from Speaker Classification II: Selected Papers, Springer Lecture Notes in Artificial Intelligence, 2007.

How is individuality expressed in voice? An introduction to speech production & description for speaker classification

Volker Dellwo, Mark Huckvale, Michael Ashby

Department of Phonetics and Linguistics
University College London
Gower Street, London, WC1E 6BT
United Kingdom
v.dellwo@ucl.ac.uk, m.huckvale@ucl.ac.uk, m.ashby@ucl.ac.uk

Abstract. As well as conveying a message in words and sounds, the speech signal carries information about the speaker's own anatomy, physiology, linguistic experience and mental state. These speaker characteristics are found in speech at all levels of description: from the spectral information in the sounds to the choice of words and utterances themselves. This chapter presents an introduction to speech production and to the phonetic description of speech to facilitate discussion of how speech can be a carrier for speaker characteristics as well as a carrier for messages. The chapter presents an overview of the physical structures of the human vocal tract used in speech, it introduces the standard phonetic classification system for the description of spoken gestures and it presents a catalogue of the different ways in which individuality can be expressed through speech. The chapter ends with a brief description of some applications which require access to information about speaker characteristics in speech.

Keywords: Speech production, Phonetics, Taxonomy, IPA, Individuality, Speaker characteristics.

1. Introduction

Whenever someone speaks an utterance, they communicate not only a message made up of words and sentences which carry meaning, but also information about themselves as a person. Recordings of two people saying the same utterance will sound different because the process of speaking engages the neural, physiological, anatomical and physical systems of a specific individual in a particular circumstance. Since no two people are identical, differences in these systems lead to differences in their speech, even for the same message. The speaker-specific characteristics in the signal can provide information about the speaker's anatomy, physiology, linguistic experience and mental state. This information can sometimes be exploited by listeners and technological applications to describe and classify speakers, possibly allowing speakers to be categorised by age, gender, accent, language, emotion or

health. In circumstances where the speaker is known to the listener, speaker characteristics may be sufficient to select or verify the speaker's identity. This leads to applications in security or forensics. The aim of this chapter is to provide a framework to facilitate discussion of these speaker characteristics: to describe ways in which the individuality of speakers can be expressed through their voices.

Always in the discussion of speaker characteristics, it must be borne in mind that a spoken utterance exists primarily for its communicative value – as an expression of a desire in the mind of the speaker to make changes in the mind of the listener. The study of the communicative value of utterances is the domain of Linguistics, which we take to include knowledge of articulation, phonology, grammar, meaning and language use. The study of speaker characteristics is in a sense parallel to this, where we concentrate on what a particular implementation of an utterance within the linguistic system tells us about the person speaking.

At first glance, it may appear that we should be able to separate speaker characteristics from message characteristics in a speech signal quite easily. There is a view that speaker characteristics are predominantly low level – related to the implementation in a particular physical system of a given set of phonetic gestures, while message characteristics operate at a more abstract level, related to the choice of phonetic gestures: the syllables, words and phrases that are used to communicate the meaning of a message. However this is to oversimplify the situation. Speakers are actually different at all levels, because speakers also differ in the way in which they realise the phonetic gestures, they vary in the inventory of gestures used, in the way in which gestures are modified by context, and in their frequency of use of gestures, words and message structure. A speaker's preferred means of morning greeting may help identify them just as much as their preferred average pitch.

To build a framework in which the many potential influences of an individual on his or her speech can be discussed, we have divided this chapter into three sections: section 2 provides an overview of vocal structures in humans, section 3 introduces the conventional principles of phonetic classification of speech sounds, while section 4 provides a discussion on how and on what levels speaker characteristics find their way into the speech signal and briefly discusses possible applications of this knowledge.

2. Vocal apparatus

In this section we will give an overview of the physical structures in the human that are used in the physical generation of speech sounds. We will look at the anatomy of the structures, their movements and their function in speech. The first three sections look at the structures below the larynx, above the larynx and the larynx itself. The last section briefly introduces the standard signals and systems model of speech acoustics.

2.1. Sub-laryngeal vocal tract

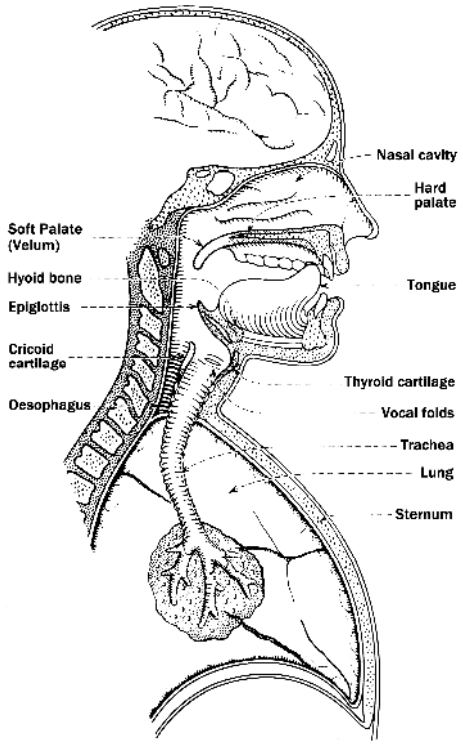


Fig. 1. Schematic diagram of the human organs of speech (Adapted from [1])

Figure 1 shows the main anatomical structures that are involved in speaking. Looking below the larynx we see the lungs lying inside a sealed cavity inside the rib cage. The volume of the air spaces in the lungs can be varied from about 2 litres to about 6 litres in adults. The volume of the chest cavity and hence the volume of the lungs themselves is increased by lowering the diaphragm or raising the rib cage; the volume is decreased by raising the diaphragm or lowering the rib cage. The diaphragm is a dome of muscle, rising into the lower surface of the lungs, and tensing it causes it to flatten out and increase the size of the chest cavity; conversely relaxation of the diaphragm or action of the abdominal wall muscles makes the diaphragm more domed, reducing the size of the cavity. The external intercostal muscles bring the ribs closer together, but since they are pivoted on the vertebrae and are floating at the lower end of the rib cage, contraction of these muscles raises the rib cage and increases the volume of the chest cavity. The internal intercostal muscles can be used to depress the rib cage, and in combination with muscles of the abdominal wall, these can act to forcibly reduce the size of the chest cavity.

