Chapter 2
Speech Production and Modelling

2.1 Introduction

The human process of speech production involves three major levels of processing. The first is a high-level conceptualisation of the *intention* to speak, including forming a message. The second level is where the abstract message is transformed into a *linguistic* form. Finally, the third stage is *articulation* of the message, which involves mapping a sequence of *phonemes* to *phonations* and the motor coordination of lungs, glottis, larynx, tongue, lips, jaw and other parts of the speech production system.

Speech coding with code-excited linear prediction (CELP) operates on the acoustic signal with a rudimentary speech production model based on physiology. In this work, it is therefore appropriate to concentrate on the third level, articulation only. Application of models for the higher levels, linguistic content or even the abstract 'message' of speech, would be beneficial in many cases, as was discussed in Sect. 1.2. For example, we could apply a speech recogniser at the encoding stage to extract both linguistic content as well as paralinguistic content such as speaker identity, style, emotional and physiological state, etc. The decoder could then reproduce a speech signal which is equivalent to the original [21]. Unfortunately, such technology is currently unavailable, of insufficient quality or much too complex for practical implementation [17, 18, 21]. In this Chapter, we will therefore discuss articulation and modelling thereof on a physiological and acoustic level.

This Chapter is though only a review of necessary prerequisites for speech coding, not a comprehensive exposition of speech production. For more information, see [2, Chap. A2] or [15, 16].

2.2 Physiology and Articulation

In short, on a physiological level, speech production begins from the lungs which contract and push out air. This airflow can be used for two types of effects. Firstly, the airflow can set the *vocal folds* into an oscillation, periodically closing and opening,

T. Bäckström, *Speech Coding*, Signals and Communication Technology,
DOI 10.1007/978-3-319-50204-5_2

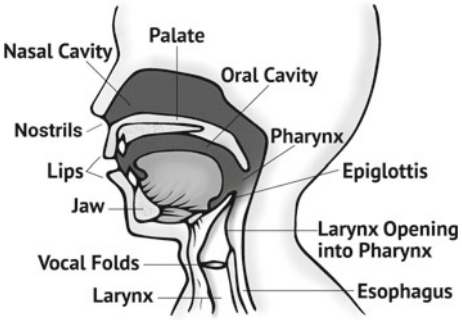

Fig. 2.1 Illustration of the human vocal apparatus used to produce speech, in a cross section of the head (© Copyright Mayra Marin, reproduced with permission)

such that the emitted airflow gains a (quasi)periodic waveform. Secondly, the airflow can cause noisy turbulences at constrictions of the *vocal tract*. The oscillating or noisy waveforms then flow through the vocal tract, whose resonances shape the spectrum of the acoustic signal. These three components, oscillating vocal folds, turbulent noise in constrictions and acoustic spectral shaping of the vocal tract, give the speech signal its defining characteristics.

Figure 2.1 illustrates the main parts of the vocal apparatus. The air flows through the *larynx* and the *glottis*, which is the orifice between the vocal folds. Airflow then proceeds through the *pharynx*, into the mouth between the *tongue* and *palate*, between the teeth and is finally emitted through the lips. A part of the air flows also through the *nasal cavities* and is emitted through the nostrils.

The most important excitation of the speech signal stems from oscillations of the vocal folds. Given the right conditions, such as airflow speed and stiffness of the vocal folds, the airflow from the lungs bring the vocal folds into an oscillation. Airflow pushes the vocal folds open and they gain momentum. As the vocal folds open, air rushes through the glottis whereby the pressure drops such that ultimately, the vocal folds are no more pushed out, but rather pulled back together due to the Bernoulli effect, until they clash into each other. As long as the airflow and stiffness of the folds are constant, this process will continue in a more or less periodic manner. One period of the glottal oscillation is illustrated in Fig. 2.2.

Speech sounds where the vocal folds are oscillating are *voiced* sounds and the process of uttering voiced sounds is known as *voicing*. This manner of articulation is called *sonorant*. Some examples of sonorant phonemes are; the vowels ‘a’ in the word ‘*a*lgebr*a*’; and the vowel ‘o’ as well as the nasal ‘n’ in the word ‘*no*’.

Figure 2.3 illustrates the vocal folds and the glottis in a view from above. Here the vocal folds are seen in their abducted, or open position, where air can freely flow through them.

Unvoiced speech excitations are produced by constricting or even stopping airflow in some part of the vocal tract, such as between the tongue and teeth, tongue and palate, between the lips or in the pharynx. This manner of articulation is thus known as *obstruent*, since airflow is obstructed. Observe that these constrictions can occur concurrently with a voiced excitation. However, speech sounds with only an unvoiced

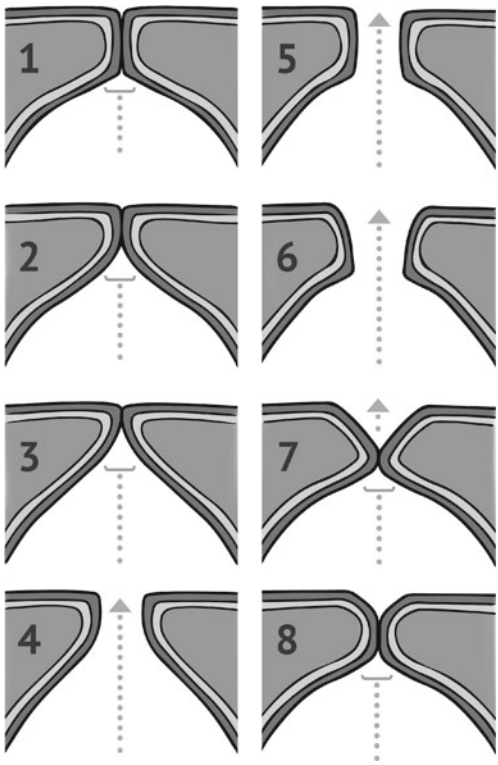

Fig. 2.2 Illustration of the cross-section of the vocal folds viewed in the anterior-posterior plane (front-back direction) during one period of the glottal oscillation (© Copyright Mayra Marin, reproduced with permission)

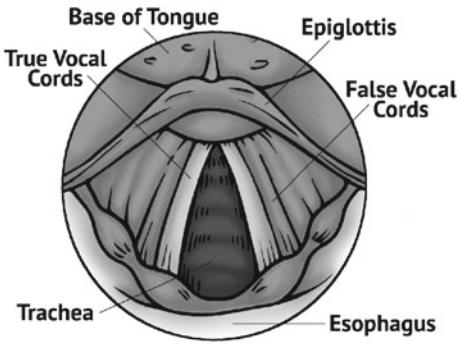

Fig. 2.3 A view on the glottis from above (© Copyright Mayra Marin, reproduced with permission)

excitation are known as *unvoiced* sounds. A constriction causes the airflow to go into a chaotic regime, which is effectively a turbulent mode. It is characterised by essentially random variations in airflow, which can be perceptually described as noise. Obstruent articulations where airflow is obstructed but not stopped are *fricatives*.

