication. In: *The Science and Psychology of Music Performance*, ed. by R. Parncutt, GE McPherson (Oxford Univ., Oxford 2001) pp. 199–218
16.146 J. Sundberg: Emotive transforms, Phonetica **57**, 95–112 (2000)
16.147 Brownlee: *The role of sentence stress in vowel reduction and formant undershoot: A study of lab speech and informal spontaneous speech, PhD thesis* (University of Texas, Austin 1996)
16.148 S.-J. Moon: *An acoustic and perceptual study of undershoot in clear and citation- form speech, PhD dissertation* (Univ. of Texas, Austin 1991)
16.149 K.N. Stevens, A.S. House: Perturbation of vowel articulations by consonantal context. An acoustical study, JSHR **6**, 111–128 (1963)
16.150 B. Lindblom: Spectrographic study of vowel reduction, J. Acoust. Soc. Am. **35**, 1773–1781 (1963)
16.151 P. Delattre: An acoustic and articulatory study of vowel reduction in four languages, IRAL-Int. Ref. Appl. VII/ **4**, 295–325 (1969)
16.152 D.P. Kuehn, K.L. Moll: A cineradiographic study of VC and CV articulatory velocities, J. Phonetics **4**, 303–320 (1976)
16.153 J.E. Flege: Effects of speaking rate on tongue position and velocity of movement in vowel production, J. Acoust. Soc. Am. **84**(3), 901–916 (1988)
16.154 R.J.J.H. van Son, L.C.W. Pols: "Formant movements of Dutch vowels in a text, read at normal and fast rate, J. Acoust. Soc. Am. **92**(1), 121–127 (1992)
16.155 D. van Bergem: *Acoustic and Lexical Vowel Reduction, Doctoral Dissertation* (University of Amsterdam, Amsterdam 1995)
16.156 W.L. Nelson, J.S. Perkell, J.R. Westbury: Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations, J. Acoust. Soc. Am. **75**(3), 945–951 (1984)
16.157 S.-J. Moon, B. Lindblom: Interaction between duration, context and speaking style in English stressed vowels, J. Acoust. Soc. Am. **96**(1), 40–55 (1994)
16.158 C.S. Sherrington: *Man on his nature* (MacMillan, London 1986)
16.159 R. Granit: *The Purposive Brain* (MIT, Cambridge 1979)
16.160 N. Bernstein: *The coordination and regulation of movements* (Pergamon, Oxford 1967)
16.161 P.F. MacNeilage: Motor control of serial ordering of speech, Psychol. Rev. **77**, 182–196 (1970)
16.162 A. Löfqvist: Theories and Models of Speech Production. In: *The Handbook of Phonetic Sciences*, ed. by W.J. Hardcastle, J. Laver (Blackwell, Oxford 1997) pp. 405–426
16.163 J.S. Perkell: Articulatory processes. In: *The Handbook of Phonetic Sciences. 5*, ed. by W.J. Hardcastle, J. Laver. (Blackwell, Oxford 1997) pp. 333–370
16.164 J. Sundberg, R. Leandersson, C. von Euler, E. Knutsson: Influence of body posture and lung volume on subglottal pressure control during singing, J. Voice **5**, 283–291 (1991)
16.165 T. Sears, J. Newsom Davis: The control of respiratory muscles during voluntary breathing. In: *Sound production in man*, ed. by A. Bouhuys et al. (Annals of the New York Academy of Science, New York 1968) pp. 183–190
16.166 B. Lindblom, J. Lubker, T. Gay: Formant frequencies of some fixed-mandible vowels and a model of motor programming by predictive simulation, J. Phonetics **7**, 147–161 (1979)
16.167 T. Gay, B. Lindblom, J. Lubker: Production of bite-block vowels: Acoustic equivalence by selective compensation, J. Acoust. Soc. Am. **69**(3), 802–810 (1981)
16.168 W.J. Hardcastle, J. Laver (Eds.): *The Handbook of Phonetic Sciences* (Blackwell, Oxford 1997)
16.169 J. S. Perkell, D. H. Klatt: *Invariance and variability in speech processes* (LEA, Hillsdale 1986)
16.170 A. Liberman, I. Mattingly: The motor theory of speech perception revised, Cognition **21**, 1–36 (1985)
16.171 C.A. Fowler: An event approach to the study of speech perception from a direct- realist perspective, J. Phon. **14**(1), 3–28 (1986)
16.172 E.L. Saltzman, K.G. Munhall: A dynamical approach to gestural patterning in speech production, Ecol. Psychol. **1**, 91–163 (1989)
16.173 M. Studdert-Kennedy: How did language go discrete?. In: *Evolutionary Prerequisites of Language*, ed. by M. Tallerman (Oxford Univ., Oxford 2005) pp. 47–68
16.174 R. Jakobson, G. Fant, M. Halle: *Preliminaries to Speech Analysis, Acoustics Laboratory, MIT Tech. Rep. No. 13* (MIT, Cambridge 1952)
16.175 B. Lindblom: Explaining phonetic variation: A sketch of the H&H theory. In: *Speech Produc-*

tion and Speech Modeling, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 403–439

16.176 B. Lindblom: Role of articulation in speech perception: Clues from production, J. Acoust. Soc. Am. **99**(3), 1683–1692 (1996)

16.177 E. Rapoport: Emotional expression code in opera and lied singing, J. New Music Res. **25**, 109–149 (1996)

16.178 J. Sundberg, E. Prame, J. Iwarsson: Replicability and accuracy of pitch patterns in professional singers. In: *Vocal Fold Physiology, Controlling Complexity and Chaos*, ed. by P. Davis, N. Fletcher (Singular, San Diego 1996) pp. 291–306, Chap. 20

16.179 J.J. Ohala: An ethological perspective on common cross-language utilization of F0 of voice, Phonetica **41**, 1–16 (1984)

16.180 I. Fónagy: Hörbare Mimik, Phonetica **1**, 25–35 (1967)

16.181 K. Scherer: Expression of emotion in voice and music, J. Voice **9**, 235–248 (1995)

16.182 P. Juslin, P. Laukka: Communication of emotions in vocal expression and music performance: Different channels, same code?, Psychol. Rev. **129**, 770–814 (2003)

16.183 J. Sundberg, J. Iwarsson, H. Hagegård: A singers expression of emotions in sung performance,. In: *Vocal Fold Physiology: Voice Quality Control*, ed. by O. Fujimura, M. Hirano (Singular, San Diego 1995) pp. 217–229