

CHAPTER 1

SPEECH ANALYSIS: THE PRODUCTION-PERCEPTION PERSPECTIVE

Li Deng[†] and Jianwu Dang[‡]

[†]*Microsoft Research*

One Microsoft Way, Redmond, WA 98052

[‡]*School of Information Science*

Japan Advanced Institute of Science and Technology

1-1 Asahidai, Nomi, Ishikawa, 923-1292

Email: {deng@microsoft.com, jdang@jaist.ac.jp}

This chapter introduces the basic concepts and techniques of speech analysis from the perspectives of the underlying mechanisms of human speech production and perception. Spoken Chinese language has special characteristics in its signal properties that can be well understood in terms of both the production and perception mechanisms. In this chapter, we will first outline the general linguistic, phonetic, and signal properties of spoken Chinese. We then introduce human production and perception mechanisms, and in particular, those relevant to spoken Chinese. We also present some recent brain research on the relationship between human speech production and perception. From the perspectives of human speech production and perception, we then describe popular speech analysis techniques and classify them based on the underlying scientific principles either from the speech production or perception mechanism or from both.

1. Introduction

Chinese is the language of over one billion speakers. Several dialect families of Chinese exist, each in turn consisting of many dialects. Although different dialect families are often mutually unintelligible, systematic correspondences (e.g., in lexicon and syntax) exist among them, making it easy for speakers of one dialect to pick up another relatively quickly. The largest dialect family is the Northern family, which consists of over 70% of all Chinese speakers. Standard or Mandarin Chinese is a member of the Northern family and is based on the pronunciation of the Beijing dialect. Interestingly, most speakers of Standard

Chinese have another dialect as their first tongues and only less than one percent of them speak without some degree of accent.

According to a rough classification, Mandarin Chinese has five vowels, three high /i, y, u/ (one of them, /y/, is rounded), one mid /ə/, and one low /a/. When the high vowels occur before another vowel, they behave as glides. The context dependency of the vowels has simpler rules than that for English.

There are 22 consonants in Mandarin Chinese. Compared with English, the distribution of consonants in Mandarin Chinese is more closely dependent on the syllable position, and the syllable structure is much simpler.

There are two types of syllables – full and weak ones – in Mandarin Chinese. The former has intrinsic, underlying tone and is long, while the latter has no intrinsic tone and is short. A full syllable may change to a weak one, losing its intrinsic tone and undergoing syllable rime reduction and shortening (similar to syllable reduction in English).

In contrast to English which has over 10,000 (mono) syllables, Mandarin Chinese has only about 400 syllables excluding tones (and 1300 including tones). Relatively simple phonological constraints can sufficiently describe the way in which many available syllables are excluded as being valid ones in Mandarin Chinese.

In addition to the specific phonological and phonetic properties of spoken Chinese outlined above, the special characteristics in its signal properties consist of tonality and fundamental frequency variations that signal the lexical identity in the language in addition to paralinguistic information. Speech analysis techniques for fundamental frequency or pitch extraction are therefore more important for Chinese than for the non-tonal languages such as English. Recent research has provided both the production and perceptual accounts for tonal variations in spoken Chinese, where the articulatory constraint on the perception processing has been quantified. To understand the underlying science for the speech processing techniques, we will first present the human mechanisms in speech production and perception.

2. Speech Production Mechanisms

2.1. *Speech Organs*

Many speech analysis techniques are motivated by the physical mechanisms by which human speech is produced. The major organs of the human body responsible for producing speech are the lungs, the larynx (including the vocal cords), the pharynx, the nose and the mouth (including the velum, the hard palate, the teeth, the tongue, and the lips), which are illustrated by the midsagittal view

shown in Figure 1. Together, these organs form a complicated “tube” extending from the lungs to the lips and/or the nostrils. The part of the tube superior to the vocal cords is called the vocal tract. During speech production, the shape of the vocal tract varies extensively by movements of the tongue, jaw, lips, and the velum. These organs are named the articulators and their movement is known as articulation.

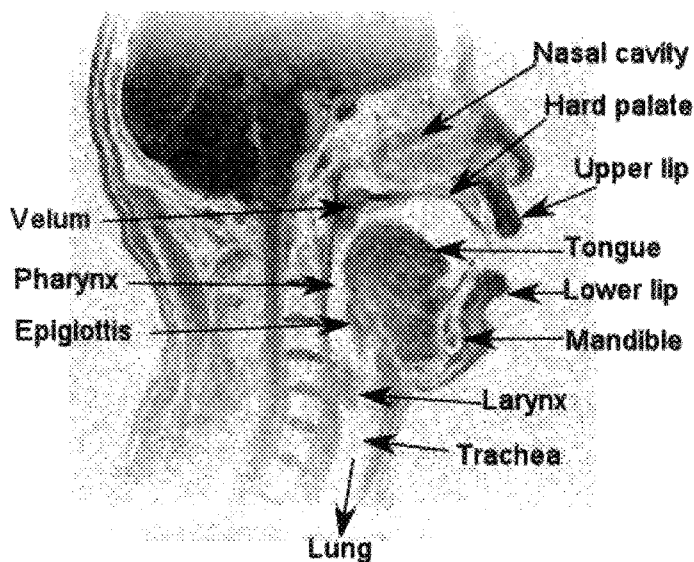


Fig. 1. MRI-based midsagittal view of the human head, showing the vocal organs.

2.2. Acoustic Filtering in Speech Production

In speech production, a time-varying vocal tract is formed by articulation. Each specific vocal tract shape has its inherent resonances and anti-resonances corresponding to its shape, and functions as a filter.^{1,2} Although both the resonances and anti-resonances are decisive factors for generating sound, the former one is of prime importance in determining the properties of speech sounds. For this reason, in speech analysis, we often treat the vocal tract as an all-pole filter. The principal resonant structure, particularly for vowels, is the essential characteristics for distinguishing one from another. That is, if a specific vocal tract with a proper source is given, the sound is uniquely determined. Note that there are many vocal tract shapes that are able to generate a given sound. This is called the inverse problem in speech processing.

To illustrate the relation between the geometry of the vocal tract and the resonance structure of vowels, we show the midsagittal view of the vocal tract shape obtained in pronouncing the Chinese vowels /a/, /i/, /u/ and /ə/ in Figure 2, which are located in the top-left, the lower-left, the top-right, and the lower-right panels, respectively. The geometric difference between vowels is clearly seen in the vocal tract. For instance, there is a wide opened cavity in the anterior part of the vocal tract with a lower tongue position for /a/, while for /i/ the posterior part is widely opened with a higher tongue position. According to the geometric structure, vowel /a/ is called *open vowel* or *low vowel*, while vowel /i/ is referred to as a *closed vowel* or *high vowel*. Geometrically, a key part in the vocal tract for determining the vowel properties is the narrowest portion of the vocal tract, namely the *constriction*. The location of constrictions is most crucial for producing a sound, and is used to describe the articulation of the vowels. As shown in the lower-right panel of Figure 2, the constrictions of /a/, /i/ and /u/ are located at the vertices of the triangle and /ə/ is in its center. Note that the constriction of the vocal tract is not as tight for vowels as that for consonants. For vowels, the location of the constriction is usually approximated using the highest position of the tongue body, namely the tongue height.

The resonances of the vocal tract are heavily dependent on the location of the constriction in the vocal tract, which can be used for distinguishing vowels. If there is no noticeable constriction in the vocal tract such as production of /ə/, the vocal tract can be roughly thought as a uniform tube whose resonances will appear at around 500 Hz, 1500 Hz, and so on, which are a multiple of the sound velocity over four times the tube length. If the constriction is located in the anterior part of the vocal tract such as the vowel /i/, the first resonance lowers down and the second one goes up, where the typical values are around 250 Hz and 2100 Hz. If the constriction is located in the anterior part of the vocal tract such as /a/, the first resonance moves upward to about 700 Hz and the second one reduces to about 1200 Hz. The resonant system of the vocal tract can be considered as a filter that shapes the spectrum of the sound source to produce speech. After a voiced source is modulated by the vocal tract filter, a vowel sound would be obtained from the lip radiation, where the resonant modes are known as *formants*. According to convention, formants are numbered from the low-frequency end, referred to as F1, F2, F3, etc. F1 and F2 roughly determine the major phonetic properties of the vowels, and the other formants mainly give the detailed information on timbre and individual differences.