Changes in the size of the chest cavity affect the size of the lungs and hence the pressure of the air in the lung cavities. A reduction in pressure draws in air through the mouth or nose, through the pharynx, larynx and trachea into the lungs. A typical inspiratory breath for speech has a volume of about 1.5 litres, and is expended during speech at about 0.15 litres/sec [2]. One breath may be used to produce up to 30 seconds of speech. An increase in the pressure of air in the lungs forces air out through the trachea, larynx, pharynx, mouth and nose. To produce phonation in the larynx, the lung pressure has to rise by at least 300Pa to achieve sufficient flow for vocal fold vibration. A more typical value is 1000Pa, that is 1% of atmospheric pressure.

Pressure is maintained during speech by a control mechanism that connects stretch receptors in the trachea, bronchioles and lung cavities to the muscles that control chest cavity volume. The stretch receptors provide information about the physical extension of the lung tissues which indirectly measures lung pressure. At large volumes the natural elasticity of the lungs would cause too high a pressure for speaking, so nerve activation on the diaphragm and external intercostal muscles is required to maintain a lower pressure, while at low volumes the elasticity is insufficient to maintain the pressure required for speaking, so nerve activation on the internal intercostal muscles and abdominal wall muscles is required to maintain a higher pressure.

2.2. The larynx

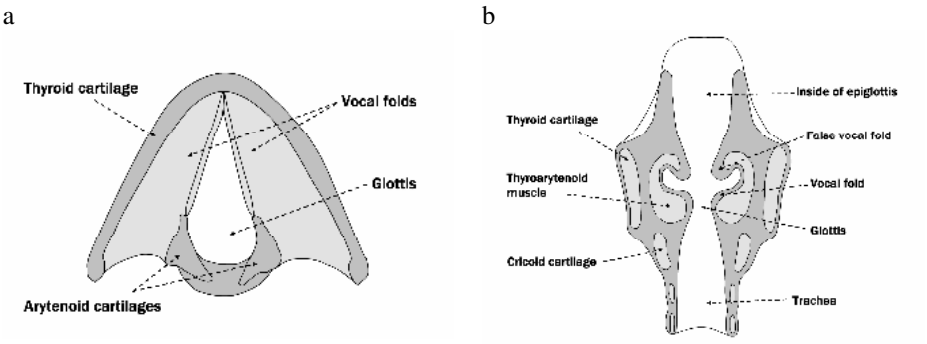


Fig. 2. Schematic diagrams of the larynx: (a) superior view, showing vocal folds, (b) vertical section, showing air passage.

The larynx is the major sound generation structure in speech. It sits in the air pathway between lungs and mouth, and divides the trachea from the pharynx. It is suspended from the hyoid bone which in turn is connected by muscles to the jaw, skull and sternum. This arrangement allows the larynx to change in vertical position. The larynx is structured around a number of cartilages: the cricoid cartilage is a ring that sits at the top of the trachea at the base of the larynx; the thyroid cartilage is a V shape with the rear legs articulating against the back of the cricoid cartilage and the pointed front sticking out at the front of the larynx and forming the "Adam's apple" in

the neck; the two arytenoid cartilages sit on the cricoid cartilage at the back of the larynx.

The vocal folds are paired muscular structures that run horizontally across the larynx, attached close together on the thyroid cartilage at the front, but connected at the rear to the moveable arytenoid cartilages, and forming an adjustable valve. For breathing the folds are held apart (abducted) at their rear ends and form a triangular opening known as the glottis. Alternatively, the arytenoids can be brought together (adducted), pressing the folds into contact along their length. This closes the glottis and prevents the flow of air. If the folds are gently adducted, air under pressure from the lungs can cause the folds to vibrate as it escapes between them in a regular series of pulses, producing the regular tone called "voice". Abduction movements of the vocal folds are controlled by contraction of the posterior cricoarytenoid muscles, which cause the arytenoids to tilt and hence draw the rear of the vocal folds apart. Adduction movements are controlled by the transverse interarytenoid muscles and the oblique interarytenoid muscles which draw the arytenoids together, also the lateral cricoarytenoid muscles which cause the arytenoids themselves to swivel in such a way as to draw the rear of the folds together.

The open glottis position gives voiceless sounds, such as those symbolised [s] or [f]; closure produces a glottal stop, symbolised [ʔ], while voice is used for all ordinary vowels, and for many consonants. Commonly, consonants are in voiced-voiceless pairs; for example, [z] is the voiced counterpart of [s], and [v] the voiced counterpart of [f].

As well as adduction/abduction, the vocal folds can change in length and tension owing to movements of the thyroid and arytenoid cartilages and of changes to the muscles inside the vocal folds. These changes primarily affect the rate of vocal fold vibration when air is forced through a closed glottis. The cricothyroid muscles rock the thyroid cartilage down and hence stretch and lengthen the vocal folds. Swivelling of the arytenoid cartilages with the posterior and lateral interarytenoid muscles also moves the rear of the folds relative to the thyroid, and changes their length. Within the vocal folds themselves, the thyroarytenoid muscle can contract in opposition to the other muscles, and so increase the tension in the folds independently from their length.