When airflow is temporarily stopped entirely to be subsequently released, it is known as a *stop*, and when the stop is released into a fricative, it is an *affricative*.

Table 2.1 Manners of articulation

Obstruent	Airflow is obstructed
Stop	Airflow is stopped and then released, also known as *plosives*
Affricative	Airflow is stopped and released into a fricative
Fricative	Continuous turbulent airflow through a constriction
Sonorant	Vocal folds are in oscillation
Nasal	Air is flowing through the nose
Flap/Tap	A single contraction where one articulator touches another, thus stopping airflow for a short moment
Approximant	Articulators approach each other, but not narrowly enough to create turbulence or a stop
Vowel	Air is flowing freely above the vocal folds
Trill	Consonants with oscillations in other parts than the vocal folds, such as the tongue in 'r'

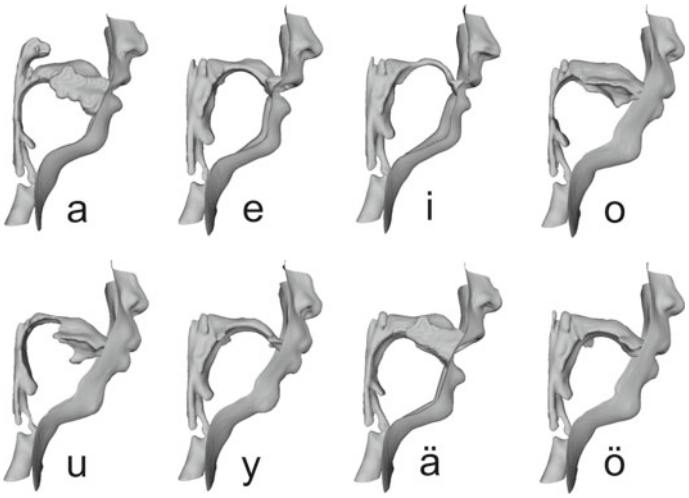

Fig. 2.4 Illustration of vocal tract shapes in 8 different vowels of Finnish (© Copyright Jarmo Malinen, reproduced with permission)

As examples of obstruents; the stop 'p' in the word '*p*op' corresponds to an unvoiced obstruent; the stop 'b' in the word '*b*us' corresponds to a voiced obstruent; and the fricative 'h' in the word '*h*ello' is also an unvoiced obstruent.

In summary, some of the main manners of articulation are listed in Table 2.1.

Finally, important characteristics of speech signals are defined by shaping the vocal tract. Figure 2.4 illustrates vocal tract shapes in the Finnish language. The

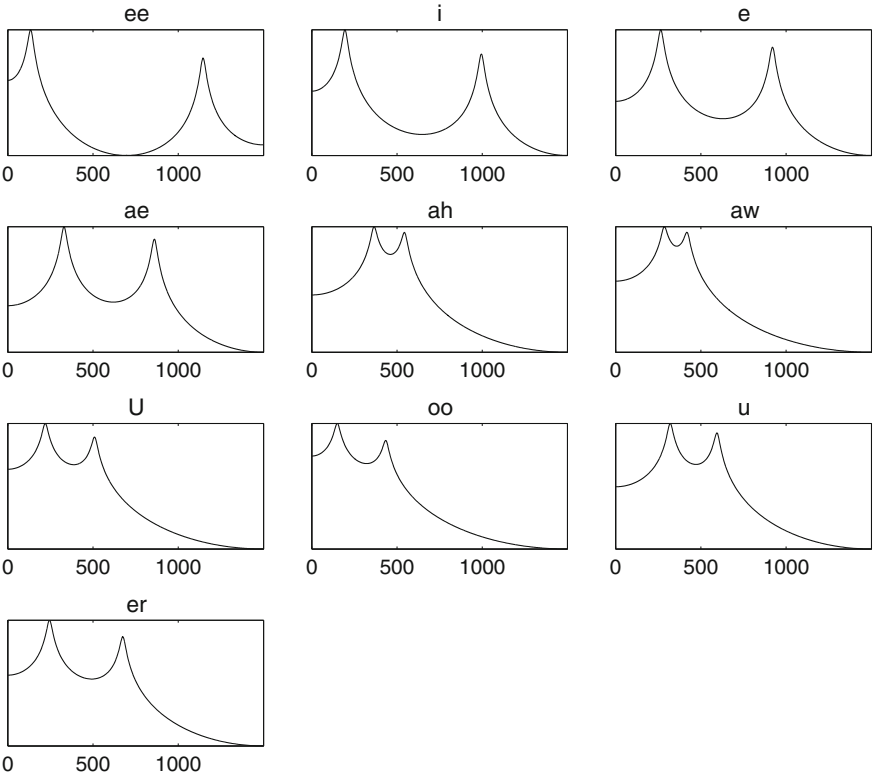

Fig. 2.5 Illustration of prototype spectral envelopes for American English vowels, showing the characteristic peaks of the first two formants, F1 and F2, averaged over 76 male American English speakers, depicted on a logarithmic magnitude scale. (Formant frequencies and magnitudes extracted from [19, p. 320])