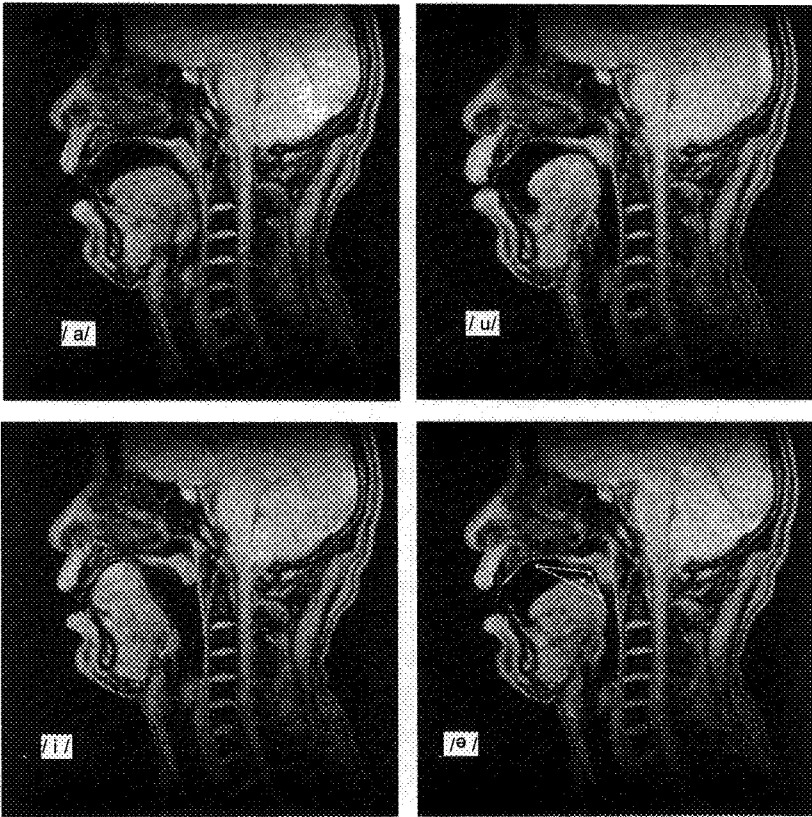


Fig. 2. Midsagittal view of the vocal tract in producing Chinese vowels: /a/ top-left, /i/ lower-left, /u/ top-right, and /ə/ lower-right. Constrictions of /a/, /i/ and /u/ are located in the vertexes of the triangle and /ə/ in the center.

2.3. Excitation Source in Speech Production

To understand speech production, especially the production of consonants, we also need to understand the sources of the sounds in addition to filters. The source or energy for speech production is the steady stream of air that comes from the lungs as we exhale. This air flow is modulated in a variety of ways and is transformed to the acoustic signal in the audio frequency range. In generating speech sounds, such energy can be classified to three different types of sources: quasi-periodical pulse train for vowels, turbulence noise for fricatives, or burst for plosives. Although the different source types may be combined with one another, the quasi-periodical pulse is the major source for voiced sounds (i.e., vowels or voiced consonants). In voiced sounds, the number of circles per second in the quasi-periodical pulse train is referred to as *fundamental frequency*, or F_0 ,

which is perceived as a pitch that describes how high a tone is. Since Chinese is a tone language, the fundamental frequency and its variance are very important for understanding Chinese words. We now describe in some detail the voiced sound source, where the speech organ, the larynx, plays a central role.

The glottis, which is the airspace inside the larynx, is opened by the separation of the left and right folds during normal breathing, where the air stream is inaudible. When the vocal folds get closer to one another, the air pressure in the subglottal part increases, and then the air stream becomes an audible pulse train due to vocal folds' vibration. The vocal folds are soft tissue structures contained within the cartilaginous framework of the larynx. The location of the larynx has been shown in Figure 1 and the structure of the larynx is shown in Figure 3, where (a) shows a side view and (b) is the top view. The anterior parts of the vocal folds are attached together at the front portion of the thyroid cartilage, while the posterior parts connect to the arytenoid cartilages on the left and right sides separately. To produce a sound, the arytenoid cartilages adduct (move together) the vocal folds and make the point of division between the subglottal and supraglottal airways. This action sets the vocal folds into rapid vibration as the air pressure in the subglottal part increases, and thus makes the air flow audible, namely *phonation*. For stopping the phonation the arytenoid cartilages abduct (move apart) the vocal folds, as the shape shown in Figure 3(b).

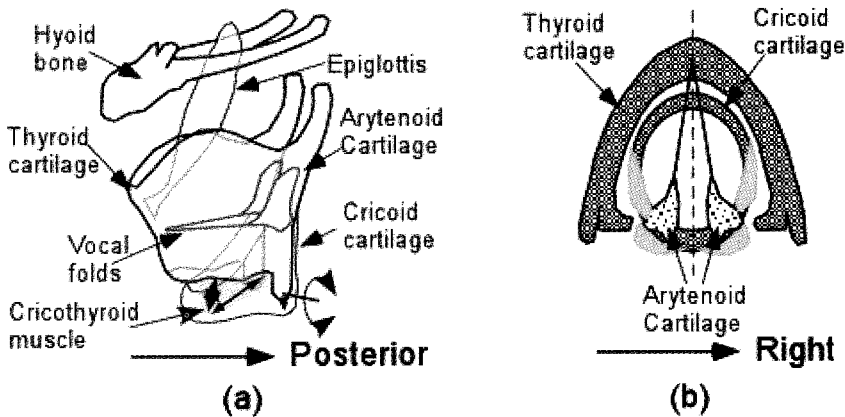


Fig. 3. The structure of the larynx (a) a side view and (b) the top view.

The structure of the vocal folds is often described by the cover–body concept. It suggests that the vocal folds can be roughly divided into two tissue layers with different mechanical properties. The cover layer is comprised of pliable, non-contractile mucosal tissue that serves as a sheath around the body-layer. In contrast, the body layer consists of muscle fibers (thyroarytenoid) and some

ligamentous tissue. For a phonation, the arytenoid cartilages move the vocal folds together, the vocal folds start a rapid vibration as the air pressure in the subglottal part increases. Suppose in the initial state the vocal folds on the left and right sides are in contact and the airway is closed. An idealized cycle of vocal fold vibration begins with a lateral movement of the inferior portion of the cover surface which continues until the left and right sides separate, thereby opening the airway. Once the maximum lateral displacement is achieved, the lower portion of the vocal folds begins to move medially with the upper portion following. Eventually, the lower portions on the left and right sides collide with each other and again closing the airway. Medial displacement continues as the upper portions also collide. The entire process then repeats itself cyclically at the fundamental frequency of vibration (F_0). Once in vibration, the vocal folds effectively convert the steady air flow from the lungs into a series of flow pulses by periodically opening and closing the air space between the vocal folds. This stream of flow pulses provides the sound source for the excitation of the vocal tract resonances in vowels. That is, the vocal folds are capable of converting a steady, flowing stream of air into vibratory motion.

One of the methods for controlling the F_0 is manipulating the lung pressure. In general, the F_0 increases as the lung pressure increases, and vice versa. For a tone language such as Chinese, there is a large scale change in the F_0 within a syllable. An easy way to generate Chinese tones, especially for a low tone, is by manipulating the lung pressure.

The second major source of sound in speech production is the air turbulence that is caused when air from the lungs is forced through a strict constriction in the vocal tract. Such constrictions can be formed in the region of the larynx, as in the case of [h] sounds in English (in Chinese, constrictions are around the velar location), and at many other places in the tract, such as between various parts of the tongue and the roof of the mouth, between the teeth and lips, or between the lips. The air turbulence source has a broad continuous spectrum, and the spectrum of the radiated sound is affected by the acoustics of the vocal tract, as in the case of voiced sounds. Sustainable consonant sounds that are excited primarily by air turbulence, such as [s, f], are known as *fricatives*, and hence the turbulent noise is often referred to as *frication*.