Generally, changes in length, tension and degree of adduction of the vocal folds in combination with changes in sub-glottal pressure cause changes in the loudness, pitch and quality of the sound generated by phonation. Normal (modal) voice produces a clear, regular tone and is the default in all languages. In breathy voice (also called murmur), vibration is accompanied by audible breath noise. Other glottal adjustments include narrowing without vibration, which produces whisper, and strong adduction but low tension which produces an irregular, creaky phonation.

2.3. Supra-laryngeal vocal tract

Immediately above the larynx is the pharynx, which is bounded at the front by the epiglottis and the root of the tongue. Above the pharynx, the vocal tract branches into

the oral and nasal cavities, see Fig. 1. The entrance to the nasal cavity is controlled by the soft palate (or velum) which can either be raised, to form a closure against the rear wall of the pharynx, or lowered, allowing flow into the nasal cavity and thus out of the nostrils. The raising of the soft palate is controlled by two sets of muscles: the tensor veli palatini and the levator veli palatini which enter the soft palate from above. Lowering of the soft palate is controlled by another two sets of muscles: the palatopharyngeus muscle and the palatoglossus muscle which connect the palate to the pharynx and to the back of the tongue respectively.

Air flowing into the oral cavity can eventually leave via the lip orifice, though its path can be controlled or stopped by suitable manoeuvres of the tongue and lips. The main articulators which change the shape and configuration of the supra-laryngeal vocal tract are the soft palate, the tongue, lips and jaw.

The upper surface of the oral cavity is formed by the hard palate, which is domed transversely and longitudinally, and is bordered by a ridge holding the teeth. In a mid-sagittal view, the portion of this behind the upper incisors is seen in section, and generally referred to as the alveolar ridge. The lower surface of the oral cavity consists of the tongue, a large muscular organ which fills most of the mouth volume when at rest. Various parts of the tongue can be made to approach or touch the upper surface of the mouth, and complete airtight closures are possible at a range of locations, the closure being made not only on the mid-line where it is usually visualised, but extending across the width of the cavity and back along the tongue rims. The position and shape of the tongue are controlled by two sets of muscles: the extrinsic muscle group lie outside the tongue itself and are involved in the protrusion of the tongue, the depression of the tip of the tongue, the forward-backward movement of the tongue and the raising and lowering of the lateral borders of the tongue. The intrinsic muscles lie within the body of the tongue and are involved in flattening and widening the tongue, lengthening and narrowing the tongue, and also raising and lowering the tongue tip. Together the many sets of muscles can move the bulk of the tongue within the oral cavity and change the shape of the remaining cavity, which in turn affects its acoustic properties.

The available space in the oral cavity and the distance between the upper and lower teeth can be altered by adjusting the jaw opening. Raising the jaw is performed mainly by the masseter muscle which connects the jaw to the skull, while lowering the jaw is performed by muscles that connect the jaw to the hyoid bone.

At the exit of the oral cavity, the lips have many adjustments that can affect the shape of the oral opening and even perform a complete closure. Lip movements fall into two broad categories: retrusive/protrusive movements largely performed by the orbicularis oris muscles that circle the lips, and lateral/vertical movements performed by a range of muscles in the cheeks that attach into the lips, called the muscles of facial expression.

2.4. Sound generation

To a very good approximation, we can describe the generation of speech sounds in the vocal tract as consisting of two separate and independent processes. In the first process, a constriction of some kind in the larynx or oral cavity causes vibration

and/or turbulence which gives rise to rapid pressure variations which propagate rapidly through the air as sound. In the second process, sound passing through the air cavities of the pharynx, nasal and oral cavities is modified in terms of its relative frequency content depending on the shape and size of those cavities. Thus the sound radiated from the lips and nostrils has properties arising from both the sound source and the subsequent filtering by the vocal tract tube. This approach is called the source-filter model of speech production.

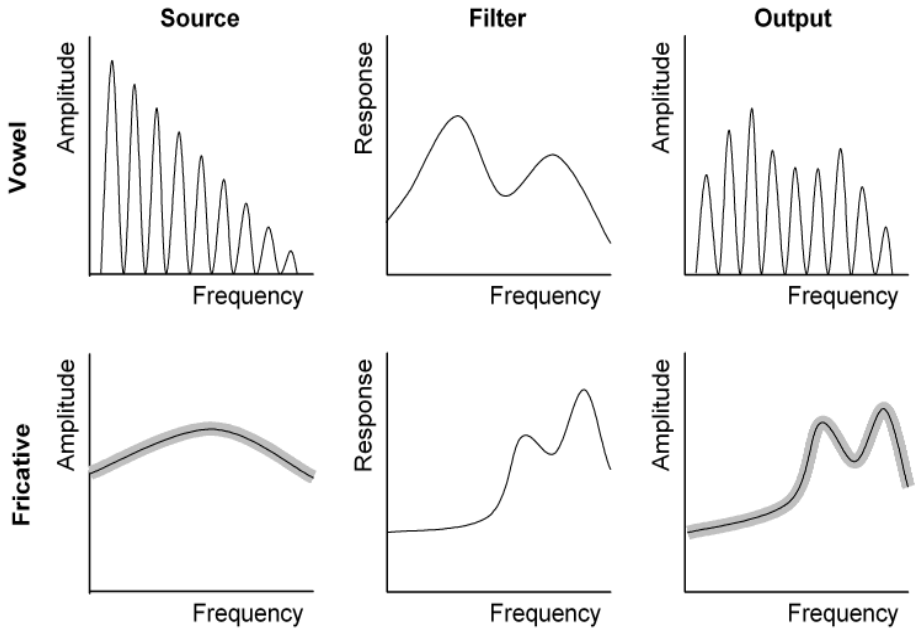


Fig. 3. Frequency domain diagram of the source-filter explanation of the acoustics of a voiced vowel (upper) and a voiceless fricative (lower). Left: the source spectrum, middle: the vocal tract transfer function, right: the output spectrum.

