11. Formant Estimation and Tracking

D. O'Shaughnessy

This chapter deals with the estimation and tracking of the movements of the spectral resonances of human vocal tracts, also known as formants. The representation or modeling of speech in terms of formants is useful in several areas of speech processing: coding, recognition, synthesis, and enhancement, as formants efficiently describe essential aspects of speech using a very limited set of parameters. However, estimating formants is more difficult than simply searching for peaks in an amplitude spectrum, as the spectral peaks of vocal-tract output depend upon a variety for factors in complicated ways: vocal-tract shape, excitation, and periodicity. We describe in detail the formal task of formant tracking, and explore its successes and difficulties, as well as giving reasons for the various approaches.

11.1 Historical

In this chapter, we deal with a focused problem of speech analysis – trying to identify some very specific aspects of speech that have been found to be of great use in a wide variety of speech applications. These parameters of speech are called formants. They are generally viewed to be the resonances of the vocal tract (VT), which often appear in spectral displays (such as spectrograms) as regions of high energy, slowly varying in time as the vocal tract moves (Fig. 11.1).

Formants are useful in the coding, recognition, synthesis, and enhancement of speech, as they efficiently describe essential aspects of speech using a very limited set of parameters. For coding, if speech can be reduced to formant parameters to represent the VT shape (and a few other parameters to represent VT excitation), then very efficient coding is possible [e.g., 2.4 kbps linear predictive coding (LPC)]. Standard coders (e.g., in cellphone technology) use LPC with 10 or so coefficients to characterize the VT; it is feasible instead to represent the VT with even fewer parameters if formants are properly employed.

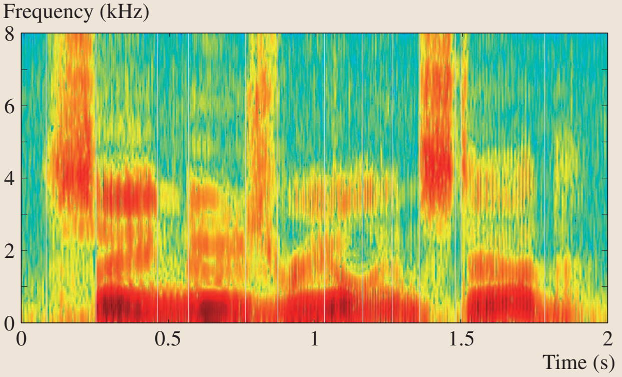

Fig. 11.1 Wideband spectrogram of an adult male speaker saying 'Say newsreel instead'. *Light vertical lines* (added manually) denote phoneme boundaries

Springer Handbook of Speech Processing
Benesty, Sondhi, Huang (Eds.) • © Springer 2008

In the 1970s, formants were the primary focus of automatic speech recognizers (ASRs), as it is necessary in ASR (as in coding) to greatly reduce the information present in a speech signal (e.g., 64 kbps for toll-quality speech in the basic telephone network), without sacrificing useful information about VT shape [11.1]. As many studies of human speech production and perception have identified formants as prime candidates to represent the VT spectrum efficiently, they were popular parameters to try to estimate from speech signals. It was found, however, that tracking formants reliably was not an easy task. Since the mid-1980s, ASR has primarily relied on the mel-frequency cepstral coefficients (MFCCs) instead of formant-based parameters to represent VT information. The advantage of the MFCC approach has been an automatic way to reduce the amount of information in a Fourier transform (FT) of a frame of speech (which is always assumed to reasonably capture the essential information about VT shape at any specific point in time; a frame is a short section of speech, e.g., 20 ms) to a small set of parameters, e.g., 10–16. The data reduction factor is about the same as for LPC, except that the MFCC is able to utilize some auditory factors in warping frequency scales to model the human ear better than LPC can (i. e., a mel or bark scale: linear up to 1 kHz, and logarithmic thereafter). Nonetheless, there still remains interest in formants for ASR purposes, as MFCC and LPC tend to suffer significantly when increasing amounts of noise are present in the received speech signal. Both MFCC and LPC take global approaches to speech analysis, which makes it difficult to separate noise from speech in corrupted signals.

The MFCC and LPC have been so popular for speech recognition and coding, respectively, because they are parameters that are obtained by simple mathematical rules (algorithmic transformations not subject to discrete, and hence nonlinear, decisions). Such transformations reduce the size of speech representations. A Fourier display does little data reduction as it occupies about as much data space as the untransformed speech. The MFCC and LPC are more compressed, but still leave room for further compression, as is possible with formants. *Features* such as formants must be estimated using error-prone methods, as they can only be obtained by applying (possibly faulty) decisions in the data reduction process. They achieve greater data reduction, at the cost of making errors. Given the widespread use of MFCC for ASR, some formant trackers employ the MFCC as the spectral input to their algorithms [11.2, 3], even though the MFCC tend to smooth spectra in ways that may obscure complicated formant structure. (Because ASRs using MFCCs rarely track formants, any formant tracking errors that might result from such an approach would not be pertinent for ASR, but it is nonetheless useful to see how well one can track formants using such a popular spectral estimation method.)

If instead we return to an original (possibly noisy) spectrum and properly examine the short-time Fourier transform (STFT) to find formants amid any interfering noise, there is still potential to carry out better ASR with formants [11.4]. Recently, formants have seen increasing use in text-to-speech (TTS) synthesis applications [11.5], as the trend has been to employ very large databases of small speech units (extracted from the speech of a single professional training speaker). In TTS, given a text to pronounce, a phonetic sequence is automatically determined, allowing access to these units. A major issue in recent research has been the efficient determination of which units to use, which requires searching a large space of alternatives (typically, through a Viterbi search) making many cost estimations about the spectral similarity between units. If these units are characterized via formants, these costs are often easier to compute than if other parameters are used.

Finally, aids for people with speech difficulties are often designed around the essential aspects or features of speech such as formants. In some speech aids, information additional to the normal speech input may be available to assist in the formant tracking task. Attaching a laryngeal sensor to a speaker's throat can provide detailed information about the VT excitation in a relatively unobtrusive fashion, and appears to assist in some formant tracking methods [11.6].

Certainly, a visual display capturing the mouth of the speaker corresponding to the input speech can help both ASR and formant tracking, as such imagery yields important information about (at least) mouth opening. As most speech applications do not have access to imagery, but only to the speech itself, we will not assume image assistance in this section. There are several formant trackers in commercial applications; a common one freeware package is Wavesurfer [11.7].

11.2 Vocal Tract Resonances

While formants are commonly viewed as peaks in the speech spectrum, we must be more rigorous in defining them, as spectral displays of speech vary greatly in the number and types of peaks seen (Fig. 11.2). Formants are usually understood to be broad spectral peaks in an STFT of speech, corresponding to the underlying vocal tract resonances (VTRs). Such basic resonances can often be calculated from VT area functions, if available, e.g., via X-rays or electromyography. Few speech applications have access to such data, however, and thus resonance estimations must usually be based on analysis of the speech coming from the mouth. However, such resonances are only present in output speech energy to the extent that the VT system is excited with sufficient energy at those frequencies. We must recall that speech output is the convolution of the VT excitation waveform and the impulse response of the VT (or, equivalently, the speech spectrum is the product of these two spectral representations), and thus the nature of the VT excitation is a major factor in whether VTRs form visible peaks in the output speech STFT.