The third type of sound source results from the build-up of pressure that occurs when the vocal tract is closed at some point for a stop consonant. The subsequent plosive release of this pressure produces a transient excitation of the vocal tract which causes a sudden onset of sound. If the vocal folds are vibrating during the pressure build-up, the plosive release is preceded by a low-level sound, namely a *voice bar*, which is radiated through the walls of the vocal tract. If the

vocal folds are not vibrating during the closure, it needs time to build up a pressure difference between the supraglottal and subglottal portions by exhausting the stored air behind the closure. The time from releasing the closure to vocal folds vibration is called as the voice onset time (VOT). Since Mandarin Chinese has no voiced stop consonant, the VOT seems to be more crucial for perceiving Chinese stop consonants. The plosive release approximates a step function of pressure, with a consequent -6 dB/octave spectrum shape, but its effect is very brief and the resultant excitation merges with the turbulent noise at the point of constriction, which normally follows the release. Note that although there is no concept of VOT for fricatives, the length of the friction is an important cue for perceiving such consonants.

In connected speech, muscular control is used to bring all of these sound sources into play. This is accomplished with just the right timing for them to combine, in association with the appropriate dimensions of the resonant system, to produce the complex sequence of sounds that we recognize as a linguistic message. For many sounds (such as [v, z]), voiced excitation from the vocal folds occurs simultaneously with turbulent excitation. It is also possible to have turbulence generated in the larynx during vowel production to achieve a breathy voice quality. This quality is produced by not closing the arytenoids quite so much as in normal phonation, and by generating the vibration with a greater air flow from the lungs. There will then be sufficient random noise from air turbulence in the glottis combined with the periodic modulation of the air flow to produce a characteristic breathiness that is common for some speakers. If this effect is taken to extremes, a slightly larger glottal opening, tense vocal folds and more flow will not produce any phonation, but there will then be enough turbulence at the larynx to produce whispered speech.

Comprehensive and classical descriptions of the source mechanisms in speech production can be found in references.³⁻⁵

3. Speech Perception Mechanisms

3.1. *Hearing Organs*

Figure 4 illustrates the structure of the human ear, which is the “front-end” of hearing and speech perception. The outer ear consists of the pinna, which is the visible structure of the ear, and the passage known as the auditory canal. The ear canal, an air-filled passageway, is open to the outside world at one end. At its internal end, it is closed off by the eardrum (the tympanic membrane). Acoustic waves falling on the external ear travel down the auditory canal to reach the eardrum, or tympanic membrane, where the pinna plays a significant role of

collecting the sound due to the effect of reflections from the structures of the pinna. This effect is confined to frequencies above 3 kHz as it is only at these high frequencies that the wavelength of the sound is short enough for it to interact with the structures of the pinna. An outstanding enhancement by the pinna is seen in frequency region around 6 kHz. The pinna also plays a role in judging the direction of a sound source, especially in the vertical direction.

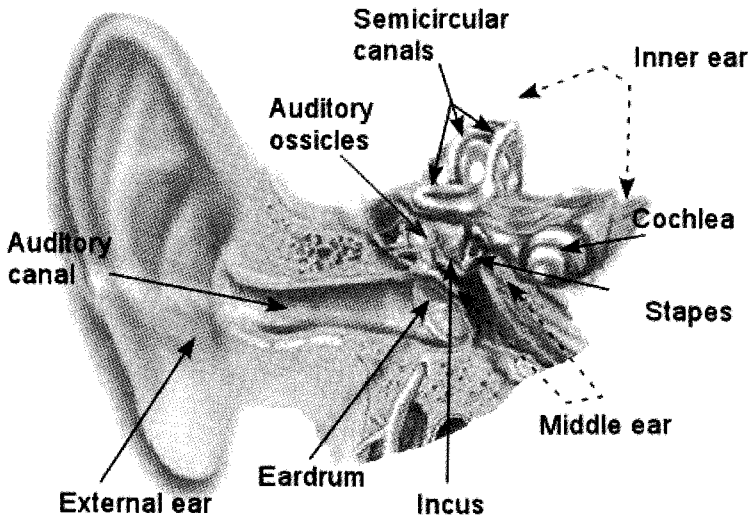


Fig. 4. A view of the outer, middle, and inner ear.

The auditory canal, with a length slightly more than two centimeters, forms an acoustic resonator with a rather heavily damped main resonance at about 3.5 kHz that is corresponding to its length. Some slight secondary resonances take place at higher frequencies. The principal effect of this resonant behavior is to increase the ear's sensitivity to sounds in the 3–4 kHz range. Thus, the pressure at the eardrum for tones near this resonance may be as much as 10 times greater than the pressure at the entrance to the ear canal. This effect enables us to detect sounds that would be imperceptible if the eardrum were located at the surface of the head.

The eardrum is driven by the impinging sound to vibrate. The vibrations are transmitted through the middle ear by the malleus (hammer) to the incus (anvil) which, in turn, is connected to the stapes (stirrup). The footplate of the stapes covers the oval window, which is the entrance to the fluid-filled cavity that makes up the inner ear. The cochlea is the main structure of the inner ear. The ossicles vibrate with a lever action, and enable the small air pressure changes that

vibrate the eardrum to be coupled effectively to the oval window. In this way the ossicles act as a transformer, to match the low acoustic impedance of the eardrum to the higher impedance of the input to the cochlea.

Although the pinna and the ossicles of the middle ear play an important role in the hearing process, the main function of processing sounds is carried out within the inner ear, also known as the cochlea, and in higher levels of neural processing.

As shown in Figure 5, the cochlea is a coiled, tapered tube containing the auditory branch of the mammalian inner ear. At its end there are the semicircular canals, that has the main function of balance control, and not hearing. Figure 5 shows a section through one turn of the spiral, and it is divided along its length into three parts by two membranes, where the core component is the organ of Corti, the sensory organ of hearing. The three parts are known as the scala vestibuli, the scale media and the scala tympani. The interior of the partition, the scala media, is filled with endolymph, a fluid similar in composition to the fluid within body cells. The scala vestibuli lies on one side of the partition, and the scala tympani on the other. Both regions are filled with perilymph, a fluid similar in composition to the fluid surrounding body cells. The helicotrema, an opening in the partition at the far, or apical, end of the cochlea, allows fluid to pass freely between the two cavities. One of the membranes, Reissner's membrane, is relatively wide, and serves to separate the fluids in the seals media and the scala vestibuli but has little effect acoustically. The other membrane, the basilar membrane (BM), is a vital part of the hearing process. As shown in the figure, the membrane itself occupies only a small proportion of the width of the partition between the scale media and the scala tympani. The remainder of the space is occupied by a bony structure, which supports the organ of Corti along one edge of the BM. Rather surprisingly, as the cochlea becomes narrower towards the helicotrema, the basilar membrane actually becomes wider. In humans, it is typically 0.1 mm wide at the basal end, near the oval window, and is 0.5 mm wide at the apical end, near the helicotrema.

The stapes transmits to the oval window on the outside of the cochlea, which vibrates the perilymph in the scala vestibuli. If the stapes is given an impulsive movement, its immediate effect is to cause a distortion of the basal end of the BM. These initial movements are followed by a traveling wave along the cochlea, with corresponding displacements spreading along the length of the BM. However, the mechanical properties of the membrane in conjunction with its environment cause a resonance effect in the membrane movements; the different frequency components of the traveling wave are transmitted differently, and only

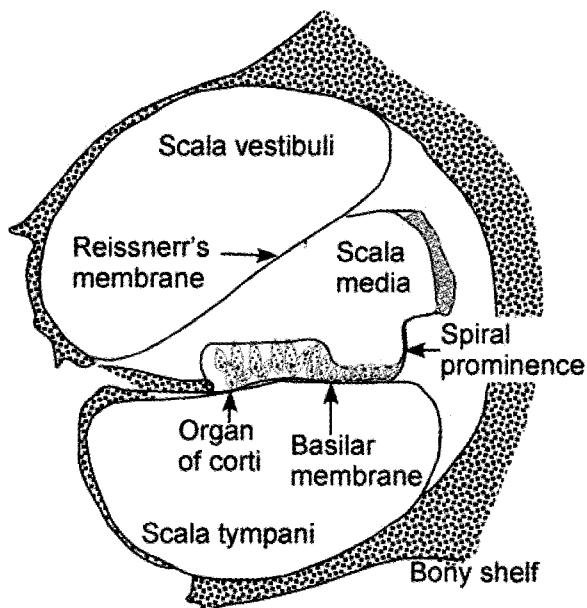


Fig. 5. Cross-section of one turn of the cochlear spiral.

the lowest audio frequency components of the wave cause any significant movement at the apical end. For very low frequencies (below 20 Hz), the pressure waves propagate along the complete route of the cochlea - up the scala vestibuli, around the helicotrema and down the scala tympani to the round window. These low frequencies do not activate the organ of Corti and are below the threshold for hearing. Higher frequencies do not propagate to the helicotrema but are transmitted through the endolymph in the cochlea duct to the perilymph in the scala tympani. Note that a very strong movement of the endolymph due to very loud noise may cause hair cells to die.

