1 Introduction

The respiratory cycle

One of the most remarkable features of phonation is the disruption of the normal respiratory cycle. Outside phonation, the normal cycle of respiration presents a comparable duration for both the inspiration and the expiration phases (top of Fig. 1.1), whereas during phonation the expiratory phase is usually much longer than the inspiratory phase.

Indeed, the phonation process results from the air flow generated by the lung compression during the respiration–expiration phase. This air flow generates the subglottal pressure needed to produce the vibration of the vocal folds for voiced sounds (vowels, voiced consonants), friction for fricative consonants, and intraoral pressure to allow the production of stop consonants. As a result, phonation is possible only during the expiration phase.

While producing speech, the speaker has to optimize the duration of both the inspiration and the expiration phases of the respiratory cycle, so that the expiration phase is the longest possible and the inspiration phase the shortest possible (bottom of Fig. 1.1). The latter should induce an acceptable duration of silence that fits with the specific conditions of the speech act (usually at a syntactic boundary). On the other hand, expiration should correspond to the speaker's estimation of the appropriate quantity of expired air necessary to produce the planned sequence of syllables, with the desired parameter values of rhythm, intensity, and laryngeal frequency. As prosodic unit, the *breath group* contains the prosodic objects produced during the expiration phase, i.e. between the consecutive inspiration phases.

The air consumption, and therefore the cycle of respiration, depends on the speaker's emotional state, and more specifically on the energy needs related to the speaker's emotions. The respiratory cycle will be longer at rest, and will accelerate when the activity level increases, as it requires more oxygen, until it gets heavily reduced by physical effort (swimming, carrying heavy objects, racing, etc.) to a time span that allows only a short phonation time.

The speaker's emotionally depressed state consumes less energy with less pulmonary air volume and allows a slower phonation rhythm while producing

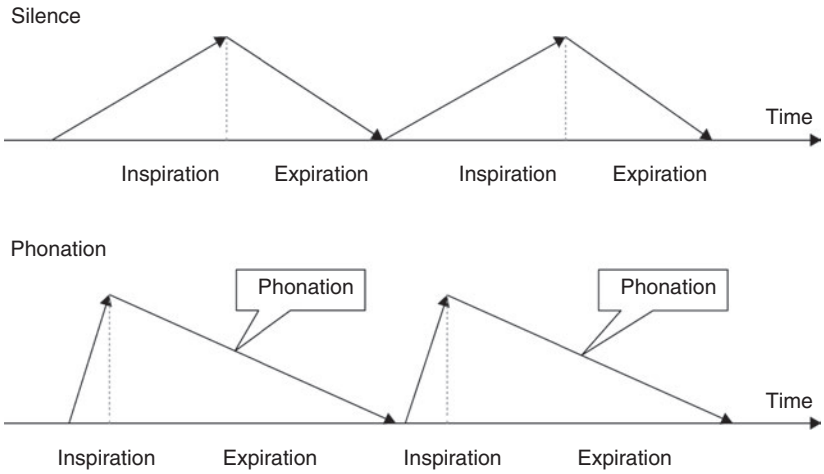

Figure 1.1 Respiration cycle, without phonation (top) and with phonation (bottom).

the same number of syllables that a speaker would in a more neutral emotional state. On the other hand, some types of anger and fear consume a lot of physiological energy. This state leads to shorter phonation sequences, which may not even reach the duration of a single sentence, or of a complete syntagm, and which may end with an unexpected (for the listener) respiratory pause of considerable duration.

An example of an out-of-breath speaker phonation cycle, when the speaker needs inspiratory pauses that are longer and more frequent than usual, is given in Figure 1.2.

It is clear, then, that in order to master her/his speaking activity, the speaker must constantly control her/his physiological state, which is itself conditioned by an emotional state, in order to control the air volume inspired to the lungs and to maintain a sufficient subglottal pressure during expiration. The larger the pulmonic air debit, the shorter the phonation time, as, for example, during phonation with high acoustic intensity, a higher than usual laryngeal frequency, or with a large amount of melodic variation. Conversely, a low debit of pulmonic air will result in low-intensity speech and a lower laryngeal frequency with reduced melodic variation.