As we are primarily interested in formant estimation during strong sonorants, we first look at VT excitation for vowels, where the excitation consists of glottal puffs of air modulated by the vocal folds with a lowpass nature (about −12 dB/octave fall-off). As a result, the first formant (F_1) normally has the highest intensity, and the amplitude decreases at about −6 dB/octave for formants at higher frequencies (−6 is the net result, after including the +6 dB/octave radiation effect at the lips). Spectrograms often show clear bands of formant energy for the first 3–5 formants in vowels, but this situation varies greatly depending on the spectrum of each vowel [e.g., /u/ has a greater fall-off, with little energy above F_2 (second formant), as its F_1 and F_2 are quite low in frequency, whereas /i/, with a much higher frequency F_2, has much more energy in the F_3–F_4 region]. The style of speech also has effects: breathy voice has a greater rate of fall-off, and hence weaker formants at high frequencies, than a shouting voice. Few speech applications have shown interest in the estimation of formants above F_4, as higher formants are quite weak (especially relative to any background noise) and have much less perceptual relevance than F_1–F_4. Indeed, estimations of F_3 and F_4 are often difficult owing to the low energy in their frequency ranges (e.g., for back vowels such as /u/).

Issues of weak formants are not limited to high frequencies. Nonvowel phonemes, such as nasals and obstruents, partly owing to the presence of zeros in their spectra, have varied spectral displays in terms of which resonances are visible. In the case of obstruents, a major additional factor is the different nature and location of the noise excitation, which is much higher in the VT, hence exciting only higher resonances, as will be

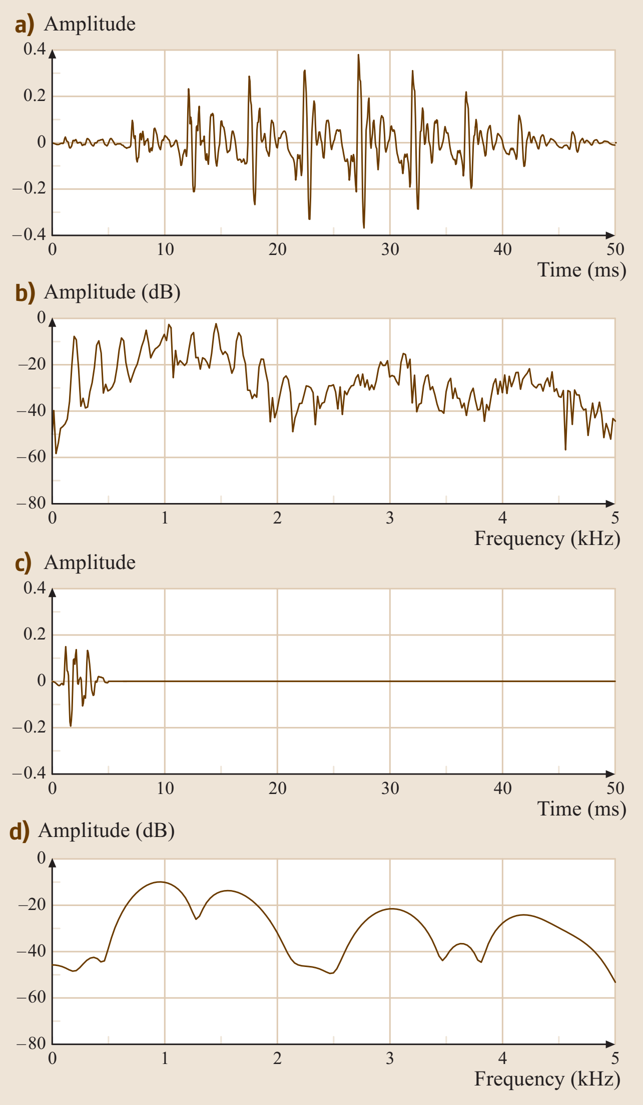

Fig. 11.2a–d Time signals and spectra of a vowel. **(a)** Speech signal weighted by a 50 ms Hamming window, **(b)** the corresponding spectrum, **(c)** speech signal weighted by a 5 ms Hamming window, and **(d)** the corresponding spectrum

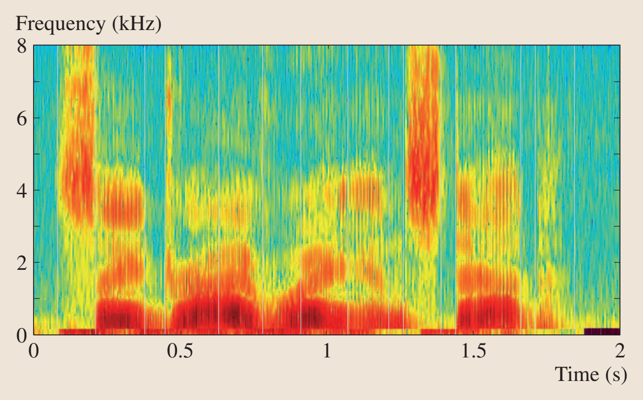

Fig. 11.3 Wideband spectrogram of the same speaker saying 'Say driveway instead'

discussed later. Thus, the varying display of bands of energy that we wish to associate with VTRs is a major factor in the difficulty of formant tracking, and not just for a minority of speech sounds.

Part of the difficulty of formant estimation and tracking has been not properly exploiting much of what is known about human speech production and perception. This failing has occurred too often in the speech research field (e.g., for many years, ASR ignored the fact that speech had an underlying message component; it was only in the 1980s that language models were commonly used for ASR). A major difficulty of tracking formants is their tendency in spectral displays to weaken, disappear, and/or merge (Figs. 11.1–11.3), as the VT undergoes changes in shape or as the VT excitation changes. As speech is inherently dynamic (e.g., 12 phonemes/second on average), the ability to track temporal changes in any parametric representation (e.g., formants) is essential. While such changes (e.g., abrupt appearance or disappearance of formants, as well as formants approaching each other so closely as to appear as one band of energy in time displays) are readily apparent, the underlying resonances of the VT are much more well behaved, albeit harder to observe. VTRs derive directly from VT shape, and are the poles of the VT transfer function. As the VT moves smoothly from one position to another, the VTRs smoothly change values [11.8]. Even when occlusions are made or released (as in stop consonants), the positions of the VTRs change smoothly. As such, VTRs would be much easier to track than formants (whose appearance on spectrograms is more varied). However, lacking visual displays of the VT (e.g., a camera focused on the speaker's mouth, or a radiographic image of the inside of the mouth), which is normally the case in almost all speech applications, we must rely on formant rather than VTR tracking. We can, nonetheless, exploit presumed knowledge about VTRs when estimating formants.

11.3 Speech Production

To better understand the foundation for formants, let us examine some aspects of human speech production. In any estimation task, it is often useful to examine important properties of the data source, rather than attack a problem with few assumptions. Too often in the past, research on formant estimation simply viewed STFTs, looked for bands of energy, and applied basic peak-picking approaches (with some constraints on continuity), without further regard for the nature or origin of the speech signal. It is true that the essential information about formants (for most speech applications) is readily available in such displays, but such an approach is like doing ASR without a language model. An essential step of any pattern recognition task, whether complete ASR or simpler formant tracking, is to reduce the input data while minimizing loss of useful information; this *preprocessing* step not only makes the process more efficient (as subsequent recognition steps treat smaller amounts of data), but also allows more-accurate estimation decisions, by focussing more closely on aspects of the data that are crucial (for the task at hand). In many cases of pattern recognition, an underlying source undergoes several (often nonlinear) mappings before a data acquisition device can capture a readily observable signal. A better understanding of these mappings can aid the parameter estimation process. In our case, the basic underlying source is a message in a speaker's brain. ASR attempts to recover that message (e.g., a text), given the observed speech signal. More specifically here we are concerned with identifying certain intermediate parameters – the formants – in this process. In any event, better knowledge about where formants originate is useful.