Phonation is periodic vibration in the larynx which starts when sub-glottal pressure rises sufficiently to push adducted folds apart. The resulting flow through the glottis causes a fall in pressure between the folds due to the Bernoulli effect, which in turn draws the folds together and ultimately causes them to snap shut, cutting off the flow and creating a momentary pressure drop immediately above the glottis. The cycle then repeats in a quasi-periodic manner at frequencies between about 50 and 500Hz depending on larynx size and larynx settings. The spectrum of this sound is rich in harmonics, extending up to about 5000Hz, and falling off at about -12dB/octave. See Fig. 3.

Apart from phonation, other sound sources are created by air-flow from the lungs becoming turbulent at constrictions in the larynx and oral cavity or at obstacles to the

air-flow. Noise sources caused by the turbulence have broad continuous spectra which vary in envelope depending on the exact place and shape of constriction. Typically, noise sources have a single broad frequency peak varying from about 2 to 6kHz, rolling off at lower and high frequencies.

The frequency response of an unobstructed vocal tract closed at the glottis and with a raised soft-palate can be well described by a series of poles (resonances) called the formants of the tract, see Fig. 3. The formant frequencies and bandwidths are commonly used to parameterise the vocal tract frequency response. However, when the soft-palate is lowered, when there are constrictions to the air-flow through the tract, or when the glottis is open, additional zeros (anti-resonances) are present.

When sound is radiated from the lips and nostrils, it undergoes another frequency shaping which effectively differentiates the signal, providing a gain of +6dB/octave to the speech signal.

3. Phonetic classification

Phonetic classification is the system of categories and descriptive labels which underlies the Phonetic Alphabet of the International Phonetic Association [3]. It regards speech as a succession of sounds (segments), and characterises the production of each such segment by specifying a relatively static target configuration. This section introduces the standard principles used by phoneticians to categorise the phonetic gestures used in speech.

3.1 Place and manner of articulation

Vowels are sounds produced with a relatively open vocal tract through which air flows with little resistance, while consonants involve some degree of obstruction to the airflow. Place of articulation refers to the location along the vocal tract where a consonantal obstruction is formed.

The terminology for place of articulation is summarised in Fig. 4, around a mid-sagittal schematic of the vocal tract. Words shown without a leading or trailing hyphen are complete place terms. So alveolar refers to a type of articulation in which the tip and blade of the tongue approach the ridge behind the upper teeth, velar to one made by the back of the tongue against the velum, and so on. More precise terminology consists of hyphenated terms on the left, which refer to 'active' articulators, paired with terms from the shaded box (which refer to 'passive' articulators).

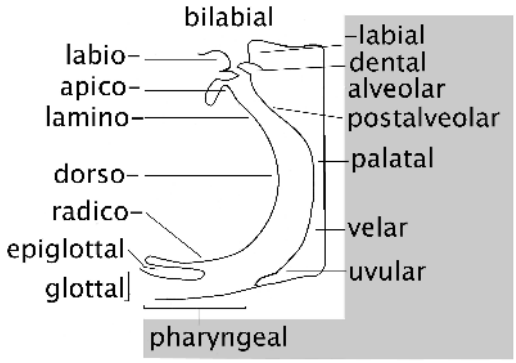


Fig. 4. Schematic of vocal tract showing terminology used to indicate place of articulation (after [4])

Manner of articulation refers to the type of obstruction used in the production of a consonant – whether, for example, the airflow is blocked completely for a brief time (yielding the manner known as plosive) or simply obstructed so that noisy turbulent flow occurs (the manner known as fricative).

Table 1. Manners of articulation (after [4])

manner	**definition**	**comments**
nasal	complete oral closure, soft palate lowered to allow air to escape nasally	
plosive	complete closure, soft palate closed	
affricate	plosive released into fricative at the same place of articulation	not always treated as a separate manner
fricative	close approximation of articulators, turbulent airflow	sibilants, having turbulence at the teeth, are an important sub-category
lateral fricative	complete closure on mid-line, turbulent flow at the side	
lateral approximant	complete closure on the mid-line, open approximation at the side	
approximant	open approximation, flow not turbulent	approximants which are within the vowel space are also called semivowels
trill	flexible articulator vibrates in the air stream	in trills and taps the brief closures do not raise intra-oral air pressure significantly
tap/flap	a single brief closure made by the tongue hitting the alveolar ridge	flaps start with the tongue retroflexed

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	’	Examples:
|	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
!	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
‖	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

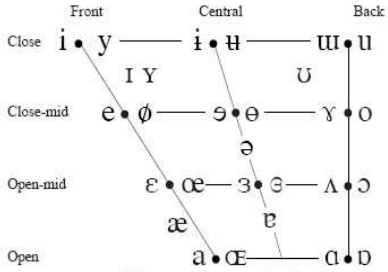
OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʢ ʡ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

k͡p t͡s

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən	
ˌ	Secondary stress		
ː	Long	eː	
ˑ	Half-long	eˑ	
˘	Extra-short	ĕ	
|	Minor (foot) group		
‖	Major (intonation) group		
.	Syllable break	ɹi.ækt	
‿	Linking (absence of a break)		

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
◌˞	Rhoticity	ə˞ a˞	◌̙	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL			CONTOUR	
é͡ or	˥	Extra high	ě or	˩˥ Rising
é	˦	High	ê	˥˩ Falling
ē	˧	Mid	ĕ	˨˦ High rising
è	˨	Low	ě	˩˨ Low rising
ȅ	˩	Extra low	ě	˩˦˩ Rising-falling
↓		Downstep	↗	Global rise
↑		Upstep	↘	Global fall

Manners of articulation are summarised in Table 1. Manners differ chiefly in the degree of obstruction, but also involved are the nasal/oral distinction and the central/lateral distinction. The rate of an articulatory manoeuvre is also relevant: for instance, if the tongue tip and blade make one brief closure against the alveolar ridge the result is called a tap, symbolised [ɾ], but a similar closure made at a slower rate will be a plosive [d].