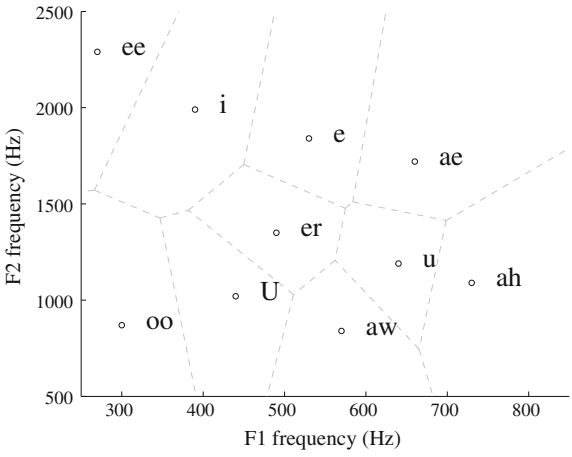

Fig. 2.6 The distribution of vowels with respect to the two first formants, F1 and F2, averaged over 76 male American English speakers. The *dashed lines* depict the approximate regions where phones would be classified to the corresponding phoneme (formant frequencies extracted from [19, p. 320])

different shapes give the tube distinct resonances, which make out the defining differences between vowels. The resonances are known as *formants* and numbered with increasing frequency, such that the first formant F1 is the resonance with the lowest frequency. Figure 2.5 illustrate the prototype shapes for English vowels. Here the characteristic peaks of formant frequencies are depicted, corresponding to the resonances of the vocal tract.

The two first formants, F1 and F2, are from a linguistic point of view the most important, since they uniquely identify or characterise the vowels. Figure 2.6 illustrates the distribution of English vowels on the axes of F1/F2 plane. We can here see that the formants are fairly evenly distributed on the two dimensional plane. It is well-known that vowels are identified mainly based on F1 and F2, and consequently, they have to be well separated in the two dimensional plane such that they can be easily identified. Conversely, would a language have vowels close to each other, they would most likely over time shift frequency such that they can be uniquely identified, as people attempt to pronounce clearly and avoid misunderstanding.

The importance of the two first formants is further demonstrated by the fact that we have well-known non-technical descriptions for vowel characteristics, which can be intuitively understood. Specifically, vowels can be described on the axes of closeness (closed vs. open) and backness (front vs. back). It should be noted, however, that even though closeness and backness have specific meanings in phonetics, the intuitive interpretations of these concepts are prone to misunderstandings. In this book, we will therefore only discuss formants, since their interpretation is unambiguous.

Observe that the formant frequencies F1 and F2 describing vowels are unique for each language. Still, since the vocal tract is shorter for females and children than adult males, the frequencies are shifted higher in comparison to male, while the closeness and backness remain relatively constant.

2.3 Phonemes

2.3.1 Vowels

As described before, vowels are sonorant phonations, that is, the vocal folds exhibit a periodic excitation and the spectrum is shaped by the vocal tract resonances, the formants. The two first formants define the vowels and their average locations are listed in Table 2.2. The third formant is less important from a linguistic perspective, but is essential to reproduce natural sounding vowels.

The table lists vowels[1] with their corresponding symbols in the international phonetic alphabet (IPA) as well as their symbols (or combination of symbols) using the speech assessment methods phonetic alphabet (SAMPA) set. In the following, we will use the backslash notation /·/ to denote IPA symbols, such as /t/ and /θ/.

[1] This is a representative list of vowels, but in no way complete. For example, diphthongs have been omitted, since for our purposes they can be modelled as a transition between two vowels.

Table 2.2 Formant locations of vowels identified by their International Phonetic Alphabet (IPA) symbol as well as the computer readable form SAMPA [23–25]

Vowel		Formant (Hz)			Examples
IPA	SAMPA	F1	F2	F3	
i	i	290	2300	3200	cit**y**, s**ee**, m**ea**t
y	y	280	2150	2400	*German:***ü**ber, R**ü**be
ɨ	1	290	2200	2500	ros**e**'s
ʉ	}	330	1500	2200	r**u**de
ɯ	M	330	750	2350	*Irish:* **ao**l
u	u	290	595	2390	thr**ou**gh, y**ou**, thr**ew**
I	I	360	2200	2830	s**i**t
Y	Y	400	1850	2250	*German:* f**ü**llt
ʊ	U	330	900	2300	p**u**t, h**oo**d
e	e	430	2150	2750	*German:* G**e**nom, M**e**than, B**ee**t
ø	2	460	1650	2100	*French:* p**eu**
ə	@	500	1500	2500	**a**bout, **a**rena
ɘ	@\	420	1950	2400	*Dutch:* **i**k
ɵ	8	520	1600	2200	*Australian English:* b**i**rd
ɤ	7	605	1650	2600	*German:* m**ü**ssen
o	o	400	750	2000	*German:* **O**fen, R**o**man
ɛ	E	580	1850	2400	b**e**d
œ	9	550	1600	2050	*German:* H**ö**lle, g**ö**ttlich
ɜ	3	560	1700	2400	b**i**rd
ɞ	3\	580	1450	2150	*Irish English:* b**u**t
ʌ	V	700	1350	2300	r**u**n, w**o**n, fl**oo**d
ɔ	O	540	830	2200	l**a**w, c**au**ght, **a**ll
æ	{	770	1800	2400	c**a**t, b**a**d
ɐ	6	690	1450	2300	*German:* od**er**
a	a	800	1600	2700	h**a**t
ɶ	&	570	1550	1800	*Swedish:* h**ö**rt
ɑ	A	780	1050	2150	f**a**ther
ɒ	Q	650	850	2000	n**o**t, l**o**ng, t**a**lk

Table 2.3 Table of consonants used in English (From http://en.wikipedia.org/wiki/Help:IPA_for_English http://en.wikipedia.org/wiki/X-SAMPA)