In human speech production, the lungs (as a pressure source) push air past the vocal folds, which often modulate the airflow to create the excitation sound for the VT, which in turn acts as a filter to amplify certain frequencies while attenuating others. The VT is the most important component in speech production, as it controls the formants and its variation leads directly to the perception (by listeners) of the individual phonemes (sounds) of speech. A tubular passageway composed of muscular and bony tissues, the VT, via its shape, modifies the spectral distribution of energy in glottal sound waves. Different sounds of speech are primarily distinguished by their periodicity (voiced or unvoiced), spectral shape (which frequencies have the most energy), and duration. The vocal folds specify the voicing feature, but the major partitioning of speech into sounds is accomplished by the VT via spectral filtering. The VT is often modeled as an acoustic tube with resonances called formants, but it can have antiresonances as well. In z-transform terminology, each resonance corresponds to a complex conjugate pair of poles, whose angular position corresponds to the center frequency of the resonance (usually, we refer to this as the formant value) and whose radial position corresponds to the resonance bandwidth. (In formant tracking, the amplitudes of the resonance peaks are usually of considerably less interest than the center frequencies and bandwidths, as amplitudes are greatly influenced by the spectral tilt of the glottal excitation, as well as by the proximity of nearby formants; listeners seem to take account of this by utilizing rather little relative peak amplitude information in speech perception.) The formants are often abbreviated as F_i, e.g., F_1 is the formant with the lowest frequency. (F_0 is associated with the fundamental frequency of the vocal folds during voiced speech, and is not a resonance. F_0 is quite visible in narrowband spectrograms, as speech consists of F_0 and its many harmonics, modulated by the VT transfer function, i. e., the spectral envelope.)

Figure 11.2 shows displays of the amplitude of an STFT for a typical vowel. If, as is usual, we use a time window (i. e., the length of the STFT) of about 20–25 ms, perhaps shaped by a Hamming window to reduce edge effects [11.1], the display will show a large amount of detail, corresponding to many aspects of the windowed speech signal. (Typical FFT lengths are 256, 512, or 1024 samples, which correspond to common frame durations at sampling rates around 16 kHz; as a result, sample points in the STFT are about 30 Hz apart, which means that each harmonic has a few values, and the STFT is highly variable in frequency.) Most notable are the harmonics, peaks spaced every F_0 Hz (Fig. 11.2b), which correspond to multiples of the vocal fold vibration rate. (These equally spaced peaks have finite width, corresponding to the inverse of the window duration; if one were to theoretically reiterate a vowel's pitch period indefinitely, we would see an actual line spectrum.) Superimposed on this base periodic set of spectral lines is the VT transfer function or spectral envelope, as the STFT is the product of the excitation spectrum (F_0 lines) and the VT filter response. It is this spectral envelope, and more particularly its formant peaks, that interest us the most in this Section. Note, however, that the formants are not immediately accessible in the FT, as the spectral envelope is essentially only specified at multiples of F_0. Typically, one may smooth the STFT to render the formants more evident.

Another way to suppress the spectral effects of the (interfering) harmonics is to use an STFT with a shorter time window (i. e., one shorter than the pitch period) (Fig. 11.2d). A disadvantage of this approach is that the shorter window needs to be placed synchronously with the pitch period for optimal formant estimation, and any corrupting noise in the speech signal has greater negative effects on formant estimation, as a much shorter set of speech data is used (owing to the shorter window).

Antiresonances correspond to zeros in the output speech, and occur owing to aspects of glottal excitation or to the existence of side-branches in the airflow path in the VT. This section examines formants, which correspond to the resonances. As such, we will not explicitly try to track the zeros, which are considerably more difficult to locate reliably (very recently, particle filtering has been used to track the poles and zeros of the VT [11.9]). Luckily, zeros are much less relevant for most speech applications.

Figures 11.1 and 11.3 show typical wideband spectrograms. Many speech applications make reference to such spectrograms, although few actually include these three-dimensional displays explicitly in their speech analysis (e.g., LPC and the MFCC are much more commonly used in coding and in ASR, respectively). They clearly show the formant bands of energy (where darkness indicates intensity) as a function of both time and frequency. In wideband spectrograms, use of a short time window yields much better time resolution (e.g., about 3 ms, making the vocal fold closure excitations readily visible as increases in energy every pitch period). As a result, frequency is typically smoothed automatically over a range of 300 Hz (roughly the inverse of 3 ms, and chosen to exceed most people's F_0, and hence include two or more harmonics within the window's low-pass filter range of smoothing). In such a display, the time signal is covered roughly continuously, thus obviating the need for pitch-synchronous analysis; a frame-based analysis method, as in ASR applications, would, on the other hand, need to address that issue.

11.4 Acoustics of the Vocal Tract

Speech is produced when air passes through the VT, which can be modeled as an acoustic tube of variable cross-sectional area $A(x,t)$, approximately closed at the glottal end and roughly open at the lips [A varies in space ($x=0$ at the glottis and $x=L$ at the lips; L will be assumed to be 17 cm) and in time t]. VT length varies greatly among speakers (about 13 cm for women, and less for children); we choose 17 cm as the nominal value here as this is an average value for men, and yields simple values for the average positions of the formants: 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz, ... etc. (as we will see later). Glottal area is small relative to typical A values, although the glottal end of the VT is truly closed only during glottal stops and during the closed phases of voicing. Lip rounding or closure often narrows the acoustic tube at the lips. A tube (e.g., the VT) closed at one end and open at the other resembles an organ pipe and is called a quarter-wavelength resonator, as the frequencies at which the tube resonates are those where sound waves traveling up and down the tube reflect and coincide at the ends of the tube. VTRs can be heuristically computed using only the boundary conditions of the VT, along with the phase relationship between the pressure and volume velocity in the traveling sound waves. Formant frequencies match the boundary conditions for pressure P (relative to atmospheric pressure) and volume velocity U: a closed end of the tube makes $U=0$, whereas $P\approx 0$ at an open end. P is 90° out of phase with U, owing to the inductance and capacitance of the VT. Resonances occur at frequencies F_i, $i=1,2,3,\dots$, where $|U|$ is maximum at the open end of the VT and $|P|$ is maximum at the closed end. Such frequencies have wavelengths where the VT length l is an odd multiple of a quarter-wavelength; hence, at 500 Hz, 1500 Hz, 2500 Hz, ... etc..

A uniform VT is only a good model for a schwa vowel /ə/. With other sounds, $A(x)$ is a complicated function of space along the VT, and as a result, the formants move to other frequencies. Normally, the deviations are within a range of a few (or several) hundred Hz. Thus, for a 17 cm VT, F_1 is usually in the range 300–800 Hz; F_1 tends to be low when the VT is relatively closed (e.g., for most consonants, and for vowels with a raised tongue – high vowels), and high for low vowels. F_2 is usually in the range 700–2200 Hz; F_2 tends to be high when the tongue is relatively forward, and low when the tongue is more to the rear. In general, the range of F_2 is greater than for the other formants. F_3 is usually in the range 1800–2800 Hz; F_3 tends to be high when the tongue is relatively forward and high, and low when the tongue is retroflexed (as in /ɪ/). Higher formants are usually progressively weaker in intensity, and less relevant for most applications. (Synthesizers that use formants often fix F_4 near 3500 Hz and F_5 near 4500 Hz; variations in these numbers appear to have little useful perceptual effect.) Thus, formant estimators focus on tracking F_1–F_3.