3.2 The IPA chart

Almost any sound may be voiceless or voiced regardless of its place or manner of production; and places and manners may be (with some restrictions) combined. The IPA chart takes the form of an array, with the columns being places of articulation, and the rows being manners. Voiceless and voiced symbols are put in that order in each cell. Blank cells on the IPA chart correspond to possible though unattested sound types, while shaded cells show impossible combinations

3.3 Vowel classification

For vowels, the arched tongue body takes up various positions within the oral cavity. In the vowel [i] the tongue body is well forward in the mouth, beneath the hard palate, whereas in [u] it is pulled back. Both [i] and [u] have the tongue relatively high in the oral cavity, while the vowel [a] requires it to be lowered (the jaw may open to assist). This provides a two-dimensional vowel "space" in the oral cavity, with the dimensions high-low (also called close-open) and front-back. The lips provide a third independent factor. They may form a spread orifice, as in [i], or be rounded and protruded into a small opening, as in [u]. Using tongue position for [i] but adding lip-rounding in place of lip-spreading yields a vowel which the IPA symbolises [y] as heard in a French word such as *rue* [ry] "street".

The IPA presents the vowel space as a quadrilateral of standardised proportions, based partly on X-ray studies of tongue position during sustained vowel production. Steady-state vowels can be represented as points within this space and symbolised appropriately. Diphthongs (such as those heard in English *sound* or *noise*) can be represented as a movement within the space.

3.4 Further aspects of vowel classification

Certain languages use nasalization to differentiate otherwise similar vowels. A nasalised vowel is produced with a lowered velum, adding the acoustic resonances of the nasal cavity and giving a distinct auditory effect. For example, French [sɛ] *sait* "(he/she) knows" has a non-nasalised (oral) vowel, while [sɛ̃] *saint* "saint" has the nasalized counterpart.

Vowels may also have 'r-colouring' or rhotacisation, produced by a modification of tongue shape, typically by combining a curled-back tongue-tip gesture with an otherwise normal vowel. It is heard in North American English in such vowels as that in *nurse*, or the second syllable of *letter*.

3.5 Multiple articulation

Languages may make use of pairs of segments which are alike in voice, place, and manner but distinct in sound because of an accompanying secondary adjustment. In the RP variety of English, for instance, a voiced alveolar lateral consonant which precedes a vowel (as in *look*) is different from one which occurs after a vowel (such as *cool*). The second one has a raising of the back of the tongue (the sound is said to be velarised). For English the difference is automatically conditioned and carries no meaning but in many languages (for example, Russian) this type of difference is applied to numerous pairs of consonants, and utilised to create linguistic contrasts. Apart from this velarization, other common types of secondary articulation found in languages are labialisation (the addition of a labial stricture, usually lip-rounding), and palatalisation (simultaneous raising of the front of the tongue towards the palate). Constriction of the pharynx gives pharyngealisation, which is used in the "emphatic" consonants of Arabic.

If there are two simultaneous gestures of equal degree by independent articulators, the result is termed double articulation. For example the Yoruba word for "arm" (part of the body) is [akpá], where [kp] indicates a voiceless plosive formed and released at the bilabial and velar places simultaneously. This is termed a labial-velar plosive. The widespread approximant sound [w] is also a labial-velar double articulation.

3.6 Non-pulmonic airstreams

The egressive pulmonic airstream is the basis of speech in all languages, but certain languages supplement this with non-pulmonic airstreams – mechanisms which produce short-term compressions or rarefactions effected in the vocal tract itself, used for certain of their consonant sounds. Non-pulmonic sound types found in languages are ejectives, implosives and clicks. Ejectives, symbolised with an apostrophe [pʼ tʼ kʼ tʃʼ] are the commonest type. Their articulation resembles that of ordinary voiceless plosives and affricates, but they are produced with a closed glottis, which is moved upwards during the production, shortening the vocal tract and compressing the air trapped behind the articulatory constriction. Release of the articulatory closure takes place (generally with a characteristic auditory effect, which can be relatively powerful) and the glottal closure is then maintained for at least a further short interval. The speaker will then generally return to the pulmonic airstream for the following sound. This mechanism can be called the egressive glottalic mechanism. By definition, the vocal folds are closed, so all ejectives must lack vocal fold vibration.

Implosives are made by moving the closed glottis down rather than up, giving the ingressive glottalic mechanism. The implosives commonly encountered in languages

are voiced, rather than being simple reversals of the ejective type. In these, egressive lung air passes between the vibrating vocal folds at the same time as the larynx is in the process of lowering. They are symbolised with a rightwards hook attached to the symbol for a voiced plosive, as in [ɓ ɗ ɠ].