IPA	SAMPA	Examples
b	b	**b**uy, ca**b**
d	d	**d**ye, ca**d**, **d**o
ð	D	**th**y, brea**th**e, fa**th**er
ʤ	dZ	**g**iant, bad**g**e, **j**am
f	f	**ph**i, ca**ff**, **f**an
g	g	**g**uy, ba**g**
h	h	**h**igh, a**h**ead
j	j	**y**es, **y**acht
k	k	s**k**y, **c**ra**ck**
l	l	**l**ie, s**l**y, ga**l**
m	m	**m**y, s**m**ile, ca**m**
n	n	**n**igh, s**n**ide, ca**n**
ŋ	N	sa**ng**, si**nk**, si**ng**er
θ	T	**th**igh, ma**th**
p	p	**p**ie, s**p**y, ca**p**
r	r	**r**ye, t**r**y, ve**r**y (trill)
ɹ	r\	**r**ye, t**r**y, ve**r**y (approximant)
s	s	**s**igh, ma**ss**
ʃ	S	**sh**y, ca**sh**, emo**t**ion
t	t	**t**ie, s**t**y, ca**t**, a**t**om
tʃ	tS	**Ch**ina, ca**tch**
v	v	**v**ie, ha**v**e
w	w	**w**ye, s**w**ine
z	z	**z**oo, ha**s**
ʒ	z	equa**t**ion, plea**s**ure, vi**s**ion, bei**g**e

2.3.2 Consonants

Roughly speaking, all phonemes which are not vowels are consonants. In the following, we will present the most important consonant groups, which correspond to the manners of articulation presented in Table 2.1. Examples of consonants are listed in Table 2.3.

Stops In stops, the airflow through the vocal tract is completely stopped by a constriction and subsequently released. A stop thus always has two parts, a part where air is not flowing and no sound is thus emitted, and a release, where a burst of air causes a noisy excitation. In addition, stops are frequently combined with a subse-

quent phoneme, whereby the transition to the next phoneme begins practically from the start of the burst.

Examples of stops are /t/ in '**t**ie' and /p/ in 'ca**p**'

Fricatives and Affricatives Fricatives are consonants where airflow is partly obstructed to cause a turbulent noise, shaped by the vocal tract. Affricatives begin as a stop but releasing into a fricative.

An example of a fricative is the /ʃ/ in '**s**hy' and an example of an affricative is the /tʃ/ in 'ca**tch**'.

Nasals, Laterals and Approximants In this group of consonants airflow is partly blocked or diverted through an unusual route.

In nasals, air flows through the nose (at least partly) instead of the mouth, such as the /n/ in 'ca**n**'.

In laterals, the tongue blocks airflow in the middle of the mouth but air can proceeds on the sides of the tongue. A typical lateral is /l/ in '**l**ie'.

In approximants, airflow is restricted but not quite enough to create a turbulent airflow. Examples of approximants are the /w/ and /r/ in the word '**wa**r', where /r/ is pronounced without the trill.

Trills Trills such as a rolling /r/ are characterised by an oscillation of some other part than the vocal folds, most commonly of the tongue, but also possible with the lips. Note, however, that in most accents of English, trills are not used but approximants are used instead.

Examples of trills are the /r/ in the German word '**R**athaus' or the English word '**r**eally' with a Scottish accent.

2.4 Intonation, Rhythm and Intensity

The linguistic content of speech is practically always supported by variations in intonation, intensity and speaking rhythm. Here intonation refers to the time contour of the fundamental frequency, rhythm to the rate at which new phonemes are uttered and intensity to the perceived loudness of the speech signal (closely related to the energy of the signal). By varying the three factors, we can communicate a variety of paralinguistic messages such as emphasis, emotion and physical state.

For example, the most important word of a sentence (or other segment of text) is pronounced in most languages with a high pitch and intensity as well as a slower speed. This makes the important word or syllable *really* **stand** `out` from its background, thus ensuring that the important part is perceived correctly.

Emotions are, similarly, to a large part communicated by variations in these three parameters. The reader can surely imagine the speaking style which communicates anxiousness (rapid variations in the fundamental frequency F_0, high speed and intensity), boredom (small variations in F_0, low speed and intensity), sadness, excitement, etc.

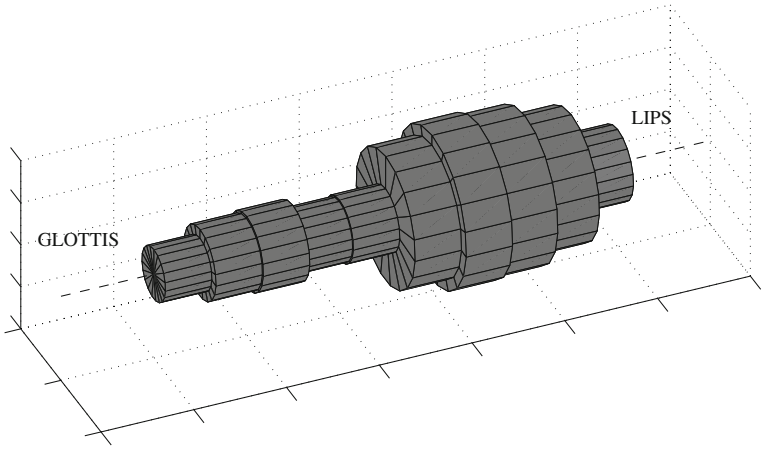

Fig. 2.7 Illustration of the tube model of speech production

Sometimes especially intonation also plays a linguistic role. For example, a sentence with a question is, depending on language, often finished with a rapidly rising pitch, while a statement has a constant or sinking pitch. Thus the main difference between "Happy?" and "Happy!" is the pitch contour. Moreover, some languages use pitch contours to distinguish words. Such languages are known as tonal languages and they are especially common in Asia.

2.5 Vocal Tract Models

In Figs. 2.1 and 2.4 we can see that the vocal tract is a twisted and variable-shaped tube, composed of tissue with a wide range of properties (tongue is soft, but the teeth are hard). A common assumption is that this complex-shaped tube can be modelled with a straight, round, lossless and piece-wise constant radius tube, without significant reduction in modelling capacity. In essence, the assumption is that although a tube model is not necessarily physically accurate, it does provide sufficient generality to encompass the significant properties of the vocal tract. The typical tube model of speech production is illustrated in Fig. 2.7.

The reason that the tube model is so prevalent in speech processing is that it is simple yet effective. The constant-width parts of the tube can be modelled by simple delays, while at the discontinuities, a portion of the signal is reflected, while another portion continues to the next segment. Since the system is assumed lossless, the reflection coefficient depends solely on the ratio of the cross-section areas of two two tube segments. Concatenating successive tube segments results in a lattice-form filter structure, which can, equivalently, be expressed as a linear predictive filter, which is described in more detail in Chap. 4. Due to this representation as a simple

filter, the tube model is very efficiently handled with discrete-time signal processing methods [16]. This approach is generally attributed to Kelly and Lochbaum, who worked with one-dimensional models of the vocal tract in the 1960s [10, 13].