Pulmonic air expiration requires a control of the vocal folds tension (the word *tension* stands here for the complex muscular mechanisms controlling the positioning and the elongation of the vocal folds) in order to compensate for the diminution of the lung air volume during expiration. As this air volume diminishes, the subglottal pressure mechanically diminishes as well, since an

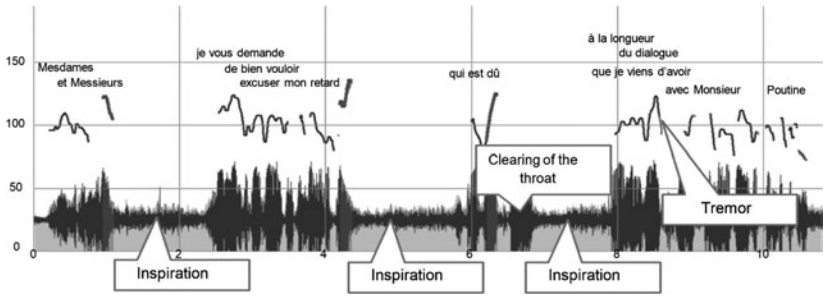

Figure 1.2 An example of an out-of-breath speaker (NS), that is, when a speaker needs inspiratory pauses that are longer and more frequent than usual: *Mesdames et Messieurs # je vous demande de bien vouloir excuser mon retard # qui est dû # à la longueur du dialogue que je viens d'avoir avec Monsieur Poutine* [NS, 2007] ("Ladies and Gentlemen # I ask you to excuse my delay # which is due # to the long dialog I had with Mr. Poutine").

adequate and complete lung compression is impossible to achieve. This drop must be compensated for by the speaker, according to his or her expectation of the number of syllables to be pronounced in the planned breath group (the interval between two consecutive inspiration phases). This mechanism is not always completely compensated for by the muscles controlling the expiration, which in turn may partly explain the origin of the frequently observed *declination line* of the laryngeal frequency, i.e. the tendency of the laryngeal frequency to be higher at the beginning than at the end of the expiration phase.

The source-filter model of phonation

The speech production mechanism is constrained by the speaker's specific physiological and emotional state and, at the same time, by phonological, syntactic, and semantic constraints on language functions. A particularly simple model frequently used in speech processing represents the phonation mechanism by two separate processes, a sound source and a filter, that shapes the speech spectral characteristics of the source (Fig. 1.3).

The speech source whose characteristics are deemed to represent at the same time not only the vocal folds vibrations for voiced sounds (vowels and voiced fricatives such as [v], [z], [ʒ]), but also the friction noise used to produce consonants such as [f], [s], [ʃ] (despite the fact that the noise source is not actually localized in the glottis). The filter in this source-filter model, which represents the shaping action of the vocal tract on the sound spectrum produced by the source, possesses characteristics allowing for the amplitude of the

Figure 1.3 Source-filter model of phonation.