Clicks are widespread as paralinguistic noises (such as the "tut-tut" of disapproval) but found as linguistic sounds in relatively few languages. They are suction sounds formed by enclosing a volume of air in the mouth and then enlarging that volume by tongue movement, with a consequent reduction in pressure. The back of the enclosed volume is formed by the tongue in contact with the velum; the front closure may be completed by the lips, or by the tongue tip and blade. Clicks are categorised as bilabial, dental, (post)alveolar, palatoalveolar and alveolar lateral. The basic clicks mechanism is voiceless, but the remainder of the vocal tract may perform a wide range of click accompaniments, including voicing, aspiration, voice-plus-nasality, glottal closure and others

3.7 Beyond the segment

Length, stress and pitch are classified as suprasegmental (or prosodic) features on the current version of the IPA chart, This means that that they do not apply to single segments, but to sequences of segments, or entire syllables.

Languages often distinguish short and long vowels, and (less commonly) short and long consonants (the latter equivalently termed geminates). Vowels paired as "long" and "short" within a language do not necessarily differ only in duration. In English, for example, the "short " vowel of *bit* is typically lower, and centralised, compared with the "long" vowel of *beat*, and there are similar quality differences in the other pairs.

Duration, loudness and pitch are all relevant to the marking of "stress", which is generally considered to be a property of a whole syllable. Languages which make use of stress typically use it to render one syllable within a polysyllabic word more prominent. The position of this syllable may be fixed for a particular language (so in Czech, the initial syllable is stressed, whereas in Polish the penultimate syllable is stressed) or alternatively free to occupy various positions and thus differentiate words (so in English the noun *import* is stressed on the first syllable, but the verb is stressed on the second). Stresses are important to listeners in the task of parsing the incoming string into words.

The IPA also provides a range of marks for indicating pitch levels and movements on syllables. In tone languages (such as the various kinds of Chinese), the pitch level or pitch movement applied to each syllable is fixed as part of the makeup of each word. In addition, all languages (whether or not they employ lexical tone) make use of intonation, which is pitch variation applied to utterances of phrase length. It typically shows the grouping of words into "chunks" of information, the relative importance of words within the phrase, and affects interpretation (for example by marking utterances as questions or statements). Intonation can be modelled by locating a small number of tone levels (or movements) at specific points in a phrase,

from which the overall pitch contour can be derived by interpolation. It is very common for the endings of intonation patterns (terminal contours) to carry particular significance in interpretation.

4. Expressions of individuality

The production of a spoken utterance can be described in terms of a sequence of processing stages in the speaker, starting from a desire to achieve a communicative goal, and ending with sound generation in the vocal system. In this section we will build on the discussion of the human organs of speech given in section 2, and the discussion of how they are exploited to create different phonetic gestures given in section 3, to present a catalogue of the ways in which speaker-specific characteristics influence these generation stages.

In section 4.1 we'll look at how a speaker uses language to achieve a particular communicative goal – here a speaker will show preferences for which language to use, which words to choose, which grammatical structures are most appropriate for the circumstances. In section 4.2 we'll look at the phonological stage of production - given an utterance, the speaker must plan which phonetic segments and which form of intonation and rhythm would be most appropriate. In section 4.3 we'll look at how the sequence of segments must in turn be realised as a continuous and dynamic series of phonetic gestures whereby the articulators move to realise the phonological form. In section 4.4 we'll see how the movements of the articulators, particularly the jaw, lips, tongue, soft palate and larynx creates sounds which themselves carry speaker information as well as the spoken message. In section 4.5 we'll see how all of these stages can be influenced by the mental and physiological state of the speaker, for example whether they are emotionally aroused, or whether they are intoxicated. In section 4.6 we'll describe ways in which speaker characteristics can be used to place an individual as a member of a number of groups. Finally, in section 4.7 we'll introduce two main application areas for information about speaker identity extracted from speech.

4.1 Individuality in language and language use

Since there are about 5000 different languages in the world (the number varies depending on the definition of 'language') the choice of language used by a speaker can already be considered a distinguishing characteristic. While a few languages have large populations of speakers, many are very small, and often geographically isolated. Icelandic for example, has only approximately 280 000 speakers in the world. Many speakers also have competence in more than one language, so a speaker may be able to decide which language best suits a given circumstance. When a speaker uses a second language, it is also very common for their language use to be influenced by properties of their first language.

Even within a language, there can be variations in dialect – relative differences in the frequency of use of words and grammatical forms as well as variations in the pronunciation of words. For example, Scottish speakers of English might more

frequently use the word *wee* to mean small, where other English speakers may use *little*. Or *I don't know* in British English, might be produced as *Me no know* in Jamaican English. Dialects do not only vary in a geographical sense; since language changes with time, older speakers may use different forms to younger speakers. Similarly dialect use may be indicative of a social grouping, perhaps related to socio-economic class, education or gender. Thus speaker A may tend to use *that's right* where speaker B prefers to say *OK* or speaker C says *cool*. Such words will then occur in the spoken discourse of a speaker with a higher frequency and are thus an individual feature of this particular speaker.

Of course different speakers will also react differently in different situations, so the very way in which a speaker chooses to use language to communicate in a specific situation can also be indicative of their identity.

4.2 Individuality in the sound system

After an utterance has been constructed as a sequence of words, it is mapped into a set of speech sounds spread out across time. One component of this process deals with which phonetic segments are needed (this is called the segmental component), and the other is related to the stress pattern, timing and intonation of the sequence (this is called the supra-segmental component).

On a segmental level, the sound inventory used in the mental lexicon to represent the pronunciation of words can vary from speaker to speaker even for one language. This is one part of what is called the 'accent' of the speaker. For example most British English accents differentiate the words *Kahn*, *con* and *corn* using three different back open vowel qualities; however many American English accents use only two vowels in the three words (e.g. *Kahn* and *con* become homophones). Similarly, older speakers of English may differentiate the word *poor* from the word *paw* through the use of a [ʊə] diphthong that is not used by younger speakers.