Almost all languages employ the *cardinal* vowels of /i/, /a/, /u/, and these three often represent VT shapes that are far from a neutral schwa (uniform VT) shape, i. e., other vowels (if present for a given language) have more intermediate shapes, and therefore formant frequencies that are closer to the neutral values (500, 1500, 2500 Hz). Tracking formants is typically more difficult for more extreme VT shapes, as some formants often appear to merge in such cases. (Examples are seen in Figs. 11.1 and 11.3, where formants change intensity suddenly and make apparent jumps, usually at phoneme boundaries, but certainly not at all phoneme boundaries.) Thus, let us examine further how VT shape relates to such vowels. The vowel /a/ can be roughly modeled by a (lower) narrow tube (representing the pharynx) opening relatively abruptly into a wide (upper) tube (the oral cavity). Assuming a 17 cm VT, for simplicity suppose that each tube has a length of 8.5 cm; then, each tube would produce the same set of resonances, at odd multiples of 1 kHz (1000, 3000, 5000 Hz, ...). Each tube is a quarter-wavelength resonator, since its back end is relatively closed and its front end is relatively open (i. e., the two tubes are half-versions of the full schwa VT, and thus have formants at twice the original values). As with all 17 cm models of the VT, this two-tube model has the same number (one) of formants per kHz, on average, but each one is moved by 500 Hz.

11.4.1 Two-Tube Models for Vowels

At any boundary between sections of the VT, whenever the change in areas is sufficiently abrupt (as with /a/, near the velum), the acoustic coupling between cavities is small and the interaction between cavity resonances will be slight; each section then controls its own number of the overall set of formants. Due to some acoustic coupling, formants never approach each other by less than about 200 Hz; thus, e.g., F_1 and F_2 for /a/ are not both at 1000 Hz, but rather near these values: $F_1 = 900$ Hz, $F_2 = 1100$ Hz, $F_3 = 2900$ Hz, and $F_4 = 3100$ Hz. The reverse holds for /i/: a wide pharyngeal tube narrowing abruptly into the oral cavity tube. Theory would have $F_1 = 100$ Hz, $F_2 = 1900$ Hz, $F_3 = 2100$ Hz, and $F_4 = 3900$ Hz, but in practice, F_1 is closer to 280 Hz, and F_4 often approaches F_3 (making a group of 2–3 formants around 2 kHz). Actual observed values for real versions of /a/ and /i/ will show deviations from these model numbers due to modeling inaccuracies; nonetheless, these simple models give reasonably accurate results and are easy to interpret physically.

The third cardinal vowel /u/ has a more-complicated analysis: anytime that the lips are rounded (as in /u/), all formants lower in frequency. This can be roughly seen by considering a VT mostly closed at both ends; the VT then becomes a half-wavelength resonator, as its boundary conditions are the opposite of the normal open-mouth model (e.g., the volume velocity is minimized at both (closed) ends of the VT, whereas it is maximized at the mouth in the open-mouth version). The theoretical locations for such a VT model are $F_1 = 0$, $F_2 = 1000$ Hz, $F_3 = 2000$ Hz, etc. Again, in practice, F_1 rarely goes so low, and approaches 200 Hz as the lips are almost closed. Examining such an analysis, one may be tempted to conclude that formants do not deviate more than about 400 Hz from the nominal neutral values of 500, 1500, 2500 Hz. Empirical evidence, however, shows that F_2, in particular, has a rather wider range (and F_3 goes as low as 1900 Hz). Again, all values given here assume a typical man's VT; for people of other sizes, one generally scales frequency values by the average length of the VT, relative to 17 cm; as VTs are not linear versions of each other, there may be significant deviations from such simple modeling.

11.4.2 Three-Tube Models for Nasals and Fricatives

We cannot easily extend the simple results of the two-tube VT model to more-complicated VT shapes for other phonemes. Nonetheless, some approximate extensions are feasible, and help understanding of the task of formant estimation. Nasal consonants require analysis of a three-tube VT model, as the nasal cavity is involved when the velum is lowered for such consonants. The entire system can be modeled as three tubes (pharyngeal, nasal, and oral) joined at one point (the velum), with acoustic circuits for the three in parallel. The poles of such a model are specified by [11.1]: $1/Z_\mathrm{p} + 1/Z_\mathrm{m} + 1/Z_\mathrm{n} = 0$, where $Z_\mathrm{p} = -\mathrm{i}Z_{0\mathrm{p}}\cot(\beta l_\mathrm{p})$, $Z_\mathrm{m} = -\mathrm{i}Z_{0\mathrm{m}}\cot(\beta l_\mathrm{m})$, and $Z_\mathrm{n} = \mathrm{i}Z_{0\mathrm{n}}\tan(\beta l_\mathrm{n})$. Z_p, Z_m, and Z_n refer to the acoustic impedance seen in the velar region of the VT, in the direction of the pharynx, nasal passages, and mouth, respectively, while $Z_{0\mathrm{p}}$, $Z_{0\mathrm{m}}$, and $Z_{0\mathrm{n}}$ are the characteristic impedances of these respective cavities. The mouth and pharyngeal tubes have closed acoustic terminations, while the nasal tube is open (at the nostrils). Dimensional similarity of the pharyngeal and nasal tubes allows a simplification: if $l_\mathrm{p} = l_\mathrm{n} = 10.5$ cm, each of $1/Z_\mathrm{p}$ and $1/Z_\mathrm{n}$ has periods of about 1.6 kHz and the function $1/Z_\mathrm{p} + 1/Z_\mathrm{n}$ has infinite values about every 800 Hz.

The mouth tube for nasal consonants is often significantly shorter than the other tubes (e.g., 3–7 cm). As a result, nasal consonants are characterized by: (a) formants every 800 Hz (due to the longer pharynx+nasal tube than the normal pharynx+mouth tube), (b) wider formant bandwidths, and (c) zeros in the spectrum. When airflow from the lungs reaches the velum junction, it divides according to the impedances of the mouth and nasal tubes. Spectral zeros occur at frequencies where $Z_\mathrm{m} = 0$, which results in no airflow into the nasal tube and thus no nasal speech output. Solving $Z_\mathrm{m} = -\mathrm{i}Z_{0\mathrm{m}}\cot(2\pi F_i l_\mathrm{m}/c) = 0$ for F_i yields zeros at odd multiples of $c/4l_\mathrm{m}$. The mouth tube for /m/ is about 7 cm, which gives zeros at 1.2, 3.6, 6.0 kHz, ... Shorter tubes for /n/ and /η/, about 5 and 3 cm, respectively, mean fewer zeros below 5 kHz: only one each at 1.7 and 2.8 kHz, respectively. Besides the poles due to the pharynx and nasal cavities, which occur every 800 Hz, nasal spectra have *pole–zero pairs* due to the mouth cavity; i. e., each zero is close in frequency to a mouth cavity pole. In spectra, nasal consonants appear as sounds with relatively steady formants, weaker than for vowels and with less coarticulation, as movements within the VT are muffled by oral tract closure.

Unlike the sonorants, all obstruents (except the glottal /h/) have an excitation source in the (upper) oral cavity. Air passing through a narrow constriction there creates frication noise just in front of the opening, which excites the remainder of the oral cavity in front of the constriction. This is usually modeled as random (Gaussian) noise with an approximately flat spectrum. For voiced obstruents (e.g., /v/ and /z/), the noise is modulated to be pulse-like, as the vocal folds vibrate; in spectrograms, this adds a *voice bar* at very low frequencies (0–150 Hz). This bar is weak enough not to be confused with formants, as well as being outside the range of F_1.