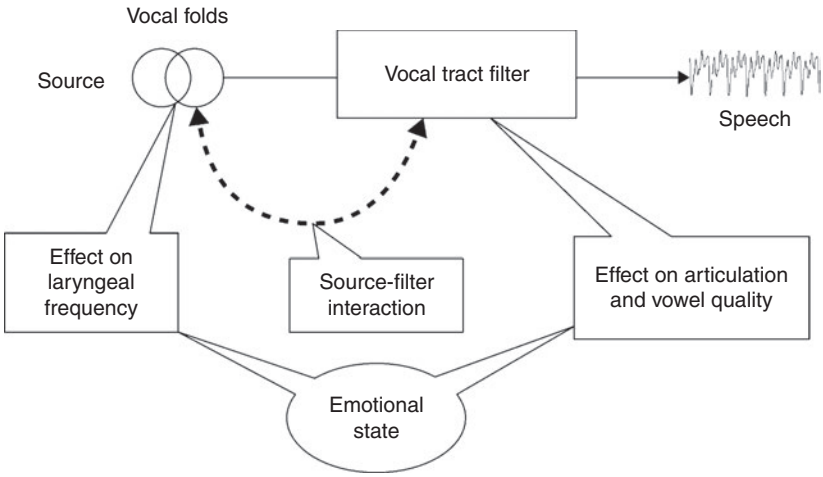

Figure 1.4 Interactions in the source-filter model between phonation and emotions.

harmonics of voiced sounds to be shaped, on the one hand, and of noise regions of the fricatives (both for voiced and unvoiced) on the other hand. Stop consonants such as [p], [t], [k] are simply not taken into account in this model, although voiced stops [b], [d], [g] are partially represented by their voiced character. However, this is partially justified, as the perception of stop consonants is based essentially on spectral transitions on the vowel (if any) that follows (spectral loci).

If the speaker's emotional state has to be taken into account in a model, it is essential to consider the interactions necessarily existing between the source and the filter (Fig. 1.4). Indeed, the emotional state has an effect on the physiological mechanism of phonation, as it affects the respiration cycle, the volume of air inspired (resulting in the speech rate), the subglottal pressure, and the tension of the vocal folds, which, in turn, determines the laryngeal frequency and the voice pitch. The position of the articulators is also modified, conditioning vowel quality. This emotional state affects the muscular tension

responsible for the positioning of the articulators, which are modeled by the filter. It also produces secondary effects on the source characteristics (for example, on the control of the laryngeal frequency and the position of the glottis in the vocal tract).

Emotions

One can easily say that there are as many categories of emotions as there are authors dealing with the subject. For example, in what may appear as a continuum, Eckman (1999) distinguishes the following basic categories: Joy, Sadness, Disgust, Fear, Anger, and Surprise, with secondary emotions resulting from a mixture of these basic emotions. Shame, for example, can be considered as a mixed emotion, combining fear and anger directed at oneself. Eckman's categories of emotion, like many other systems, obviously pertains to the terminology of emotions, often influenced or even determined by categories existing in the language of the researchers (cf. color terminology or snow quality in Inuktitut, etc.).

Physiological constraints linked to various emotions were often studied (e.g. Sauleau 2010). Factors prone to influence phonation are salivation, muscular tension, perspiration, or more globally, blood pressure, and cardiac frequency.

The physiological parameters controlling phonation affected by emotions are essentially as follows:

a. Energy, which acts on voicing and vowel quality;
b. Tension of the vocal folds, which determines the melodic height as well as vowel quality;
c. Articulation, another factor affecting vowel quality;
d. Speech rate, responsible for the tense or lax mode of articulation;
e. The degree of voicing, characterizing the noise/source ratio (breathy voice);
f. Breath insertion, as an index of irritation, pleasure, fear (iconic value);
g. Uncontrolled muscular movements (shivering) acting on the laryngeal frequency as well as on vowel quality;
h. Regulation of the respiration cycle, which determines the position and the length of pauses.

Dominance of an emotional state occurs when linguistic rules and constraints are not fulfilled in the realization of vowels, consonants, and the prosodic structure.

However, emotion affects the whole phonation process (laryngeal source and vocal tract), as well as all of the syllables, whereas the dialectal or idiosyncratic variations pertain essentially to stressed (prominent) syllables. An extreme case of this process is shown in Figure 1.5. The borderline cases correspond to the "hot" anger and extreme stress, for which emotion disturbs all or some aspects of the phonological realizations of prosodic markers, and at the other end of the scale,