For the purpose of hearing, the frequency-selective BM movements must be converted to a neural response. This transduction process takes place by means of the inner hair cells in the organ of Corti. Research over the last couple decades has resulted in a detailed biophysical understanding of how the hearing receptors of the inner ear function. It is the hair cells, particularly the inner hair cells that convert mechanical vibrations of the basilar membrane into electrical signals that are carried by neurons of the auditory nerve to higher levels of the central nervous system. This remarkable process is controlled by a unique structural feature of the hair cell. The bottom of each hair cell rests on the basilar membrane, while the stereocilia extend from the top of the hair cell. The stapes vibrates the endolymph in the scala media, thus causing movements of the hair

bundles of the hair cells, which are acoustic sensor cells that convert vibration into electrical potentials. The hair cells in the organ of Corti are tuned to certain sound frequencies, being responsive to high frequencies near the oval window and to low frequencies near the apex of the cochlea. However, the frequency selectivity is not symmetrical: at frequencies higher than the preferred frequency the response falls off more rapidly than for lower frequencies. The response curves obtained by von Békésy were quite broad, but more recent measurements from living animals have shown that in a normal, healthy ear each point on the BM is in fact sharply tuned, responding with high sensitivity to a limited range of frequencies. The sharpness of the tuning is dependent on the physiological condition of the animal. The sharp tuning is generally believed to be the result of biological structures actively influencing the mechanics of the cochlea. The most likely structures to play this role are the outer hair cells, which are part of the organ of Corti. The magnitude of the BM response does not increase directly in proportion with the input magnitude: although at very low and at high levels the response grows roughly linearly with increasing level, in the mid-range it increases more gradually. This pattern shows a compressive non-linearity, whereby a large range of input sound levels is compressed into a smaller range of BM responses.

The ear as the “front-end” of hearing and speech perception system described above provides the input to the “back-end” or higher levels of the human auditory system. We now provide a functional overview of this “back-end”.

3.2. Perception of Sound

Much of the knowledge we gained in the area of sound perception, including both speech and non-speech sounds, has been based on psychoacoustic experiments where human subjects are exposed to acoustic stimuli through earphones or a loudspeaker, and then these subjects tell something about the sensations these stimuli have produced. For instance, we can expose the subject to an audible sound, gradually decreasing its intensity and ask for an indication of when the sound is no longer audible. Or we may send a complex sound through the headphone to one ear and ask the subject to adjust the frequency of a tone entering the other ear until its pitch is the same as that of the complex sound.

In speech perception, one of the most famous experiments has been that on the McGurk effect.⁶ In the experiment, a video is made by composing a talking head with a sound of “BA”, where the video of the talking head was producing sound “GA”. The subjects are asked to alternate between looking at the talking head while listening, and listening with their eyes closed. Most adult subjects

(98%) think they are hearing “DA” when they are looking at the talking head. However, they are hearing the sound as “BA” when closing their eyes. The above McGurk effect has a few interesting implications regarding sound and speech perception:

- (i) Listening and perception are carried out at different levels. The subjects perceive a sound not only according to the acoustic stimuli but also based on additional information. (Regarding the nature of such “additional information”, some researchers claim that the objects of speech perception are articulatory gestures while others argue that the objects are auditory in nature.)
- (ii) The same general perceptual mechanisms underlie the audio-visual interactions dealing with speech or with other multi-sensory events.
- (iii) The McGurk effect has been considered by the protagonists of the motor theory of speech perception as supporting the idea of a specialized perceptive module, independent of the general auditory module.
- (iv) The McGurk effect is often taken as evidence for gestural approaches to speech perception because it provides an account for why the auditory and visual information are integrated during perception.

3.3. Pitch and Fundamental Frequency

Pitch is defined as the aspect of auditory sensation in terms of which sounds may be ordered on a scale running from “low” to “high”. Pitch is chiefly a function of the frequency of a sound, but it also depends on the intensity and composition. At the sound pressure level (SPL) of 40 dB, the pitch of a 1,000 Hz sinusoid is 1,000 Hz. The half of the pitch is 500 Hz and the twice of the pitch is 2,000 Hz. In the frequency region above 1,000 Hz, the pitch unit (Mel) is almost proportional to the logarithm of the frequencies. That is why the Mel scale is extensively used in many applications of the signal processing such as the Mel Frequency Cepstrum Coefficient (MFCC) in speech analysis and recognition.

Although we have defined a pitch scale in terms of pure tones, it is obvious that more complex sounds, such as musical notes from a clarinet, spoken words, or the roar of a jet engine, also produce a more or less definite pitch. Following the Fourier series, we can consider complex waveforms to be made up of many components, each of which is a pure tone. This collection of components is called a spectrum. For speech sounds, for example, the pitch depends on the frequency of the spectrum’s lowest component. In most cases, the pitch is used to describe the fundamental frequency of a voiced sound, which is the vibration frequency of the vocal folds.

Strictly speaking, the pitch is a subjective quantity, while the fundamental frequency is a physical quantity. They should be used distinctively because they are not always consistent with each other even in their values. As mentioned above, pitch is not only a function of the frequency of a sound, but also dependent on the intensity and composition. This phenomenon is clearly seen in producing and perceiving Chinese tones. For instance, you can produce a low tone (Tone 3) of Mandarin Chinese by lowering the power of the sound instead of the fundamental frequency. Experiments have shown that many foreign students who can pronounce good Chinese tones often manipulate the sound power when generating the low tone.

4. Relationship between Speech Production and Perception Mechanisms

Producing and perceiving speech sounds are basic human activities in speech communication. A speaker can be considered as an encoder in a speech production system and a listener as a decoder to accomplish speech perception. Further, speech information is exchanged not only between the speaker and a listener but also internally within the human brain because a speaker also acts as a listener of his own speech. This is the case especially in acquiring a new language. During such a learning process, the loop in the human brain between speech production and perception must be closed, as a type of internal “speech chain”.

4.1. *Speech Chain*

The concept of a speech chain⁷ is a good place to start thinking about speech and the relationship between its production and perception. This chain includes the following links in which a thought is expressed in different forms as it is born in a speaker’s mind and eventually giving rise to its understanding in a listener’s mind. The various components of the speech chain are illustrated in Figure 6.

In the speech production side of the speech chain, the speaker first decides to say something. This event takes place in the higher level of the mind/brain. The desired thought passes through the language center(s) of the brain where it is given expression in words which are assembled together in the proper order and given final phonetic, intonational, and durational form. The articulation-motor center of the brain is planning a speech motor program which executes over time by conveying firing sequences to the lower neurological level, which in turn impart motion to all of the muscles responsible for speech production:

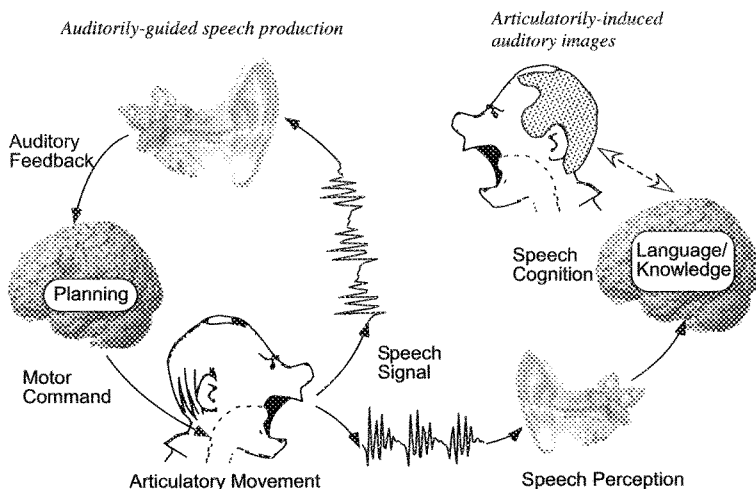


Fig. 6. The speech chain linking speaker and listener.

the muscle contractions, a stream of air emerges from the lungs, passes through the vocal cords where a phonation type (e.g. normal voicing, whispering, aspiration, creaky voice, or even no shaping at all) is developed, and receives its final shape in the vocal tract before radiating from the mouth and the nose and through the yielding wall of the vocal tract. The vibrations caused by the vocal apparatus of the speaker radiate through the air as a sound wave.