11.4.3 Obstruents

As noted above, formants are dynamic, varying greatly in time and intensity. They often come close together, and at other times fade in or out. Such changes are due to VT movements, and to the nature of VT excitation. Voiced excitation is glottal and usually relatively strong, but its intensity decreases with increasing frequency, owing to the low-pass nature of glottal puffs of air. As a result, a voiced speech spectrum falls off with frequency at about -6 dB/octave, which leads to weaker high-frequency formants. For most obstruent sounds (i. e., stops and fricatives) the VT excitation occurs much higher in the VT than at the glottis (except for the fricative /h/), and as a result, a much shorter VT is excited. This (often much) shorter VT leads to much higher resonances appearing in speech spectra, or more precisely, to the lower-frequency resonances being canceled by antiresonances of the back cavity of the VT. Many researchers do not refer to details of obstruent spectra as *formants* (i. e., they reserve the term formants for spectra of sonorant phonemes); to describe obstruents, they instead note general details such as the approximate cutoff frequency below which little energy occurs. The justification for this is that listeners pay much less attention to detailed aspects of the spectra in obstruents, and speakers appear to exercise less control over the positioning of resonances there. Hence, we focus our attention on formant estimation for sonorants.

11.4.4 Coarticulation

The motion of the VT during speech causes very dynamic patterns in formants, i. e., we often see rapid movements of formants. This phenomenon is called coarticulation, because the articulatory configurations of neighboring phonemes affect the articulation of each phoneme. Even though text is written with discrete letters, and phoneticians note that speech consists of a sequence of individual phonemes, the actual speech signal is without obvious phoneme boundaries, owing to coarticulation. Speakers smoothly move their VT from positions appropriate for each phoneme in turn, spending as much time on each phoneme as is deemed appropriate. As a result, it is often hard to segment speech into discrete units for analysis, as would be very useful in ASR. In some cases, such a division is easy, as when excitation changes abruptly with the start or end of vocal fold vibration (i. e., a voiced–unvoiced transition), or when the oral portion of the VT closes or opens (e.g., lip closure). More often, the transition between phonemes is more subtle, as the tongue and lips move between positions appropriate for successive phonemes.

Each phoneme has a nominal or *target* VT shape that a speaker would more or less assume if the sound were to be produced in isolation. In context, however, the speaker thinks ahead, and is continually moving various parts of the VT to accomplish the dynamic sequence of phonemes. Thus, coarticulation effects can extend over several phonemes; e.g., in the word 'strew' (/stru/, phonetically), the lips round in anticipation of the /u/ during the earlier /s/.

In many applications, it is useful to divide speech into segments of linguistic relevance, e.g., words, syllables, or phonemes. For example, for speech synthesis-by-rule (TTS), continuously spoken speech from a training speaker must be segmented into such units for storage (for later concatenation, as needed for a specified input text). As manual segmentation is tedious, forced alignment is often imposed on such training speech, given the assumed corresponding text. As a result, we have applications where formant tracking is simplified by having a priori knowledge of what phoneme sequence was actually spoken [11.10]. Such studies report quite high accuracy of forced alignment to phonemic boundaries in speech of a known text, to within about 40 ms [11.11]. In more-general ASR applications, of course, we do not know beforehand what was said, and formant estimation must rely on general principles.

During slow speech, the VT shape and type of excitation may not alter for periods of up to 200 ms. In these sections of speech, formant tracking is usually much easier, as the formants vary little in either position or amplitude. Most of the time, however, the VT changes more rapidly, as phonemes last, on average, about 80 ms. Coarticulation and changing F_0 can render each pitch period different from its neighbor. Nonetheless, a basic assumption of speech analysis is that the signal properties change relatively slowly with time. This allows examination of a short time window of speech (e.g., multiplying the speech signal by a Hamming window) to extract parameters presumed to remain fixed for the duration of the window. Most techniques thus yield parameters averaged over the course of the time window. To model dynamic parameters, we divide the signal into successive windows (analysis frames). Slowly changing formants in long vowels could allow windows as large as 100 ms without obscuring the desired parameters via averaging, but rapid events (e.g., stop releases) need short windows of about 5–10 ms to avoid averaging spectral transitions with the steadier spectra of adjacent sounds.

11.5 Short-Time Speech Analysis

A basic tool for spectral analysis is the spectrogram, which converts a two-dimensional speech waveform (amplitude versus time) into a three-dimensional pattern (amplitude/frequency/time). With time and frequency on the horizontal and vertical axes, respectively, amplitude is noted by the darkness of the display (Fig. 11.1). Peaks in the spectrum appear as dark horizontal bands. The center frequencies of these bands are generally considered to be the formant frequencies (subject to the discussion below of how to handle merged formants, i. e., single, wider bands that display two or more resonances that have come close for certain periods of time; e.g., in the middle of Fig. 11.3, F_1 and F_2 combine at very low frequencies in /w/). Voiced sounds cause vertical marks in the spectrogram due to an increase in the speech amplitude each time the vocal folds close. The noise in unvoiced sounds causes rectangular dark patterns, randomly punctuated with light spots due to instantaneous variations in energy. Spectrograms portray only spectral amplitude, ignoring phase information, on the assumption that phase is less important for most speech applications. We will thus ignore phase in our discussion of formants.

In the spectrogram, the amplitude of the STFT $|S_n(\mathrm{e}^{\mathrm{i}\omega})|$ is plotted with time n on the horizontal axis, frequency ω (from 0 to π) on the vertical axis (i. e., 0 to $F_s/2$ in Hz, F_s being the sampling frequency), and with magnitude indicated as darkness, typically on a logarithmic scale (e.g., decibels). Two different display styles are typical: wideband and narrowband, with wideband displays used mostly for formant tracking.

Wideband spectrograms display individual pitch periods as vertical striations corresponding to the large speech amplitude each time the vocal cords close. Voicing is readily seen in the presence of these periodically spaced striations. Fine time resolution here permits accurate temporal location of spectral changes corresponding to VT movements. A wide filter bandwidth smooths the harmonic amplitudes under each formant across a range of (typically) 300 Hz, displaying a band of darkness (of width proportional to the formant's bandwidth) for each formant. The center of each band is a good estimate of formant frequency. Formant detectors generally prefer spectral representations that smooth the fine structure of the harmonics while preserving formant structure. Traditional wideband spectrograms use a window of about 3 ms, which corresponds to a bandwidth of 300 Hz and smooths harmonic structure (unless $F_0 > 300$ Hz, which occurs with children's voices). Narrowband spectrograms, on the other hand, generally use a window with a bandwidth of approximately 45 Hz and thus a duration of about 20 ms, which allows resolution of individual harmonics (since $F_0 > 45$ Hz) but smooths speech in time over a few pitch periods.

11.5.1 Vowels

Vowels are voiced and have the greatest intensity of all phonemes. They normally range in duration from 50 to 400 ms [11.1]. Vowel energy is mostly concentrated below 1 kHz and falls off at about −6 dB/oct with frequency. Spectral displays thus often use pre-emphasis to boost higher frequencies to facilitate formant tracking; e.g., LPC treats all frequencies the same, so LPC analysis without pre-emphasis would tend to have poorer spectral estimates at higher frequencies. Other formant methods that rely on peak-picking of spectra would likely have similar difficulties without pre-emphasis. As such boosting also raises the level of any background noise, it should be noted that formant estimation for high frequencies where noise dominates will necessarily be less accurate. Because of the −6 dB/oct fall-off, few formants above F_4 are reliable in many formant estimation methods. Pre-emphasizing the speech is usually done by differencing in discrete time:

$$y(n) = s(n) - as(n-1) ,$$

where a is typically 0.9–1.0. While this may greatly attenuate frequencies below 200 Hz, such low frequencies are rarely of interest in most speech applications. Vowels are distinguished primarily by the locations of their first three formant frequencies (F_1, F_2, and F_3).