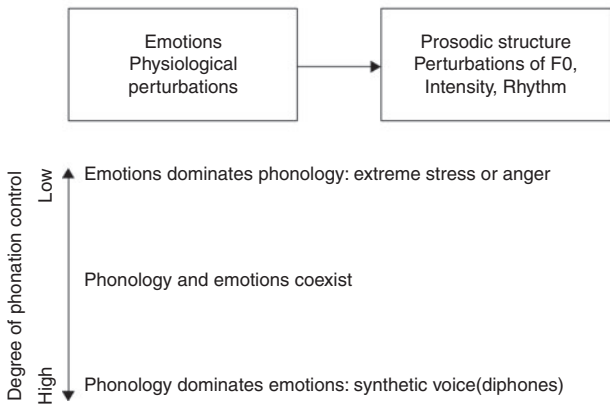

Figure 1.5 Extreme cases of the emotion–phonology relationship: emotion dominates phonology (extreme stress or anger), and phonology dominates emotions (diphone speech synthesis). In the case of prosody, emotions influence acoustical parameters such as fundamental frequency (F0), intensity, and rhythm.

synthetic speech based on diphones, totally deprived of emotional content (contrary to speech synthesis by corpus which necessarily presents traces of emotions of speakers who are involved in building the corpus).

Verbal communication by speech always includes an emotional component, and can be placed between these two extreme cases, affecting at various levels realization of speech units. The realization of melodic contours can also vary according to the socio-geographical origin of the speakers as well as their idiosyncratic characteristics, but the speaker's emotional state can affect the prosodic structure in various ways, as in the following:

a. Interruption of a stress group (a sequence of syllables with only one stressed syllable) due to a perturbed control of the respiration cycle, in particular during the expiration phase (e.g. in the case of extreme fear or anger). Stress groups can then become difficult or impossible to identify by the listener, particularly in the case of erroneous syntactic grouping.
b. Sequences of incomplete melodic contours (melodic variations on stressed syllables), without final conclusive contour (abandons).
c. Insufficient acoustic contrasts in the realization of melodic contours preventing a correct identification by the listener, resulting, for example, in a "flat" prosodic structure (with only one level leading to an enumeration structure. see Chapter 5).

In conclusion, the realizations of linguistic units, and in particular prosodic ones, due to emotional variations is purely phonetic and does not affect the linguistic functions ensured by these units, except in extreme cases (Martin, 2014a).

Voiced and unvoiced speech sounds

As mentioned earlier, speech sounds are produced by a variety of mechanisms that involve various noise sources: (1) the vibration of the vocal folds for vowels and some "voiced" (except, of course, whispered) consonants; (2) a narrowing of the vocal tract forcing expiratory air to enter a turbulent regime (called "fricatives" consonants [f], [s], or [ʃ]); (3) a micro-explosion caused by a sudden release of the closure of the vocal tract which increase[s] the pressure upstream of the closure (such as stop consonants [p], [t], or [k]); and (4) a micro-implosion caused by the sudden release of a location of a closure in the vocal tract in which a depression is produced by reduction of its volume (clicks). These modes can be combined (with the exception of the implosion), and when the vocal folds are involved by their vibration this is called "voiced sound." If this is not the case, the sound will be called "unvoiced."

Vowels produced by vocal folds vibrations are always voiced (but may be devoiced in some circumstances or whispered). However, only consonants generated with vibration of the vocal folds, such as [b], [d], and [g] for stop consonants, [v], [z], and [ʒ] for fricatives, and [m], [n], [ɲ], and [ŋ] for nasals, are voiced.

Laryngeal frequency

The successive cycles of slow opening and rapid closing of the vocal folds produce harmonics whose frequencies are integer multiples of the frequency of vibration of the vocal folds, called laryngeal frequency. Strictly speaking, a frequency can thus be associated with any segment of a voiced speech sound, assuming that the frequency value remains constant, which is, of course, never the case. In fact, the term *frequency* is a bit confusing and strictly corresponds to the inverse of the cycle time of vibration of the vocal folds, itself often called the laryngeal period (while laryngeal vibration is a quasi-periodic phenomenon rather than strictly periodic).