In the the side of speech perception, the sound wave eventually strikes the eardrums of listeners as well as of the speaker himself/herself, and is first converted to mechanical vibration on the surface of the tympanum membranes, and then is transformed to fluid pressure waves in the medium via the ossicles of the middle ear, which bathes the basilar membrane of the inner ear, and finally to firings in the neural fibers which combine to form the auditory nerve. The lower centers of the brainstem, the thalamus, the auditory cortex, and the language centers of the brain all cooperate in the recognition of the phonemes which convey meaning, the intonational and durational contours which provide additional information, and the vocal quality which allows the listener to recognize who is speaking and gain insight into the speaker's health, emotional state, and intention in speaking. The higher centers of the brain, both conscious and subconscious, bring to this incoming auditory and language data, all the experience of the listener in the form of previous memories and understanding of the current context, allowing the listener to "manufacture" in his or her mind a more or less faithful "replica" of the thought which was originally formulated in the speaker's consciousness and to update the listener's description of the current

state of the world. The listener may in turn become the speaker, and vice versa, and the speech chain will then operate in reverse.

Actually, in speech communication, a speaker is also playing the role of a listener, which forms another loop besides that between the speaker and the listener as shown in Figure 6. For many years, a number of experiments have been conducted to investigate such a relationship between speech production and perception in order to explore the nature of the speech chain.

4.2. Hypothesis of Articulatory-Auditory Linkage

Speech production can be considered as a forward process in the speech chain while speech perception is an inverse process. In these two processes, speech production and perception are interlinked with each other at both the motor planning and acoustic transmission levels. In addition to the traditional use of electromyography (EMG), a number of new technologies such as functional Magnetic Resonance Imaging (fMRI), Magnetoencephalography (MEG), and Positron Emission Tomography (PET) were developed in the last decades to discover the brain functions in speech production and perception. The research in this area has found curious resemblance of motor and acoustic patterns of the vowel sounds in speech. This not only provides physiological evidence that the motor pattern of vowel articulation is compatible with the acoustic pattern, but also suggests an important aspect of high-level speech organization in the brain: Vowel representations in the motor and auditory spaces are also compatible.

As a general assumption, the high-level motor organization utilizes sensory information as a guide to motor control. This sort of sensori-motor integration typically accounts for visually-aided hand movements, where the object's coordinates in the visual space are mapped into kinematic parameters in the motor space. Although speech articulation does not always undergo the same scheme, there may be a case for the claim that articulatory control is guided with reference to auditory input. The assumption of motor-to-auditory compatibility may not be unreasonable in such a situation. Therefore, at least as far as vowels are concerned, the compatibility of the sensori-motor patterns can be considered as a unique underlying characteristic of speech processes. This view provides an extended account for the earlier motor theory of speech perception.^{8,9}

In summary, human speech processing is a closed-loop control in which articulatorily-induced auditory images and sensory-guided motor execution communicate in the brain during speech. Due to this linkage, humans have a robust capability to adapt them in adverse environments and to extract the relevant linguistic and paralinguistic information. Many engineering techniques

for analyzing speech have been motivated, directly or indirectly, by the speech production and/or perception mechanisms discussed in this section.

5. Introduction to Speech Analysis

The previous sections in this chapter focused on how humans process speech. The remaining sections will address one area of applications where computers are used to automatically extract useful information from the speech signal. This is called speech analysis. One main theme of this chapter is that the commonly used speech analysis techniques, which we will describe in the remainder of this chapter, can be provided with a new angle based on how they are motivated by and corrected to the human speech production and/or perception mechanisms. The basic material in the remaining sections of this chapter comes from references [10,11] with extensive re-organization and re-writing.

Speech analysis is sometimes called (linguistically independent) speech parameter or feature extraction, which is of direct use to other areas of speech processing applications such as coding and recognition. From the knowledge gained in human production that articulators are continuously moving over time, it is clear that a speech signal changes its characteristics continuously over time as well, especially when a new sound is encountered. Therefore, speech analysis cannot be performed over the entire stretch of time. Rather, it is carried out on short windowed segments during which the articulators and the resulting vocal tract shape are relatively stable. We call this type of analysis, motivated directly by the nature of speech production, short-time analysis.

Within each window, the analysis proceeds in either the time or frequency domain, or by a parametric model. In this section, we discuss only the time-domain analysis, deferring other more intricate analyses to later sections.

5.1. Short-Time Analysis Windows

Speech is dynamic or time-varying. Sometimes, both the vocal tract shape and pertinent aspects of its excitation may stay fairly constant for dozens of pitch periods (e.g., to 200 ms). On the other hand, successive pitch periods may change so much that their name “period” is a misnomer. Since the typical phone averages only about 80 ms in duration, dynamic coarticulation changes are more the norm than steady-state sounds. In any event, speech analysis usually presumes that signal properties change relatively slowly over time. This is most valid for short time intervals of a few periods at most. During such a short-time window of speech, one extracts parameters or features, each representing an

average over the duration of the time window. As a result of the dynamic nature of speech, we must divide the signal into many successive windows or analysis frames, allowing the parameters to be calculated frequently enough to model dynamic vocal-tract features. Window size is critical to good modeling. Long vowels may allow window lengths up to 100 ms with minimal loss of detail due to the averaging, but stop explosions require much shorter windows (e.g., 5–10 ms) to avoid excess averaging of rapid spectral transitions. In a compromise, typical windows last about 20–30 ms, since one does not know *a priori* what sound one is analyzing.

Windowing means multiplication of a speech signal by a window, yielding a set of new speech samples weighted by the shape of the window. The simplest window has a rectangular shape, which gives equal weight to all samples of the speech signal and limits the analysis range to width of the window. A common is the Hamming window, which is a raised cosine pulse, or the quite similar Hanning window. Tapering the edges of the window allows its periodic shifting (at the update frame rate) along without having large effects on the resulting speech parameters due to pitch period boundaries.

5.2. Short-Time Average Zero-Crossing Rate

Speech analysis that attempts to estimate spectral features usually requires a Fourier transform. However, a simple measure called the zero-crossing rate (ZCR) provides basic spectral information in some applications at low cost. For a speech signal, zero crossing takes place whenever the waveform crosses the time axis (i.e., changes algebraic sign). For all narrowband signals (e.g., sinusoids), the ZCR can accurately measure the frequency where power is concentrated.

The ZCR is useful for estimating whether speech is voiced. Voiced speech has mostly low-frequency power, owing to a glottal excitation spectrum that falls off at about -12 dB per octave. Unvoiced speech comes from broadband noise excitation exciting primarily high frequencies, owing to the use of shorter vocal tracts (anterior to the constriction where noise is produced). Since speech is not narrowband, the ZCR corresponds to the average frequency of primary power concentration. Thus high and low ZCR (about 4,900 and 1,400 crossings/sec) correspond to unvoiced and voiced speech, respectively.

5.3. Short-Time Autocorrelation Function

Another way to estimate certain useful features of a speech signal concerns the short-time autocorrelation function. Like the ZCR, it serves as a tool to access

some spectral characteristics of speech without explicit spectral transformations. As such, the autocorrelation function has applications in F0 estimation, voiced/unvoiced determination, and linear prediction. In particular, it preserves spectral amplitude information in the speech signal concerning harmonics and formants, while suppressing (often undesired) phase effects.