As well as differences in the number and identity of segments in the sound inventory, accents may also differ in terms of the distribution of those segments across words in the lexicon. Both Northern and Southern British English have the vowels of *trap* and *palm*, but for the word *bath*, the Northern speakers use the first vowel, while Southern speakers use the second. Another frequently observed phonological variation across English accents is rhoticity – whether the accent expresses post-vocalic 'r' letters in the spelling of words as [r] sounds in their pronunciation.

On the supra-segmental level, speakers can vary in the way in which they seek to use intonation to select sentence functions (e.g. questions versus statements) or in the way in which particular words and phrases are highlighted (e.g. putting a focus on particular elements). For example, "uptalk" (the use of a rising terminal intonation on utterances that might otherwise be expected to have a simple fall) has recently become characteristic of younger speakers of English across a range of locations (Britain, America and Australia/New Zealand).

4.3 Individuality in controlling the speech production process

Given the phonological form of the utterance at the segmental and supra-segmental levels, the next stage in the process of speaking is the execution of the utterance through movements of the articulators. This process involves a sophisticated neural control system for the co-ordination of the many muscles involved in moving the tongue, jaw, lips, soft palate and larynx through the utterance. The complexity of this task introduces many possibilities for the expression of speaker characteristics.

A common way of modelling the motor control problem in speaking is to think of each phonological segment as specifying an articulatory gesture involving one or more articulators. In this model, speaking is then executing phonetic gestures in sequence. Fundamentally however, one gesture overlaps in time with the next, so that gestures can influence each other's form, a process called coarticulation. Importantly, the degree of coarticulation depends on the degree of gestural overlap, which may in turn depend on the amount of time available for the gestures to complete. Thus a speaker who speaks more quickly, or who decides to de-accent part of a particular utterance may show more coarticulatory behaviour than another. The consequences of coarticulation may be that the articulators do not reach the intended target position for the segment, or even that the segment has no measurable physical effect on the final production.

Speakers also vary in the particular form of gesture they use to implement a given underlying sound segment. This is another aspect of accent and gives rise to a lot of variation across individuals. Vowels are particularly variable: the exact gesture and hence the exact sound quality used to realise a particular vowel segment can vary widely. So, for example, the *trap* vowel in American English accents can vary widely in height. Consonants are affected too; for example, a recent innovation in English is the use by some speakers of a labio-dental approximant [ʋ] to implement phoneme /r/, more usually realized as a postalveolar approximant [ɹ].

The particular articulation used to realise a speech sound also varies according to context, and of course different speakers vary too in how much they are affected by the context. For example, some Southern British English speakers realise the velarised or "dark" variety of phoneme /l/ as a back rounded vowel rather than as a velarized lateral consonant in words like [bɪod] *build*.

Larynx settings are a rich source of speaker variation. As we have seen, changes in the degree of adduction and tension of the vocal folds in combination with changes in sub-glottal pressure lead to variations in voice quality. Speakers vary in their preferred, or default voice quality – particularly the degree of breathiness in the voice. Speakers can also vary the voice quality they use depending on the context, so that a breathier voice may be used for intimate communication, while increased vocal effort may be required for a noisy environment. Creaky voice can have discourse use, for example to mark the ends of topics or of dialogue turns.

Variations in vocal fold tension are used to manipulate the pitch of voiced sounds, and are the main means of implementing intonational changes in speech. Again the default pitch and pitch range used to realise particular intonation changes will vary from speaker to speaker. The mean pitch used by a speaker can vary according to the communicative context: people famously raise their pitch to talk to small children.

The range used for an utterance can be quite small – giving a monotonous pitch – or quite large – giving a dramatic quality to the speech.

Speaking rate also varies from speaker to speaker, and this not only has consequences for the duration of individual syllables: a faster rate may also increase the degree of coarticulation between adjacent gestures.

4.4 Anatomical influences on individuality

The physical size of the organs of speech is a significant source of inter-speaker variation in the speech signal. The length of the vocal tract affects its acoustic properties as a resonance chamber and hence how it functions in shaping sound sources (see section 2.4). The length and mass of the vocal folds in the larynx influence the default pitch, pitch range and voice quality available to the speaker.

The influence of vocal tract size can be seen by considering the frequency response of a simple tube, closed at one end as we change its length. For a tube of length 17.6cm – the typical length of an adult male vocal tract – the first three resonances of the tube are close to 500, 1500 and 2500Hz. These frequencies are similar to the formant frequencies for a central vowel quality. These resonant frequencies scale inversely with the length of the tube, so that a 10% increase in the length of the tube leads to a 10% decrease in the resonant frequencies. This explains why adult female formant frequencies are higher than men on average, since a typical adult female vocal tract is shorter than that of an adult male. Of course there is considerable inter-speaker variation in vocal tract length, and some women have longer vocal tracts than some men. In addition, vocal tract length can be varied within a single speaker through adjustments to the height of the larynx and to the degree of lip protrusion.

Vocal folds vary across individuals in both size and mass, and this impacts the range of frequencies for which they can vibrate. Post-pubertal men have longer and thicker folds with a lower modal frequency compared to women and children. The range of frequencies available is indicated by the range used in singing, which for men is about 87-415Hz, while for women it is 184-880Hz. While it is possible to achieve vibrational frequencies outside these ranges, this usually involves changes in the quality of vibration: irregular creaky voice at the lower end, and falsetto voice (made with tense, rigid folds) at the upper. Individual speakers vary in both the range of frequencies they are capable of producing, and in the range of frequencies used in everyday speech. More typically, speakers use a range of only about one octave of fundamental frequency in normal speaking.

Phonation builds on the capability of the respiratory system to provide a large volume of air at a suitable, steady sub-glottal pressure. Respiratory volumes vary across individuals, and hence the quantity of speech that can be produced on one breath varies. This is also strongly influenced by the efficiency of phonation, with weaker adduction causing greater air loss.