Fundamental frequency and melodic curve

The term *fundamental frequency* in speech is related to the measure of the laryngeal frequency derived from Fourier spectral analysis. This type of analysis decomposes successive small segments of the speech signal (seen through a "window" time of several tens of milliseconds) into their harmonic components, by using the Fourier theorem for harmonic analysis. These components have frequencies which are integer multiples of the inverse of the time window duration. A typical value of 30 ms window duration would, for example, give Fourier harmonic frequencies of 1/0.03 = 33.3 Hz, 66.6 Hz, 99.9 Hz, etc. The

longer the time window, the finer the frequency analysis resolution, at the expense of a lower time resolution due to the use of longer windows. A 1 second window would give an excellent frequency resolution of 1 Hz, but would be unsuitable for speech as many events may occur in a single second of speech. The commonly retained value of 30 ms results in a sufficient frequency resolution to "capture" the fundamental frequency of voiced speech sounds by interpolation. This value compares with the number of frames per second commonly used in movies to capture body movements (typically 24, 25, or 30 frames per second).

The speech fundamental frequency, F0 (denoted F "zero"), not to be confused with the Fourier fundamental frequency, corresponds to the first harmonic component found in the Fourier analysis of the signal, but also, by definition, corresponds to the frequency difference between two consecutive speech harmonics. As this analysis needs a rather long time window to be effective, the actual value of the laryngeal period may fluctuate during the time window needed for the analysis. By moving the analysis time window along the time axis, values for each successive position of the time window are obtained. These plotted values, whose ordinate corresponds to the fundamental frequency (vertical axis) and the abscissa the time (horizontal axis), form a melodic curve (Fig. 1.6).

It appears that the melodic curve has a much tormented shape with numerous ups and downs in frequency, and is also interrupted in some places. These interruptions correspond to the absence of fundamental frequency value, due in turn to the absence of voicing (unvoiced speech sounds or silence), at least if the measure is reliable, which is not always the case in adverse recording conditions (e.g. low signal to noise ratio). We observe, for example, a rather

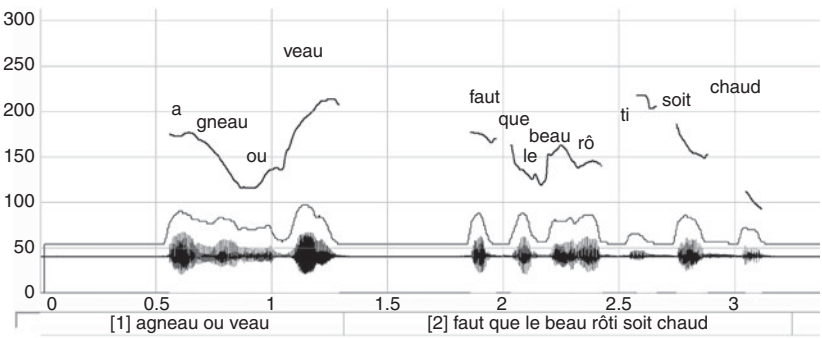

Figure 1.6 An example of melodic curve, interrupted at segments without voicing (including pauses and silence), with the fundamental frequency (top), intensity (middle), and wave (bottom) curves.

large interruption of the melodic curve in Figure 1.6 corresponding to a silent pause (between the French words *veau* and *faut*) and another corresponding to the location of the voiceless consonant [k] in the word *que*, at a time equal to 2 seconds.

Intensity

Besides the melodic curve resulting from the successive values of F0 plotted along the time axis, it is also customary to display an intensity curve by measuring the intensity or the amplitude of each speech segment inside a time window. The unit of measurement is usually the decibel (dB), a logarithmic value relative to some arbitrary reference defined in the instruments or within the software used. The most commonly used value for the measurement is relative to the global intensity detected within the time window used in Fourier analysis.