For F0 estimation, an alternative to the autocorrelation is the average magnitude difference function (AMDF). The AMDF has minima for the interval value near multiples of the pitch period (instead of the peaks for the autocorrelation function).

6. Speech Analysis Based on Production Mechanisms

6.1. Introduction to LPC

The most prominent method in speech analysis, linear predictive coding (LPC), has been directly motivated by our understanding of the physical properties of the human speech production system. LPC analysis has had a successful history for more than 30 years. The term “linear prediction” refers to the mechanism of using a linear combination of the past time-domain samples to approximate or to “predict” the current time-domain sample, using a compact set of LPC coefficients. If the prediction is accurate, then these coefficients can be used to efficiently represent or to “code” a long sequence of the signal (within each window). Linear prediction expressed in the time-domain is mathematically equivalent to modeling of an all-pole resonance system. This type of resonance system has been a rather accurate model for the vocal tract when vowel sounds are produced.

LPC is the most common technique for low-bit-rate speech coding and its popularity derives from its simple computation and reasonably accurate representation of many types of speech signals. LPC as a speech analysis tool has also been used in estimating F0, formants, and vocal tract area functions. One drawback of LPC is its omission of the zero components in several types of speech sounds with glottal source excitation and multiple acoustic paths in nasals and unvoiced sounds.

The most important aspect of the LPC analysis is to estimate the LPC coefficients from each of the windowed speech waveforms. The estimation technique has been well-established and can be found in any standard textbook on speech processing. Briefly, a “normal equation” is established using the least-squares criterion. Then highly-efficient methods are used to solve it to obtain the estimated LPC coefficients.

6.2. Choice of the LPC Order

The choice of the LPC order in speech analysis reflects a compromise of representation accuracy, computation time, and memory requirement. It can be shown that a perfect representation can be achieved in the limiting case where the order approaches infinity. This is only of theoretical interest, since the order rarely goes very high, due to increasingly excessive cost of computation. In practical applications, the LPC order is chosen to assign enough model poles to represent all formants (at two poles per resonance) in the bandwidth of the input speech signal. An additional 2–4 poles are usually assigned (e.g., the standard for 8 kHz sampled speech is 10 poles) to account for windowing effects and for weaknesses in the all-pole model. The all-pole model ignores zeros and assumes an infinitely-long stationary speech sound; thus assigning only enough poles to model the expected number of formants risks the case that poles may be used by the model to handle non-formant effects in the windowed spectrum (such is seen often in LPC modeling). The non-formant effects derive mostly from the vocal-tract excitation (both glottal and fricative) and from lip radiation. In addition, zeros are regularly found in nasalized sounds. Nasal sounds theoretically have more resonances than vowels, but we rarely increase the LPC order to handle nasals, because most nasals have more than one resonance with little energy (due to the effects of zeros and losses).

The prediction error energy can serve as a measure of the accuracy of an LPC model. The normalized prediction error (i.e., the energy in the error divided by the speech energy) decreases monotonically with LPC order (i.e., each additional pole in the LPC model improves its accuracy). With voiced speech, poles in excess of the number needed to model all formants (and a few for zero effects) add little to modeling accuracy but such extraneous poles add increasingly to the computation.

6.3. Introduction to Voicing Pitch Extraction

Fundamental frequency F_0 or pitch parameters associated with voiced speech characterize the most important aspect of the source mechanism in human speech production. F_0 is especially relevant to tone languages such as spoken Chinese, since it provides the informational basis for linguistic and lexical contrast.

Although automatic F_0 estimation appears fairly simple at first glance, full accuracy has so far been elusive, owing to the non-stationary nature of speech, irregularities in vocal cord vibration, the wide range of possible F_0 values, interaction of F_0 with vocal tract shape, and degraded speech in noisy environments. F_0 can be estimated either from periodicity in the time-domain or

from harmonic spacing in frequency. Spectral approaches generally have higher accuracy than time-domain methods, but they need more computation.

Because the major excitations of the vocal tract occur when the vocal cords close for a pitch period, each period starts with high amplitude and then has an amplitude envelope that decays exponentially with time. Since the very lowest frequencies dominate power in voiced speech, the overall rate of decay is usually inversely proportional to the bandwidth of the first formant. The basic method for pitch period estimation is a simple search for amplitude peaks, constraining the peak-to-peak interval to be consistent in time (since F0 varies slowly as constrained by articulators). Because speakers can range from infants to adult males, a large pitch-period range from about 2 ms to 20 ms is possible.

Input speech is often low-pass-filtered to approximately 900 Hz so as to retain only the first formant, thus removing the influence of other formants, and simplifying the signal, while retaining enough harmonics to facilitate peak-picking. F0 estimation in the time domain has two advantages: efficient calculation and direct specification of pitch periods in the waveform. This is useful for applications when pitch periods need to be manipulated. On the other hand, F0 values alone (without explicit determination of the placement of pitch periods) suffice for many applications, such as vocoders.

When F0 is estimated spectrally, the fundamental itself and, more often, its equally-spaced harmonics can furnish the main clues. In time-domain peak-picking, errors may be made due to peaks corresponding to formants (especially F1), misinterpreting the waveform oscillations due to the F1 as F0 phenomena. Spacing between harmonics is usually more reliable as an F0 cue. Estimating F0 directly in terms of the lowest spectral peak in the speech signal can be unreliable because the speech signal is often bandpass filtered (e.g., over telephone lines). Even unfiltered speech has a weak first harmonic when F1 is at high frequency (as in low vowels). While often yielding more accurate estimates than time-domain methods, spectral F0 detectors require much more calculation due to the required spectral transformation. Typical errors include: 1) misjudging the second harmonic as the fundamental and 2) the ambiguity of the estimated F0 during aperiodicities such as in voice creak. A given estimation method often performs well on some types of voice but less so for other types.

6.4. *Pitch Estimation Methods*

We estimate F0 either from periodicity in the time domain or from regularly-spaced harmonics in the frequency domain. Like many pattern recognition algorithms, most pitch estimators have three components: a preprocessor to

simplify the input signal (eliminate information in the signal that is irrelevant to F0), a basic F0 extractor to form the F0 estimate, and a postprocessor to correct errors. The preprocessor serves to focus the remaining data towards the specific task of F0 determination, reducing data rates by eliminating much formant detail. Since the basic pitch estimator, like all pattern recognizers, makes errors, a postprocessor may help to clean up the time series of output pitch estimates (one per frame), e.g., imposing continuity constraints from speech production theory, which may not have been applied in the basic F0 extractor often operating independently on each speech frame.

The pitch detection algorithm tries to locate one or more of the following features in the speech signal or in its spectrum: the fundamental harmonic F0, a quasi-periodic time structure, an alternation of high and low amplitudes, and signal discontinuities. The intuitive approach of looking for harmonics and periodicities usually works well, but fails too often to be relied upon without additional support. In general, pitch detectors trade off complexities in various components; e.g., harmonic estimation requires a complex preprocessor (e.g., often including a Fourier transform) but allows a simple basic extractor that just does peak-picking. The preprocessor is often just a low-pass filter, but the choice of the filter's cutoff frequency can be complicated by the large range of F0 values possible when accepting speech from many different speakers.

Frequency-domain methods for pitch detection exploit correlation, maximum likelihood, and other spectral techniques where speech is analyzed during a short-term window for each input frame.

Autocorrelation, average magnitude difference, cepstrum, spectral compression, and harmonic-matching methods are among the varied spectral approaches. Spectral methods generally have greater accuracy than time-domain methods, but require more computation.

To provide results in real-time for many applications, F0 estimators must work with little delay. Delays normally arise in part from the use of a buffer to accumulate a large frame of speech to analyze, since pitch can only be detected over intervals corresponding to pitch periods (unlike spectral envelope information, like formants, which can succeed with much shorter analysis frames). Time-domain F0 estimators often incur less delay than frequency-domain methods. The latter mostly require a buffer of speech samples prior to their spectral transformation. Many of the F0 detectors with less delay sacrifice knowledge about the timing of the pitch periods; i.e., they estimate the lengths of pitch periods without explicitly finding their actual locations. While most F0 estimators do not need to locate period times, those that do locate them are more useful in certain applications (e.g., to permit pitch-synchronous LPC analysis).