The average length of the female and male vocal tracts are 14.1 and 16.9 cm respectively [9]. The length of the tube model, on the other hand, is defined in terms of the delays in which sound propagates through tube segments. If we assume an air temperature of 35 °C, then the speed of sound is approximately $c = 350\,\mathrm{m/s}$ [19, pages 40–41]. The relationship between vocal tract length L, linear predictor length M and sampling frequency f_s is [13, p. 75]

$$f_s = \frac{Mc}{2L} \tag{2.1}$$

whereby the linear predictor length is

$$M = \frac{2 f_s L}{c}. \tag{2.2}$$

In addition, to compensate for modelling inaccuracies (inaccurate modelling of the vocal tract and omission of a glottal flow waveform model, lip radiation model and nasal cavity model), generally, a small integer is added to M. Commonly used filter lengths are thus for example $M = 10$ for $f_s = 8\,\mathrm{kHz}$ and $M = 16$ for $f_s = 12.8\,\mathrm{kHz}$. For higher sampling rates the best model order varies depending on the scenario and systems design.

Here, a very important warning has to be included in the discussion of tube models (see also page 49). Linear prediction corresponds to a tube model and as such, it can be used for synthesis applications (for derivation, see [22]). The reverse is not accurate, though. Namely, analysing a tube model from a speech signal, even in a best case scenario, is difficult if not impossible. Since this book is concerned with speech coding, which practically always employs both analysis and synthesis, linear prediction does in practice not correspond to a tube model and calling linear prediction a tube model in the context of speech coding is thus inaccurate. We should thus treat linear prediction as a composite model, which includes the vocal tract, but also other elements of the speech production system.

Instead of linear prediction, it is naturally possible to construct models which more accurately describe the physical system. For example, it is possible to use finite element methods (FEM) to model the vocal tract (such as those in Fig. 2.4 [14]), or digital wave-guides [20]. Such models are, however, of much higher complexity than the linear predictive model. Moreover, since it has been found that linear prediction is (perceptually) effective in practice, these more advanced models have not found their way to practical technologies.

2.6 Models of the Glottal Excitation

As explained above, the glottal excitation is a semi-periodic oscillation of the vocal folds, where puffs of air come through the glottis (see Fig. 2.2). Models for the behaviour of this airflow waveform are known as glottal excitation models.

For accurate physical modelling of the glottal excitation, it is essential to model the movement of the vocal folds. Reasonably accurate results can be obtained by lumped-mass models [6], where the vocal folds are assumed to consist of a small number of lumped-mass blocks, joined together by damped springs, pushed open by the airflow and clashing into the opposite side vocal folds. The glottal airflow is then modelled as a function of the opening area between the vocal folds (see Fig. 2.8).

Lumped-mass models are relatively efficient from a speech synthesis point of view, since reasonably realistic glottal airflows can be achieved with a small number of parameters. From a speech analysis perspective, they are less useful, since it is difficult to estimate the parameters of the models from acoustic signals. The difficulties include problems such as that the lumped-mass models usually involve non-linear functions, the clashing together of the vocal folds is seen as a discontinuity, most of the parameters are not directly observable and it is not clear which parameter configurations can bring the model to the desired oscillating mode. All of these features imply difficulties for estimation of the parameters.

From a speech coding perspective, it is, however, not essential to obtain the actual physical parameters, but only a model which is able to represent the distinctive features of the glottal airflow. From this perspective, we note that it is possible to obtain reasonably accurate glottal flow models by linear predictive modelling.

To understand how linear prediction can be used for glottal modelling, note that the glottal system has two types of forces displacing the vocal folds from their rest position. Namely, firstly, airflow from the lungs exerts a pressure on the closed vocal folds pushing them open. This force is relatively slow in the sense that it acts on the vocal folds to some extent during the whole open phase (that part of the period where the vocal folds are open). In contrast, secondly, the movement of the vocal folds is abruptly stopped when they clash together. This event can be instantaneous, if the vocal folds touch each other on their whole length simultaneously, or almost

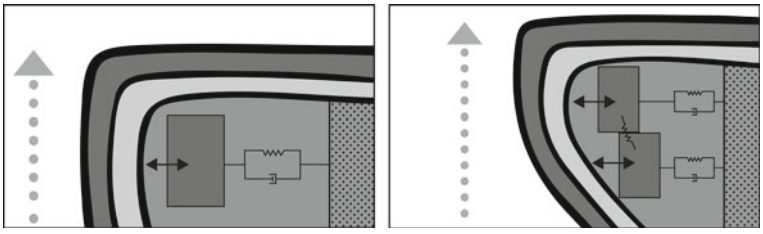

Fig. 2.8 Lumped-mass models of the glottal folds with one or two masses respectively; zoomed into Fig. 2.2. Here only the *right*-hand part of glottal folds is depicted, since the *left*-hand folds are symmetrically equal

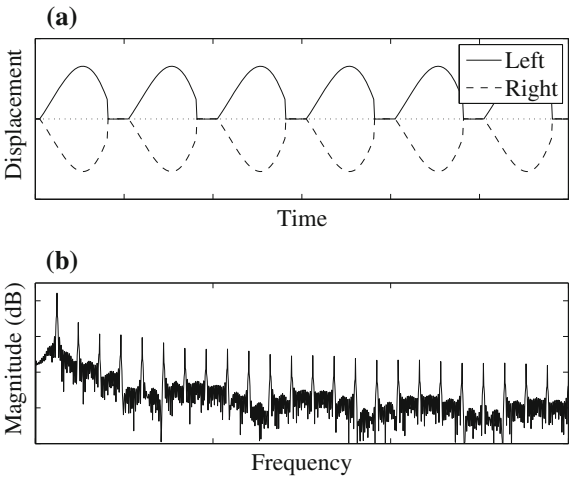

Fig. 2.9 **a** Schematic illustration of the movement of the vocal folds in terms of displacement from the *centre line*. Observe how the *left* and *right* folds clash together in the *middle*, thus causing the distinctive half-wave rectified shape of the glottal airflow. **b** The spectrum of the glottal airflow waveform, displaying the characteristic comb structure

instantaneous, if the vocal folds touch each other with a zipper-like movement. In either case, the closure of the vocal folds is only a short portion of the whole oscillation cycle. This movement is illustrated in Fig. 2.9.