11.5.2 Nasals

In nasal consonants, F_1 near 250 Hz dominates the spectrum, F_2 is usually very weak, and F_3 near 2200 Hz has the second-highest formant peak. A spectral zero, whose frequency is inversely proportional to the length of the oral cavity behind the constriction, occurs near 1 kHz for /m/, near 2 kHz for /n/, and above 3 kHz for the velar nasal. Spectral jumps in both formant amplitudes and frequencies coincide with the occlusion and opening of the oral tract for nasals. These abrupt changes cause difficulties for the continuity constraints for formant trackers, as the usual trend toward smooth formant movements is invalid at nasal boundaries. VTRs change abruptly when the oral cavity opens or closes. The lowering of the velum is not the principal factor in the spectral change here; it often lowers during a vowel preceding a nasal consonant, which causes nasalization of the vowel, widening the formants and introducing zeros into the spectrum. Vowel nasalization primarily affects spectra in the F_1 region.

11.5.3 Fricatives and Stops

As obstruents, fricatives and stops are very different from sonorants: aperiodic, much less intense, and often with most energy at high frequencies. Obstruents may be either voiced or unvoiced. Unvoiced fricatives have a high-pass spectrum, with a cutoff frequency approximately inversely proportional to the length of the front cavity of the VT. Thus the palatal fricatives are most intense, with energy above about 2.5 kHz; they have a large front cavity. The alveolar fricatives (e.g., /s/) lack significant energy below about 3.2 kHz and are thus less intense. The labial and dental fricatives are very weak, with little energy below 8 kHz, due to a very small front cavity. The glottal fricative /h/, despite exciting the full VT, also has relatively low intensity as its noise source at the glottis (effectively a whisper) is usually weaker than noise from oral tract constrictions.

It is not obvious how to handle formant tracking for obstruents. Traditionally, formants are well defined only for sonorant sounds, where the general rules of strong resonances, spaced roughly every 1000 Hz, apply. Except for /h/, obstruents have little energy in the low-frequency range of 0–2 kHz, where strong F_1 and F_2 (for sonorants) have most energy. Depending on the length of the VT in front of the constriction noise source, there may be little energy in most of the useful auditory range. As noted above, VTRs are always present, no matter where the excitation is; however, low-frequency VTRs are not excited in most obstruents, and thus are not accessible in speech analysis.

In a transition from a sonorant to an obstruent (the observations that follow here also apply, in reverse, for a transition from an obstruent to a sonorant), the visible formants usually show movements from spectral positions pertinent for the sonorant toward targets for the ensuing obstruent (e.g., a decrease in F_1 as the VT closes; F_2 falling for a labial obstruent or rising for an alveolar; F_3 rising for an alveolar and falling for a velar). In some cases, for a given formant, a smooth formant transition (obeying continuity) is clearly seen; when this happens, it is usually for F_3 or F_4, as these tend to be in the frequency range of overlap between sonorants and obstruents. Thus continuity constraints should vary with context, and not be applied across all frequencies equally.

Tracking formants at stop transitions is particularly interesting and difficult. It is important because crucial phonetic information is present during these brief periods, which are major factors in informing listeners of the articulation point of the stop. Other major phonetic cues are much more prominent in the speech signal, e.g., over longer durations and with greater intensity. Stops, on the other hand, do not cue their place during most of their duration, as the VT closure at that time means the only audible energy is the voice bar (if voiced).

The release of the VT occlusion at the end of a stop creates a brief (few ms) explosion of noise, which tends to excite all frequencies. Then, turbulent noise (frication) continues as the constriction opens for 10–40 ms, exciting the front cavities (usually F_2–F_4), as the VT moves toward the position for an ensuing sonorant. A velar constriction provides a long front cavity, with a low resonance near 2 kHz (F_2 or F_3). Velar resonances are higher due to a shorter front cavity. The spectrum of a labial burst is relatively flat and weak since there is essentially no front cavity to excite.

Most formant estimators either do poorly during obstruents, or simply claim that such regions do not need formant estimation. A statement such as the latter is relatively true, as many speech applications have a primary need for spectral estimation during the strong sections of speech; weaker sounds may often be modeled more simply, and their perception by listeners may be less critical to the communication task at hand. Nonetheless, knowing some details about what happens during the weaker sounds is of interest. Identifying the articulation point is mostly done via the formant transitions in adjacent sonorants, and the formant behavior at the stop release (normally considered to be part of the obstruent, and not part of the ensuing sonorant) is relevant. Furthermore, weak fricatives such as /v/ require attention.

11.6 Formant Estimation

Typical methods to estimate formants involve searching for peaks in spectral representations, usually from an STFT or LPC analysis [11.12, 13]. As LPC imposes an assumed simplified structure on the speech spectrum, it appears to have been employed most often in recent methods. In most LPC applications, one uses two poles per kilohertz of bandwidth (plus 2–4 additional poles to model other factors, such as the spectral tilt, which is due to glottal effects), on the assumption that speech contains one such formant in that range. A spectrum derived from an LPC model of N=10–16 poles is much more limited in variation (across frequency) than an STFT, which thus simplifies peak-picking. A disadvantage of using LPC is that its all-pole modeling is not perfect [11.14]; it chooses its pole positions to minimize mean-square error (MSE) for a fit of the speech to an N-pole spectral envelope. If the number of poles is not well chosen (e.g., to match the number of resonances clearly present in the speech), then the model spectrum is not as accurate as desired. For example, if there are too few poles in the model, the poles (selected via automatic analysis) have to place themselves in compromise locations between actual formants, and significant errors in formant tracking will result. Thus, it is rare that too few poles are used in LP analysis. Use of too many poles is a more likely risk; it is standard to follow the rule of thumb given above (2 poles/kHz), yet a given speech spectrum will often display fewer resonances than other phonemes (e.g., /u/ vowels and nasals often only show 2–3 formants, even in wideband applications, as higher formants tend to be very weak and thus poorly modeled via LPC, which focuses on peak energy). In this case, the extra poles (i. e., those not needed to directly model the actual resonances) locate themselves at non-formant frequencies to reduce the MSE further. Sometimes, if there are only, say, two extra poles, the additional poles will be real or have very wide bandwidth (to model the speech better via some broad spectral tilt effect). As LPC formant trackers usually examine the bandwidths of the LPC poles and reject (as potential formants) poles with wide bandwidth, such an effect is not a major problem. More-serious errors can arise when the additional poles model individual harmonics, e.g., in the strong F_1 region, as may happen in cases where individual formants have few harmonics; if F_1 has two dominant harmonics (e.g., the F_1 center frequency is located between two harmonics and the F_1 bandwidth is similar to F_0), then LPC may well assign four poles to model the two harmonics of F_1; this would lead to two formant candidates in the F_1 region, and require postprocessing to decide whether these candidates need to be merged into a single formant value.