7. Speech Analysis Methods Based on Perception Mechanisms

7.1. Filter Bank Analysis

As we have seen from Section 1.3, one basic function of the ear is to decompose the impinging sounds into a bank of outputs along the BM spatial dimension. Each point along this dimension is a filter with band-pass-like frequency selectivity. The outputs from the BM are analogous to those in the traditional filter bank analysis technique.

Filter bank analysis consists of a set of band-pass filters. A single input speech signal is simultaneously passed through these filters, each outputting a narrowband signal containing amplitude (and sometimes phase) information about the speech in a narrow frequency range. The bandwidths normally are chosen to increase with center frequency, thus following decreasing human auditory resolution. They often follow the auditory Mel scale, i.e., having equally-spaced, fixed bandwidths below 1 kHz, then logarithmic spacing at higher frequencies.

Such a filter-bank analysis tries to simulate very simple aspects of the human auditory system, based on the assumption that human signal processing is an efficient way to do speech analysis and recognition. Since the inner ear apparently transfers information to the auditory nerve on a spectral basis, the filter banks discussed above approximate this transfer quite roughly.

7.2. Auditory-Motivated Speech Analysis

Other speech analysis methods go further than filter-bank analysis in the hope that improved approximation to perception mechanisms may occur. Many of these alternative approaches use auditory models where the filtering follows that found in the cochlea more precisely.

One example of auditory-motivated analysis is that of modulation spectrogram, which emphasizes slowly-varying speech changes corresponding to a rate of approximately 4 per second. Such 250 ms units conform roughly to syllables, which are important units for perceptual organization. Modulation spectrogram displays show less rapid detail than standard wideband spectrograms. In a sense, they follow the idea of wavelets, which allow time and frequency resolution in automatic analysis to follow that of the ear.

Part of the auditory processing of sounds is adjusting for context. Since human audition is always active, even when people are not specifically listening, it is normal to ignore ambient sounds. People pay attention to unexpected auditory information, while ignoring many predictable and thus useless aspects

of sounds. These adjustments come naturally to humans, as part of the maturation of their hearing systems. In computer sound processing, however, one must explicitly model this behavior. A number of speech analysis methods used successfully in environment-robust automatic speech recognition do simulate the human auditory mechanisms to adjust to the acoustic environment (e.g., noise, use of telephone, poor microphone or loudspeaker).

Since automatic speech recognition involves comparing patterns or models, environmental variations can cause major acoustic differences which are superfluous for the recognition decision, and which human audition normalizes for automatically. Basic speech analysis methods such as filter-bank analysis, LPC, and cepstra, cannot execute such normalization. The filtering effects of RASTA provide one way to try to implement normalization and to improve the results of analysis for noisy speech. In another common approach, a mean spectrum or cepstrum is subtracted from that of each speech frame (e.g., as a form of blind deconvolution), to eliminate channel effects. It is unclear over what time range this mean should be calculated. In practice, it is often done over several seconds, on the assumption that environmental conditions do not change more rapidly. This, however, may impose a delay on the speech analysis; so the channel conditions are sometimes estimated from prior frames and imposed on future ones in their analysis. Calculating such a mean may require a long-term average for efficiency, which is difficult for real-time applications. Often the mean is estimated from a prior section of the input signal that is estimated to be (noisy) silence (i.e., non-speech). This latter approach requires a speech detector and assumes that pauses occur fairly regularly in the speech signal. As the channel changes with time, the mean must be updated.

Speech analysis methods that can simulate auditory properties and normalize the acoustic environments accordingly are still an active research area.

8. Speech Analysis Methods Based on Joint Production-Perception Mechanisms

8.1. *Perceptual Linear Prediction Analysis*

As we discussed earlier in this chapter, LPC analysis is based primarily on speech production mechanisms, where the all-pole property of the human vocal tract is used as the basic principle to derive the analysis method. It can be shown that LPC analysis is equivalent to a type of spectral analysis where the linear frequency scale is involved.

However, auditory processing of sounds in the ear exploits the Mel instead of linear frequency scale. Further, various nonlinear properties in the ear are not

exploited in the LPC analysis. Some of these auditory properties have been incorporated into the LPC analysis, resulting in the analysis method called perceptual linear prediction (PLP). PLP modifies the basic LPC using a critical-band or Mel power spectrum with logarithmic amplitude compression. The spectrum is multiplied by a mathematical curve modeling the ear's behavior in judging loudness as a function of frequency. The output is then raised to the power 0.33 to simulate the power law of hearing. Seventeen bandpass filters equally spaced in Bark or Mel scale (i.e., critical bands) map the range 0–5 kHz, for example, into the range 0–17 Bark. Each band is simulated by a spectral weighting. One direct advantage of the PLP is that its order is significantly less than orders generally used in LPC.

PLP has often been combined with the RASTA (RelAtive SpecTrAl) method of speech analysis. RASTA bandpasses spectral parameter signals to eliminate steady or slowly-varying components in the speech signal (including environmental effects and speaker characteristics) and rapid noise events. The bandpass range is typically 1–10 Hz, with a sharp zero at 0 Hz and a time-constant of about 160 ms. Events changing more slowly than once a second (e.g., most channel effects, except in severe fading conditions) are thus eliminated by the highpass filtering. The lowpass cutoff is more gradual, which smoothes parameters over about 40 ms, thus preserving most phonetic events, while suppressing noise.

8.2. Mel-Frequency Cepstral Analysis

Another popular speech analysis method that jointly exploits speech production and perception mechanisms is Mel-frequency cepstral analysis. The results of this analysis are often called Mel-frequency cepstral coefficients (MFCCs). To introduce MFCCs, we first introduce the (linear frequency) cepstral analysis method which has been primarily motivated by speech production knowledge.

The basic speech-production model is a linear system (representing the vocal tract) excited by quasi-periodic (vocal-cord) pulses or random noise (at a vocal-tract constriction). Thus the speech signal, as the output of this linear system, is the convolution of an excitation waveform with the vocal-tract's impulse response. Many speech applications require separate estimation of these individual components; hence a deconvolution of the excitation and envelope components is useful. Producing two signals from one in such a deconvolution is generally nondeterministic, but has some success when applied to speech because the relevant convolved signals forming speech have a unique time-frequency behavior.

Cepstral analysis or cepstral deconvolution converts a product of two spectra into a sum of two signals. These may be separated by linear filtering if they are sufficiently different. Let speech spectrum be $S = EH$, where E and H represent the excitation and vocal-tract spectra, respectively. Then, $\log S = \log (EH) = \log (E) + \log (H)$. Since H consists mostly of smooth formant curves (i.e., a spectrum varying slowly with frequency) while E is much more active or irregular (owing to the harmonics or noise excitation), contributions due to E and H can be separated. Inverse Fourier transform of log spectrum ($\log S$) gives the (linear frequency) cepstrum. It consists of two largely separable components, one due to the vocal tract excitation and another due to the vocal tract transfer function.

To incorporate the perceptual mechanism of Mel-spaced frequency analysis, the (linear frequency) cepstral analysis has been modified to produce MFCCs, which are widely used as the main analysis method for speech recognition. MFCCs combine the regular cepstrum with a nonlinear weighting in frequency, following the Bark or Mel scale so as to incorporate some aspects of audition. It appears to furnish a more efficient representation of speech spectra than other analysis methods such as LPC.

In computing MFCCs, a power spectrum of each successive speech frame is first effectively deformed both in frequency according to the Mel scale and in amplitude on the usual decibel or logarithmic scale. The power spectrum can be computed either via the Fourier transform or via the LPC analysis. Then the initial elements of an inverse Fourier transform are obtained as the MFCCs, with varying orders from the zero-th order to typically a 15th order.

The zero-th order MFCC simply represents the average speech power. Because such power varies significantly with microphone placement and communication channel conditions, it is often not directly utilized in speech recognition, although its temporal derivative often is. The next first-order MFCC indicates the balance of power between low and high frequencies, where a positive value indicates a sonorant sound, and a negative value a frication sound. This reflects the fact that sonorants have most energy at low frequencies, and fricatives the opposite. Each higher-order MFCC represents increasingly finer spectral detail. Note that neither MFCCs nor LPC coefficients display a simple relationship with basic spectral envelope detail such as the formants. For example, using a speech bandwidth with four formants (e.g., 0–4 kHz), a high value for the second-order MFCC corresponds to high power in the F1 and F3 ranges but low amounts in F2 and F4 regions. Such information is useful to distinguish voiced sounds, but it is difficult to interpret physically.