A remarkable intensity value corresponds to the increase in decibels resulting from doubling the amplitude of a pure tone: 10 log (2) = 3 dB (exactly 3.0102999... dB) for amplitude and 20 log (2) = 6 dB intensity. Halving the amplitude causes −3 dB amplitude and −6 dB of intensity drop. The multiplication of the amplitude by a factor of 10 corresponds to an increase in intensity of 20 log (10) = 20 dB, by a factor 100 of 40 dB, etc.

The dB unit is always a relative value. When the threshold of hearing is used for reference, the values are absolute decibels (0 dB SPL in English notations, where SPL stands for Sound Pressure Level) and decibels relative when the reference is different from this threshold. Absolute dB are thus dB relative to the threshold of audibility at 1000 Hz.

Since it is sufficient to increase or decrease the amplitude level of sound reproduction equipment in order to change the intensity curve span up or down, only relative intensity measures make sense, for example by comparing the values in dB of two consecutive vowels. Also it is not legitimate to average intensity values in dB, since this unit is logarithmic. Averages should be computed from the amplitude values, the formula to obtain an amplitude value from a dB value of a sound relative to a reference amplitude being A_{ref} is $I = 20 \log [a / A_{ref}]$.

Spectrographic analysis

The spectrogram is a three-dimensional graphical representation (time on the abscissa, frequency on the ordinate, and amplitude coded by colors or levels of gray) of the Fourier analysis of successive windows of the speech signal previously recorded. Depending on the length of the time window used, it can display harmonics (setting called narrowband) or high concentrations of

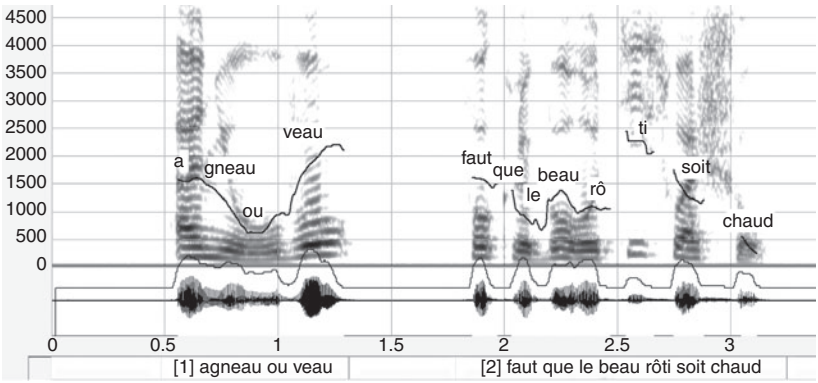

Figure 1.7 Narrowband spectrogram for visualizing harmonics corresponding to the fundamental frequency curve.

harmonic amplitude (setting called broadband) for voiced sounds. In cases of poor recording quality, a melodic curve superimposed on a narrowband spectrogram allows one to validate (or not) the reliability of the analysis, by comparing visually the evolution of the melodic curve with the harmonics: both must globally match as shown on Figure 1.7.

It should be noted that the same duration of time window is not necessarily appropriate for all voices in order to obtain a narrowband spectrogram. Harmonics will be well separated if the duration of the window is sufficient, and thus different values, e.g. 32 ms for a male voice at 130 Hz and 16 ms for a female voice to 250 Hz, may be suitable.

Syllabic duration

Syllable duration is also a prosodic parameter. The duration unit of measure is the millisecond (ms). Instrumental measurement may seem trivial, but in practice it is actually complex to do manually or automatically. Indeed, to determine syllabic segment boundaries, even by an expert versed in the joys of visual inspection of spectrograms, is far from simple and cannot be automated easily. The main reason is that the problem is ill posed, since the consonants and vowels result from continuous changes of the speaker articulatory configuration, as is the case when we walk, where moments of beginning and end gestures are not precisely defined. Likewise, the starting and ending instants of a speech event should be evaluated from the time they are perceived and the time they cease to be perceived, and these moments are not necessarily identical for a listener and for an acoustic speech analyzer.