Use of an STFT instead of an LPC spectrum avoids this latter problem, as it does not impose a specific (e.g., all-pole) model on the speech spectrum. Nonetheless, the issue of data reduction remains when using the STFT, as a typical STFT has 256–512 samples, corresponding to common FFT duration choices, for windows approximately 20–25 ms in length. One normally needs to smooth the STFT, and then do peak-picking. One can *pad with zeros*, i. e., select a much shorter range of speech samples than the FFT length (i. e., fewer samples than an estimated pitch period), to eliminate the harmonics from the spectral display, which effectively smooths the spectrum. In such a case, one normally needs to do pitch-synchronous analysis to choose at least the initial strong samples of each pitch period. Alternatively, one can smooth the FT of a full window with a low-pass filter operating in the frequency domain.

11.6.1 Continuity Constraints

It has generally been found that peak-picking methods need to be subject to continuity constraints so as to select from among multiple formant candidates, as there are often more candidates (i. e., spectral peaks) than formants [11.15]. One normally prunes away any candidate whose bandwidth is beyond the range of formants, i.e., a candidate whose bandwidth is less than 50 Hz, which is likely to be an interfering tone or an individual harmonic rather than a formant, or more than, say, 300 Hz, as formant bandwidths increase with center frequency, e.g., a roughly constant Q of around 5–6, so this upper threshold should increase with frequency. In sections of speech that appear to be sonorants, i. e., strong, voiced speech, a tracker aims to assign one formant to each possible range, i. e., roughly one formant per 1000 Hz. We need to allow for many individual cases where there are two formants below 1000 Hz (e.g., /o/ and /u/) and F_2 above 2 kHz (e.g., /i/), but within, say, the typical range of 0–4 kHz, there should be four formants (assuming a man's voice). If the signal has passed along a telephone line, then we should not expect to see more than three formants, as F_4 is often lost to the upper frequency cutoff of the phone lines (which preserve only the range of approximately 300–3200 Hz). Similarly, if the dynamic range of the spectral amplitude obtained is limited, then F_4 may in general be too weak to be clearly observed, especially if the background noise level is high enough to obscure the weaker formants, which often include F_4.

The most difficult area for designing good continuity constraints is in the temporal dynamics. Automatic tracking of formants is difficult mostly owing to rapid changes in the formant patterns when a VT closure occurs or when the VT excitation changes state (between voiced and unvoiced); such changes often occur several times per second in speech. From frame to frame, in most speech, the formants generally change slowly. Other than the abrupt formant changes (which are due to major VT changes), the most rapid formant changes are normally seen in F_2 when the tongue moves quickly in lateral motion (e.g., /ai/ and /oi/) or when the lips round/unround. The maximum rate of change for a formant is approximately 20 Hz/ms (e.g., a change of 1200 Hz over 60 ms); so any proposed formant changes exceeding this threshold should be limited to major phoneme boundaries, where all the formants change and the overall intensity level also changes abruptly.

The apparent merging of formants has been noted as a major problem for formant trackers, i. e., when two or more formants are sufficiently close to each other to present an almost solid band of energy in a given spectral display. Many sounds have two formants close enough that they may potentially appear as one spectral peak (e.g., F_1–F_2 in /a/, /o/, and /u/, and F_2–F_3 in /i/ and /r/). One should not normally rely on small rises and falls (across frequency) in a spectrum to delineate individual formants, as such small changes can easily be due to window or harmonic effects. (An LPC spectrum, of course, normally does not have such rapid, small changes, but close formants also cause difficulties with LPC as well [11.14].) Thus, a spectral peak normally requires a significant rise in spectral amplitude over a frequency range on the order of a typical formant bandwidth, in order to be declared a good formant candidate.

It is generally by applying continuity constraints that one can resolve formant merges. While formants may be quite close during the steady state of a vowel, coarticulation with adjacent phonemes generally causes sufficient formant motion of all formants that nearby frames of speech clearly show separating tracks of spectral peaks. When a formant tracker is in doubt about a wide band of speech energy in a given section of speech, it can scan left and right (i. e., before and after) to look for a possible peak that may be moving away from the main band of energy. Often, this is a case of a weaker, rising F_3 or F_4 that was temporarily *merged* with F_2 or F_3, respectively; in recent papers on formant tracking, one often sees examples of the tracker incorrectly choosing a stronger and higher-frequency spectral peak as a formant, while ignoring a weaker track (a true formant) moving away from a wide band of speech energy (in which case all of these are formants).

A number of formant trackers focus on precise estimates of the center frequencies of resonances within frequency bands that are assumed to contain a single formant [11.16]. For each frame, they divide up the spectrum into such estimated bands in an initial analysis step, then use simpler estimation methods within each band. If indeed each band has a single resonance, the precise location of its center frequency and bandwidth is simpler through the use of adaptive bandpass filters that seek to isolate individual formants, thus allowing a more-precise focus on details of the (presumed single) resonance within each chosen band of frequency [11.16]. At first glance, this might seem to beg the question of formant tracking, as we have already identified the issue of separating formants that may closely approach each other as potentially difficult. Yet reasonable results appear possible with appropriate filtering, even in noisy conditions [11.17, 18]. One approach uses a parallel formant synthesizer, as was common in older TTS systems to generate a hypothesized synthetic speech spectrum (and thus with specific, known formants), and compares that spectrum to the actual speech spectrum, minimizing a spectral distance measure (often a quadratic cost function relating the two speech spectra, and also aiming for temporal continuity) via dynamic programming (DP, often via hidden Markov models) [11.19]. While DP is popular in formant trackers [11.11], some recent versions of such an approach seem not to need DP [11.20].

11.6.2 Use of Phase Shift

A common way to track formants is to estimate speech $S(z)$ in terms of a ratio of z-polynomials (e.g., an all-pole LPC spectrum), solve directly for the roots of the denominator, and identify each root as a formant if it has a narrow bandwidth at a reasonable frequency location [11.14]). Other approaches use related phase information [from $S(z)$] to decide whether a spectral peak is a formant. When one evaluates $S(z)$ along the unit circle $z = \exp(\mathrm{i}\omega)$, a large negative phase shift occurs when ω passes a pole close to the unit circle. As formants correspond to complex-conjugate pairs of poles with relatively narrow bandwidths (i. e., near the unit circle), each spectral peak having such a phase shift is normally a formant. The phase shift approaches $-180°$ for small formant bandwidths.

In cases where two formants appear as one broad spectral peak, a modified discrete Fourier transform DFT can resolve the ambiguity. The chirp z-transform (CZT) calculates the z-transform of the windowed speech on a contour inside the unit circle. Whereas the DFT samples $S(z)$ at uniform intervals on the unit circle, the CZT may take a spiral contour anywhere in the z-plane. It may be located near poles corresponding to a spectral peak of interest and thus need to be evaluated only for a small range of frequency samples in pertinent cases. As a contour can be much closer to the formant poles than for the DFT, the CZT can resolve two poles (for two closely spaced formants) into two spectral peaks. Since formant bandwidths tend to increase with frequency, the spiral contour often starts near $z = \alpha$, just inside the unit circle (e.g., $\alpha = 0.9$), and gradually spirals inward with increasing frequency $\omega_k = 2\pi k/N$ ($z_k = \alpha\beta^k \exp(\mathrm{i}\omega_k)$, with β just less than 1). This contour would thus follow the expected path of the formant poles and can eliminate problems of merged peaks in DFT displays.

Other recent formant tracking methods are also based on phase in related ways [11.21]. Formant trackers experience increased difficulty when F_0 exceeds formant bandwidths, e.g., $F_0 > 250\,\mathrm{Hz}$, as in children's voices. Harmonics in such speech are so widely separated that only one or two appear in each formant. As a result, spectral analyzers have a tendency to label the prominent harmonics as formants, which is generally wrong, as the center frequency of a formant is rarely an exact multiple of F_0.