8.3. *Formants and their Automatic Extraction*

The main objective of speech analysis is to automatically extract essential parameters of the acoustic structure from the speech signal. This process serves the purpose of either data reduction or identification and enhancement of information-bearing elements contained in the speech signal. To determine what such information-bearing elements are, one needs to have sufficient knowledge about the physical nature of the speech signal. Such knowledge is often provided by a production model that describes how the observed speech signal is generated as the output of a vocal-tract digital filter given an input source. This type of production model decomposes the speech waveform into two independent source and filter components. Formants comprise a very important set of the information-bearing elements in view of the source-filter model, since they form the resonance peaks in the filter component of the model. At the same time, formants are also the information-bearing elements in view of speech perception, because the auditory system not only robustly represents such information, but also exploits it for distinguishing different speech sounds. Before describing the formant extraction methods, we first elaborate the concept of formants in detail.

Formants characterize the “filter” portion of a speech signal. They are the poles in the digital resonance filter or digital resonator. Given the source-filter model for voiced speech that is free of nasal coupling, the all-pole filter is characterized by the pole positions, or equivalently by the formant frequencies, F_1, F_2, \dots, F_n , formant bandwidths, B_1, B_2, \dots, B_n , and formant amplitudes, A_1, A_2, \dots, A_n . Among them, the formant frequencies or resonance frequencies, at which the spectral peaks are located, are the most important. A formant frequency is determined by the angle of the corresponding pole in the discrete-time filter transfer function.

The normal range of the formant frequencies for adult males is $F_1 = 180\text{--}800$ Hz, $F_2 = 600\text{--}2500$ Hz, $F_3 = 1200\text{--}3500$ Hz, and $F_4 = 2300\text{--}4000$ Hz. These ranges have been exploited to provide constraints for automatic formant extraction and tracking.

The average difference between adjacent formants of adult males is about 1,000 Hz. For adult females, the formant frequencies are about 20% higher than for adult males. The relationship between male and female formant frequencies, however, is not uniform and the relationship deviates from a simple scale factor. When the velum is lowered to create nasal phonemes, the combined nasal+vocal tract is effectively lengthened from its typical 17 cm vocal-tract length by about 25%. As a result, the average spacing between formants reduces to about 800 Hz.

Formant bandwidths are physically related to energy loss in the vocal tract, and are determined by the distance between the pole location and the origin of the z -plane in the filter transfer function. Empirical measurement data from speech suggest that the formant bandwidths and frequencies are systematically related. Formant amplitudes, on the other hand, vary with the overall pattern of formant frequencies as a whole. They are also related to the spectral properties of the voice source.

LPC analysis models voiced speech as the output of an all-pole filter in response to a simple sequence of excitation pulses. In addition to major speech coding and recognition applications, LPC is often used as a standard formant extraction and tracking method. It has limitations in that the vocal-tract filter transfer function, in addition to having formant poles (which are of primary interest for speech analysis), generally also contains zeros due to sources located above the glottis and to nasal and subglottal coupling. Furthermore, the model for the voice source as a simple sequence of excitation impulses is inaccurate, with the source actually often containing local spectral peaks and valleys. These factors often hinder accuracy in automatic formant extraction and tracking methods based on LPC analysis. The situation is especially serious for speech with high pitch frequencies, where the automatic formant-estimation method tends to pick harmonic frequencies rather than formant frequencies. Jumps from a correctly-estimated formant in one time frame to a higher or a lower value in the next frame constitute one common type of tracking error.

The automatic tracking of formants is not trivial. The factors rendering the formant identification process complex include the following. The ranges for formant center-frequencies are large, with significant overlaps both within and across speakers. In phoneme sequences consisting only of oral vowels and sonorants, formants smoothly rise and fall, and are easily estimated via spectral peak-picking. However, nasal sounds cause acoustic coupling of the oral and nasal tracts, which lead to abrupt formant movements. Zeros (due to the glottal source excitation or to the vocal tract response for lateral or nasalized sounds) also may obscure formants in spectral displays. When two formants approach each other, they sometimes appear as one spectral peak (e.g., F1–F2 in back vowels). In obstruent sound production, a varying range of low frequencies is only weakly excited, leading to a reduced number of formants appearing in the output speech.

Given a spectral representation $S(z)$ via the LPC coefficients, one could directly locate formants by solving for the roots of the denominator polynomial in $S(z)$. Each complex-conjugate pair of roots would correspond to a formant if the roots correspond to a suitable bandwidth (e.g., 100–200 Hz) at a frequency

location where a formant would normally be expected. This process is usually very precise, but quite expensive since the polynomial usually requires an order in excess of 10 to represent 4–5 formants. Alternatively, one can use phase to label a spectral peak as a formant. When evaluating $S(z)$ on the unit circle a negative phase shift of approximately 180 degrees should occur as the radiant frequency passes a pole close to the unit circle (i.e., a formant pole).

Two close formants often appear as a single broad spectral peak, a situation that causes many formant estimators difficulty, in determining whether the peak corresponds to one or two resonances. A method called the chirp-z transform has been used to resolve this issue.

Formant estimation is increasingly difficult for voices with high F_0 , as in children's voices. In such cases, F_0 often exceeds formant bandwidths, and harmonics are so widely separated that only one or two make up each formant. A spectral analyzer, traditionally working independently on one speech frame at a time, would often equate the strongest harmonics as formants. Human perception, integrating speech over many frames, is capable of properly separating F_0 and the spectral envelope (formants), but simpler computer analysis techniques often fail. It is wrong to label a multiple of F_0 as a formant center frequency, except for the few cases where the formant aligns exactly with a multiple of F_0 (such alignment is common in songs, but much less so in speech).

9. Conclusion

In this chapter we have provided an overview of speech analysis from the perspectives of both speech production and perception, emphasizing their relationship. The chapter starts with a general introduction to the phonological and phonetic properties of spoken Mandarin Chinese. This is followed by descriptions of human speech production and perception mechanisms. In particular, we present some recent brain research on the relationship between human speech production and perception. While the traditional textbook treatment of speech analysis is from the perspective of signal processing and feature extraction, in this chapter we take an alternative, more scientifically-oriented approach in treating the same subject, classifying the commonly-used speech analysis methods into those that are more closely linked with speech production, speech perception, and joint production/perception mechanisms. We hope that our approach to treating speech analysis can provide a fresh view of this classical yet critical subject and help readers understand the basic signal properties of spoken Chinese that are important in exploring further materials in this book.

References

1. M. Chiba, and T., Kajiyama, *The Vowel: Its Nature and Structure*, Phonetic Society of Japan.
2. G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, (1960).
3. R. Titze, "On the Mechanics of Vocal Fold Vibration," *J. Acoust. Soc. Am.*, vol. 60, (1976), pp. 1366-1380.
4. K. N. Stevens, "Physics of Laryngeal Behavior and Larynx Modes," *Phonetica*, vol. 34, (1977), pp. 264-279.
5. K. Ishizaka and J. Flanagan, "Synthesis of Voiced Sounds from a Two-mass Model of the Vocal cords," *Bell Sys. Tech. J.*, vol. 512, (1972), pp. 1233-1268.
6. H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, (1976), pp. 746-748.
7. P. Denes and E. Pinson, *The Speech Chain*, 2nd Ed., New York: W.H. Freeman and Co., (1993).
8. A. M. Liberman and I. G. Mattingly, "The Motor Theory of Speech Perception Revised," *Cognition*, vol. 21, (1985), pp. 1-36.
9. J. Dang, M. Akagi and K. Honda, "Communication between Speech Production and Perception within the Brain -Observation and Simulation-," *J. Computer Science and Technology*, vol. 21, (2006), pp. 95-105.
10. L. Deng and D. O'shaughnessy, *Speech Processing --- A dynamic and optimization-oriented approach*, Mercel-Dekker, New York, (2003).
11. D. O'shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, Piscataway, NJ, (2000).