The glottal cycle has thus one significant event per cycle, the glottal closure instant. Other forces acting on the system, like air pressure from the lungs, are more uniformly distributed over time, whereby the glottal excitation remains the only distinct event per cycle. For linear predictive modelling, we could use that excitation as the starting point, like the impulse at the start of an impulse response. However, the glottal excitation appears at the end of the glottal cycle. This excitation stops motion, whereas with linear prediction we use an impulse to start the impulse response. It then appears that the glottal excitation appears backwards in the system – we must reverse time to obtain a system where the glottal excitation corresponds to the starting point of a cycle.

To solve this discrepancy, we can simply use a backward or *anti-causal* linear predictive model. The glottal flow waveform is then simply the impulse response of a causal linear predictive model turned backwards in time. To obtain the periodic structure of the glottal waveform, this anti-causal linear predictor must then be excited by a periodic impulse train. A simple illustration of the output of an anti-causal linear predictor as a model of the glottal waveform is presented in Fig. 2.9.

Strictly speaking, also the opening force could be modelled as an excitation, but since it usually is not impulsive (as discussed above), it has a temporal structure, which will be at least partly modelled by the linear predictive model. Still, the impulse like residual(s) are the excitation of the glottal system, and these additional peaks within a period are often required to obtain an accurate model of the overall system.

Another issue is the fact that the sub-glottal part of the tube (parts of the tube which are below the vocal folds, such as the larynx and lungs), also has its acoustic resonances and that these can influence the acoustics of the system during the open

phase of the glottal period, but not during the closed phase [12]. In addition, the sub-glottal resonances can also interact with the vocal folds and modify their oscillation [1]. The sub-glottal resonances thus may influence the system in a discontinuous and non-linear manner and modelling such influences will be complicated. It can be argued that partly, the effect of sub-glottal resonances is not large and secondly, where they do influence the whole system, they are already included in the linear predictive model(s).

In any case, from a speech coding perspective, modelling the glottal flow by a linear predictor, excited by impulses, is in general quite sufficient. In fact, since the vocal tract can thus be approximated with a minimum-phase predictor and the glottal excitation with a maximum-phase predictor, these features can be employed to achieve an approximate separation between the glottal and vocal tract contributions to the acoustic signal [3–5].

2.7 Obstruent Modelling

Obstruents consist of two main categories, the fricatives and the stops, as well as their combinations, the affricatives. Recall that fricatives are formed by noise-like turbulences in constrictions of the vocal tract. If the noise really is noise-like, that is, it is a temporally uncorrelated excitation, then its modelling is relatively easy. However, it is not immediately clear whether turbulence noise is temporally uncorrelated in a statistical sense. It might have, for whatever reason, a temporal structure, or equivalently, a spectral envelope structure. Then it would be well-warranted to model it as (hardly surprisingly) a linear predictive model excited by white noise.

A second issue is that while the glottal excitation always happens at the bottom of the vocal tract and at the end of the tube model (if we disregard sub-glottal resonances), the obstruent excitations can occur in almost any part of the vocal tract. The insertion point of the noise excitation does bear an effect on the response of the entire system, such that it is not clear whether we can assume that the system is a linear predictor excited by white noise. At a minimum, the phase response of the system will be affected. However, again, we can assume that any deviation from the tube-model-excited-by-white-noise paradigm, can be modelled by another linear predictive filter.

Stops add one more dimension to noise modelling, that is, time. Stops are characterised by a combination of stop and release. Stopping the airflow is from an acoustic point of view a non-event, since it basically does not generate a new sound (although some transient effects in the vocal tract probably happen during a stop). The main acoustic event of a stop is thus paradoxically its release, when air suddenly flows through the obstruction. Clearly, this is an impulsive event, often followed by some turbulent noise, even if the obstruent would not be classified as an affricative. In other words, the excitation is a distinctive impulse followed by some level of noise especially for affricatives.

By now, it should be clear that the release might again (and it probably does) have a temporal structure, or in other words, it can be modelled by a linear predictor excited by an impulse (and trailing noise). In difference to excitation of the glottal system, impulses related to stops are usually isolated. Even trills have so low fundamental frequencies that their excitations can be treated as isolated events. Still, at onsets of vowels, the first glottal excitation of a voicing can be seen as equivalent to the release of an obstruent stop, as an isolated impulsive excitation.

2.8 Nasal Cavities

Phonemes and phones where the airflow partly or fully goes through the nasal cavities are known as *nasals*. For modelling speech production they introduce their own category of challenges. To model nasals, instead of modelling the vocal tract with just a tube model, we need a forked model, with a junction at the pharynx and additionally, optionally also for the two nostrils.

Since linear prediction is a one-dimensional digital wave-guide [20], a straightforward approach would be to apply a model where the wave has two parallel paths, both modelled with one-dimensional wave-guides. This approach would, however, again depart from the very convenient linear prediction model and introduce a significant increase in computational and modelling complexity. Fortunately, it has been found that inclusion of the nasal cavities can be implemented by a pole-zero model, that is, instead of the all-pole model of the conventional linear predictor, also include an FIR or zero-only part to the filter [1].

From a perceptual point of view, spectral zeros are not very exciting. The most important perceptual features are the high-energy regions of the spectrum, whereas low-energy areas are either masked or otherwise less important. Since the zeros of a filter produce spectral valleys, they are perceptually not very important and their approximate shape can also be modelled by an all-pole model. Therefore, even if pole-zero models would more accurately describe the physical characteristics of the system, they are usually replaced by all-pole models, because they capture the perceptually important characteristics in a simple and efficient way.

2.9 Lips

A very visible part of the speech production system are the lips and by contracting or opening them, we can change the acoustic signal significantly. After all, the perception of openness of vowels corresponds to a large part to the openness of the lips. However, the inside of the lips can be considered as a part of the vocal tract and they can be modelled by the tube model. The remaining part not included in the tube model is then the transition from the vocal tract into open air, the acoustic free field.