11.6.3 Smoothing

Many pattern estimation algorithms, including formant estimators, need to do postprocessing on the raw data that comes out of the main estimation processing step. Usually, such schemes produce estimated values once per frame (e.g., every 5–10 ms in many speech applications); this applies to ASR and to speech coders, as well as to F_0 trackers. In ASR, estimated values for VT representations are modeled by probability distributions, and decisions are made on global probabilities combining hundreds (or more) of computations; a small deviation in any one parameter has little effect, so smoothing of individual parameters in a frame is rarely needed for ASR. Here, we assume that estimated formants may be used for a variety of applications, and thus we judge performance on how well each output value matches the actual speech, and thus are less tolerant of even small errors.

While formant decisions involve continuity constraints, these constraints normally do not greatly affect localized estimations; they are instead mostly used to avoid large errors (such as missing a formant entirely). The decision in each frame is usually based on the immediate speech window, whose placement and duration are rarely synchronized for optimal estimation results. As a result, estimated formant values are often noisy, in the sense that they err slightly above or below their actual values, owing to suboptimal calculations. The estimations are usually within perceivable ranges (i. e., if we were to synthesize speech with the estimated formant values, and compare this with synthesis based on the true values, listeners would hear no difference), but it is usually preferred to smooth such noisy formant signals to present a final formant contour that is both accurate and tidy.

One initial idea here would be simply to pass any noisy formant contour, as a time signal, through a low-pass filter, choosing the cutoff frequency of the filter so as to smooth the small random deviations and not the useful formant movements related to VT movements. However, as when smoothing the output of F_0 detectors, formant estimations sometimes change rapidly between individual frames, and applying a linear low-pass filter would produce a poor result when formants do actually change abruptly; instead of a true abrupt formant jump for a nasal, the smoothed pattern would show a gradual rise or fall.

Another difficulty with linear filtering is its behavior when mistakes occur in parameter extraction. Formant and F_0 estimators sometimes produce erroneous isolated estimates (i. e., for individual frame outputs) called *outliers*, which deviate greatly from the rest of the parameter contour. Such mistakes must be corrected in postprocessing. As linear filters give equal weight to all signal samples, they would propagate the effect of such a mistake into adjacent sections of the smoothed output formant contour.

Thus, a common alternative to linear filtering is median smoothing [11.1], which preserves sharp signal discontinuities while eliminating fine errors and gross outliers. Most smoothers (linear and nonlinear) operate on a finite time window of the input signal, but linear smoothers combine the windowed samples linearly to produce the smoothed output sample, whereas median smoothing chooses a single value from among

Part B | 11.6

the window samples. In each data window, the samples are ordered in amplitude with no regard for timing within the window. The output sample is the median, i. e., the $[(N+1)/2]$th of N ordered samples (for odd N). Sudden discontinuities are preserved because no averaging occurs. Up to $(N-1)/2$ outlier samples, above or below the main contour, do not affect the output.

Median smoothers do well in eliminating outliers and in global smoothing, but do not provide very smooth outputs when dealing with noisy signals. Thus they are often combined with linear smoothers to yield a compromise smoothed output, whose sharp transitions are better preserved than with only linear filtering but with a smoother output signal than would be possible using only median smoothing.

11.7 Summary

This chapter has presented an introduction to formant tracking methods for speech. We have concentrated on the major approaches to formant tracking, generally using Fourier or LPC displays, and either peak-picking or solving for the roots in the LPC all-pole polynomial. There are also other proposed methods, such as a bank of inverse filters [11.22].

References

11.1 D. O'Shaughnessy: *Speech Communication: Human and Machine*, 2nd edn. (IEEE, Piscataway 2000)

11.2 J. Darch, B. Milner, S. Vaseghi: MAP prediction of formant frequencies and voicing class from MFCC vectors in noise, Speech Commun. **11**, 1556–1572 (2006)

11.3 R. Togneri, L. Deng: A state-space model with neural-network prediction for recovering vocal tract resonances in fluent speech from Mel-cepstral coefficients, Speech Commun. **48**(8), 971–988 (2006)

11.4 K. Weber, S. Ikbal, S. Bengio, H. Bourlard: Robust speech recognition and feature extraction using HMM2, Comput. Speech Lang. **17**(2–3), 195–211 (2003)

11.5 W. Ding, N. Campbell: Optimizing unit selection with voice source and formants in the CHATR speech synthesis system, Proc. Eurospeech (1997) pp. 537–540

11.6 J. Malkin, X. Li, J. Bilmes: A graphical model for formant tracking, Proc. IEEE ICASSP, Vol. 1 (2005) pp. 913–916

11.7 K. Sjlander, J. Beskow: WAVESURFER – an open source speech tool, Proc. ICSLP (2000)

11.8 L. Deng, L.J. Lee, H. Attias, A. Acero: A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances, Proc. IEEE ICASSP, Vol. 1 (2004) pp. 557–560

11.9 Y. Zheng, M. Hasegawa-Johnson: Formant tracking by mixture state particle filter, Proc. IEEE ICASSP, Vol. 1 (2004) pp. 565–568

11.10 D.T. Toledano, J.G. Villardebo, L.H. Gomez: Initialization, training, and context-cependency in HMM-based formant tracking, IEEE Trans. Audio Speech **14**(2), 511–523 (2006)

11.11 M. Lee, J. van Santen, B. Mobius, J. Olive: Formant tracking using context-dependent phonemic information, IEEE Trans. Speech Audio Process. **13**(5), 741–750 (2005), Part 2

11.12 S. McCandless: An algorithm for automatic formant extraction using linear prediction spectra, Proc. IEEE ICASSP **22**(2), 135–141 (1974)

11.13 G. Kopec: Formant tracking using hidden Markov models and vector quantization, Proc. IEEE ICASSP **34**(4), 709–729 (1986)

11.14 G.K. Vallabha, B. Tuller: Systematic errors in the formant analysis of steady-state vowels, Speech Commun. **38**(1–2), 141–160 (2002)

11.15 Y. Laprie, M.-O. Berger: Cooperation of regularization and speech heuristics to control automatic formant tracking, Speech Commun. **19**(4), 255–269 (1996)

11.16 K. Mustafa, I.C. Bruce: Robust formant tracking for continuous speech with speaker variability, IEEE Trans. Audio Speech **14**(2), 435–444 (2006)

11.17 A. Rao, R. Kumaresan: On decomposing speech into modulated components, IEEE Trans. Speech Audio Process. **8**(3), 240–254 (2000)

11.18 I.C. Bruce, N.V. Karkhanis, E.D. Young, M.B. Sachs: Robust formant tracking in noise, Proc. IEEE ICASSP, Vol. 1 (2002) pp. 281–284

11.19 L. Welling, H. Ney: Formant estimation for speech recognition, IEEE Trans. Speech Audio Process. **6**(1), 36–48 (1998)

11.20 B. Chen, P.C. Loizou: Formant frequency estimation in noise, Proc. IEEE ICASSP, Vol. 1 (2004) pp. 581–584

11.21 D.J. Nelson: Cross-spectral based formant estimation and alignment, Proc. IEEE ICASSP, Vol. 2 (2004) pp. 621–624

11.22 A. Watanabe: Formant estimation method using inverse-filter control, IEEE Trans. Speech Audio Process. **9**(4), 317–326 (2001)