An important clue to the effect of lip radiation is the difference between the air behaviour in the vocal tract in comparison to what the ear can perceive. The excitation of the vocal tract is always generated by the airflow from the lungs, whereby the speech production model is characterised by a flow of air. The air flow of the vocal tract is however much too small in magnitude to be perceived as an air flow, a wind, in the free field. Acoustic signals on the other hand are characterised by their pressure waveforms. Although speech production cannot produce a wind, the transition from the lips to the free field does produce local variations in air pressure and hence, the speech sound is generated.

In very broad terms this transition can thus be seen as a transition from an airflow waveform to an air pressure waveform. A conversion from flow to pressure can be approximated by a gradient, which in turn can be approximated by the first difference, that is, the IIR filter $L^{-1}(z) = \frac{1}{1-z^{-1}}$. This is naturally not a strictly stable filter and it is not applicable in real systems. Moreover, the actual physical system is more complex, but in practice the lip radiation function can be modelled by [11]

$$L^{-1}(z) = \frac{1}{1-(1-\varepsilon)z^{-1}}, \tag{2.3}$$

where $\varepsilon > 0$ is a small number. Observe that again, we have modelled a part of the speech production system with a linear predictor and that the choice to do so was not a coincidence.

2.10 System Modelling

In the previous section we found that the main characteristic constituents of a speech sound are due to

1. the acoustic shaping of the vocal tract in the form of resonances or peaks in the macro-shape of the spectrum or the spectral envelope,
2. (sonorant) voicing or oscillation of the vocal folds producing a semi-periodic excitation for the vocal tract and
3. obstructions in the vocal tract producing noise-like excitations of the vocal tract.

Recall that the two excitations, sonorants and obstruents, can occur separately or simultaneously. It is therefore necessary to design a system which allows both excitations simultaneously, for example, by adding the two excitations together and using gain factors to determine their intensity and relative balance. The composite signal then excites the vocal tract, which shapes the spectral envelope of the joint excitation. We thus need to form a model which encompasses all these parts.

We will assume that the system is linear in the sense that it can be described by a cascade of linear filters as depicted in Fig. 2.10. This linear speech production model was introduced by Fant and corroborated by acoustic radiation experiments by Flanagan [7, 8].

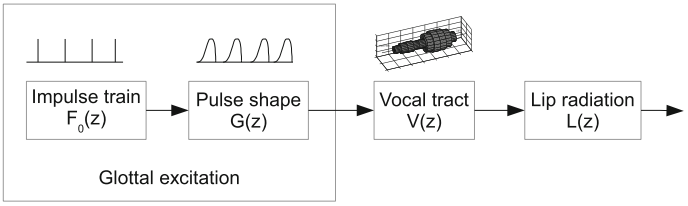

Fig. 2.10 A linear speech production model for voiced speech

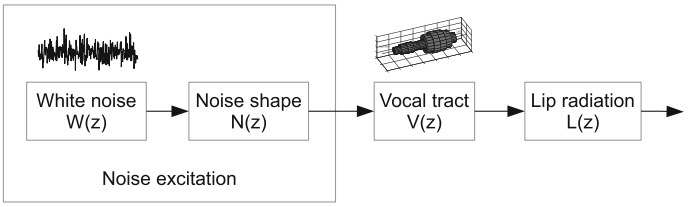

Fig. 2.11 A linear speech production model for unvoiced speech

Here, the vocal tract and lip radiation are represented by their transfer functions $V(z)$ and $L(z)$ and the glottal excitation is a combination of the impulse train corresponding to the fundamental frequency $F_0(z)$ and the wave form shape of a single glottal pulse $G(z)$. Due to linearity, the output signal is then

$$S(z) = F_0(z)G(z)V(z)L(z). \tag{2.4}$$

Similarly, we can excite the vocal tract with noise as in obstruent articulation, illustrated in Fig. 2.11. In this case, the excitation is white noise $X(z)$ possibly shaped by a filter $N(z)$, whereby the output is

$$S(z) = X(z)N(z)V(z)L(z). \tag{2.5}$$

The two excitations are assumed to be additive, $F_0(z)G(z) + X(z)N(z)$, whereby the total output is

$$S(z) = [F_0(z)G(z) + X(z)N(z)]\, V(z)L(z). \tag{2.6}$$

This approach is illustrated in Fig. 2.12.

Analytic operations on a model with two separate branches with linear prediction are complex, so we will simplify slightly the model to arrive at our final speech production model as

$$S(z) \approx [F_0(z) + X(z)]\, H(z) = [F_0(z) + X(z)]\, A^{-1}(z), \tag{2.7}$$

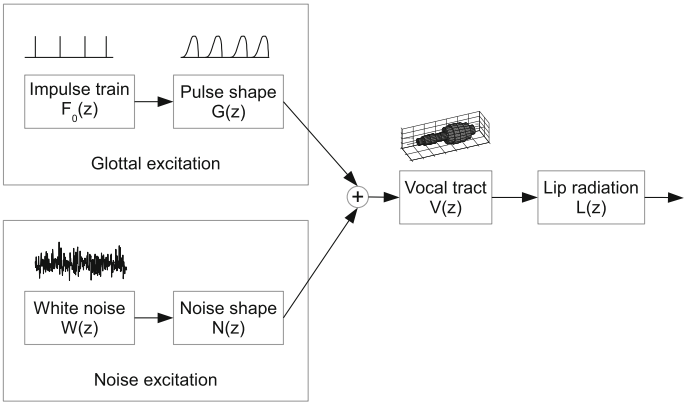

Fig. 2.12 Speech production model with combination of harmonic and noise input

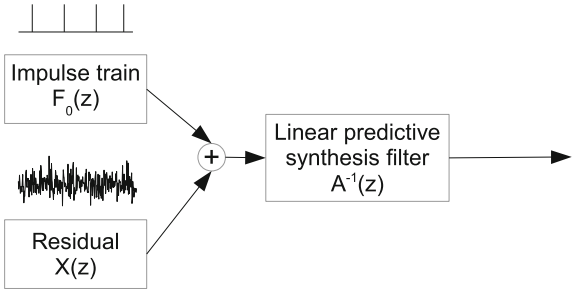

Fig. 2.13 Generic speech production model

where all filters $V(z)$, $L(z)$, $G(z)$ and $N(z)$ are merged into $H(z)$. Here we also assumed that $G(z) \approx N(z)$. Moreover, $H(z)$ is the impulse response of the linear predictor $A(z)$ and hence $H(z) = A^{-1}(z)$.

This is the 'standard' speech production model used in speech coding (and also often otherwise in speech processing) and it is illustrated in Fig. 2.13. Note that the difference between Eqs. 2.6 and 2.7 is that the former shapes the spectra of the two excitations separately, whereas the latter has equal shaping for both parts. If the noise model applied is sufficiently flexible, such that it has at least limited capability to model spectral shapes, then this assumption does not cause any noteworthy problems.

Note that it is by no means clear whether a given speech sound can be mapped to a unique constellation of the speech organs. Conversely, it is entirely possible that two or more different configurations of the speech production system could produce exactly the same speech signal. In terms of speech coding, the possibility of having multiple states which produce the same output represents over-coding and thus an inefficiency. Omitting one of the states which gives the same output would not change the quality of the system, but would reduce the states of the system and thus reduce the number of bits required to encode the signal.

However, when modelling the overall system with linear prediction, this is not a problem. The linear predictive model encodes the envelope shape of the spectrum, such that the residual is white noise. If we use a minimum mean square criterion, we will always find a unique envelope model, whereby the abovementioned danger of over-coding is removed.

In conclusion, we have found that modelling speech production by a linear predictor provides a well-warranted approximation of the underlying physiology. While it does not correspond one-to-one to a physical description, it does capture the most important features of the output signal. Moreover, the described generic speech production model happens to reproduce features of speech signals which are important for perception of speech. It is thus not surprising that linear prediction is a very effective and the most commonly used model in speech coding.

References

1. Austin, S.F., Titze, I.R.: The effect of subglottal resonance upon vocal fold vibration. J. Voice **11**(4), 391–402 (1997)
2. Benesty, J., Sondhi, M., Huang, Y.: Springer Handbook of Speech Processing. Springer, Heidelberg (2008)
3. Bozkurt, B., Doval, B., d'Alessandro, C., Dutoit, T.: Zeros of z-transform (zzt) decomposition of speech for source-tract separation. In: Proceedings International Conference Speech, Language Processing (2004)
4. Bozkurt, B., Dutoit, T.: Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In: ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (2003)
5. Degottex, G., Roebel, A., Rodet, X.: Phase minimization for glottal model estimation. IEEE Trans. Audio Speech Lang. Process. **19**(5), 1080–1090 (2011)
6. Erath, B.D., Zañartu, M., Stewart, K.C., Plesniak, M.W., Sommer, D.E., Peterson, S.D.: A review of lumped-element models of voiced speech. Speech Commun. **55**(5), 667–690 (2013)
7. Fant, G.: Acoustic Theory of Speech Production. Walter de Gruyter, Germany (1970)
8. Flanagan, J.L.: Speech Analysis: Synthesis and Perception. Springer-Verlag, New York (1972)
9. Goldstein, U.G.: An articulatory model for the vocal tracts of growing children. Ph.D. thesis, Massachusetts Institute of Technology (1980)
10. Kelly, J.L., Lochbaum, C.C.: Speech synthesis. In: Proceedings Fourth International Congress on Acoustics, vol. G42, pp. 1–4. Copenhagen, Denmark (1962)
11. Laine, U.K.: Modelling of lip radiation impedance in z-domain. In: Proceedings of the ICASSP, vol. 7, pp. 1992–1995. IEEE (1982)
12. Lulich, S.M.: Subglottal resonances and distinctive features. J. Phon. **38**(1), 20–32 (2010)
13. Markel, J.E., Gray, A.H.: Linear Prediction of Speech. Springer-Verlag, Inc., New York (1982)
14. Palo, J., Aalto, D., Aaltonen, O., Happonen, R.P., Malinen, J., Saunavaara, J., Vainio, M.: Articulating finnish vowels: results from MRI and sound data. Ling. Ural. **48**(3), 194–199 (2012)
15. Pulkki, V., Karjalainen, M.: Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics. Wiley, New Jersey (2015)
16. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signals, vol. 100. Prentice-Hall, Englewood Cliffs (1978)
17. Ramasubramanian, V.: Ultra low bit-rate speech coding: an overview and recent results. In: Signal Processing and Communications (SPCOM), 2012 International Conference on, pp. 1–5. IEEE (2012)

18. Ramasubramanian, V., Harish, D.: Ultra low bit-rate speech coding based on unit-selection with joint spectral-residual quantization: no transmission of any residual information. In: Proceedings of the Interspeech (2009)
19. Rossing, T.D.: The Science of Sound. Addison-Wesley, New York (1990)
20. Smith III, J.O.: Physical audio signal processing for virtual musical instruments and audio effects. In: Center for Computer Research in Music and Acoustics (CCRMA) (2013)
21. Tokuda, K., Masuko, T., Hiroi, J., Kobayashi, T., Kitamura, T.: A very low bit rate speech coder using hmm-based speech recognition/synthesis techniques. In: Proceedings of the ICASSP, vol. 2, pp. 609–612. IEEE (1998)
22. Vary, P., Martin, R.: Digital Speech Transmission: Enhancement, Coding and Error Concealment. Wiley, New Jersey (2006)
23. Wikipedia. Formant — Wikipedia, the free encyclopedia (2015). Accessed 1 Dec 2015
24. Wikipedia. International phonetic alphabet chart for English dialects — Wikipedia, the free encyclopedia (2015). Accessed 1 Dec 2015
25. Wikipedia. Table of vowels — Wikipedia, the free encyclopedia (2015). Accessed 1 Dec 2015

Springer

http://www.springer.com/978-3-319-50202-1

Speech Coding
with Code-Excited Linear Prediction
Bäckström, T.
2017, XXI, 240 p. 76 illus., Hardcover
ISBN: 978-3-319-50202-1