The soft palate acts as a valve that isolates the nasal cavity from the oral cavity. Differences in the effectiveness of the valve and the way this is used in speaking can lead to changes in observed nasality of a speaker's voice.

4.5 Other influences on individuality

In the previous four sections we have considered how a speaker can impose his or her individuality on speech at different processing stages in speech production. In this section we look at how changes in the state of the speaker can also affect his or her speech. We'll look at changes over time, changes in emotion or changes in pathology.

A speaker's voice does not remain constant since the speaker's vocal anatomy and physiology is affected by age. The larynx develops in children as they are growing, and its size and shape is particularly affected by hormonal changes at puberty, both for men and women. For men, the vocal folds can grow in size and mass over a short period, leading to a period of phonation instability as the speaker learns to control the new system. The vocal folds and their control are also affected by advancing age, and modal pitch, the degree of breathiness, and the degree of creakiness can all be affected.

The vocal tract itself also changes in size as a child grows, and this of course changes the range of resonant frequencies available. Control over the vocal tract, reflected in the degree of articulatory precision also develops in the first ten or so years of life. While vocal tract size remains relatively stable with advancing age, there may be significant degeneration in muscles, in the control process, and indeed in the ability to use language, such that speech becomes slower and less well articulated. Similar changes in speech can be brought about by physiological changes, such as tiredness or intoxication.

The emotional state of the speaker can have a great influence on the way in which speech is produced as well as on the content of the messages communicated. Increasing emotional arousal can raise the mean pitch and the pitch range as well as causing changes in loudness. Different emotions can have differing effects, so that it may be possible to differentiate emotional states such as anger, fear, sadness, joy and disgust, although speakers vary in exactly how these affect speech [5].

The health of the speaker can also influence his or her speech. Minor pathologies such as upper respiratory tract infections influence the larynx and the nasal cavities. Laryngitis is a swelling of the folds in response to infection which causes a lowering in pitch (due to the increase in mass of the folds), and can even prevent phonation occurring. Blocked nasal cavities can create a hypo-nasal form of speech that listeners recognise in speakers with a cold.

More serious pathological conditions, particularly stroke, can have effect on the parts of the brain responsible for speech planning and motor control – realised as aphasia and dysarthia. Damage to the vocal folds, such as swelling and the growth of nodules and polyps can also affect phonation and hence voice quality. Smoking and alcohol consumption have both been shown to cause vocal fold pathology.

4.6 From individuality to identity

We have shown that individuality can be expressed in many ways in speaking, at all levels of the message generation process. But while an individual speaker may exhibit a combination of characteristics that may make his or her own speech unique,

it is very likely that each one of the characteristics is also used by other speakers. Thus another way of describing speaker characteristics is in terms of the groups of individuals which share a given feature. And another way of defining a speaker is in terms of the groups that the speaker is a member of.

We are used to grouping speakers on the basis of categories such as language, dialect and accent. However these may be much less well defined in practice than they are in theory. What differentiates a language from a dialect is not always clear. Sometimes, geopolitical factors, like a country's borders, influence the definition of the language of a speaker. Accents may be defined in both geographical and social terms; and people can be both geographically and socially mobile. A speaker might use a blend of languages or vary their accent according to circumstance.

Even if the groups are well defined, it may not always be easy to assign a speaker to a group. The very measurements we make of speech are prone to error, and the particular speech we measure may be unrepresentative of the speaker as a whole. For example it is not always the case that we can determine the sex of a speaker from their speech; and the estimation of age or physique can be quite difficult.

On the other hand, when the context is constrained, speaker characteristics can sometimes be used quite reliably to identify individuals. So when a friend on the telephone says *Hi, it's me* then the combination of the observable speaker characteristics and the limited number of speakers known to you that might introduce themselves to you in this way may mean that the speaker can be accurately identified. This is not to say however, that you wouldn't be fooled by an impostor.

4.7 Applications

Within the field of Speech Science, much more emphasis has been placed on the scientific investigation of the linguistic content of utterances than on the investigation of speaker characteristics. To build an automatic speech recognition system that converts speech signals to text, for example, the speaker information must be discarded or ignored. Theories of production and perception focus on the strategies required to facilitate communication of words and meanings rather than on speaker identity. However two application areas for speaker information have emerged and these have led to an increasing interest in the individuality of voices. One is the field of forensic phonetics that deals with speaker identification in legal cases, the other is the technological field of speaker verification for voice access systems.

Forensic phonetics is a field in which phonetic knowledge is applied in legal cases where the identity of the speaker in a recording is disputed. Forensic phonetics distinguishes between two methodologies: identification of the speaker a) through the use of linguistically/phonetically naïve subjects, or b) through a trained expert witness [6]. In method a) a witness of the crime (e.g. a person who has received sexually harassing phone calls) is asked to identify an alleged perpetrator by his/her voice by picking the speaker out from a 'line-up' of similar voices. In method b) a trained speech expert carries out an identification between the recorded voice and that of a specific suspect. This process is often done with a combination of auditory phonetic comparisons (the expert witness judges on the basis of his/her expert perception) or

technical comparisons (analysis of fundamental frequency, formant frequencies, etc.). Speaker specific characteristics at all levels may be appropriate for forensic applications. The number of characteristics shared between recording and suspect increases the likelihood that the suspect made the recording, although it is much harder to estimate the likelihood that a person other than the suspect could have made the recording. Thus expert evidence in forensic speaker identification is better at eliminating suspects than in confirming them.

A second application that relies on speaker characteristic information is the field of speaker verification. Such systems can be used to secure access to a facility or resource, such as a building or a bank account. Typically a speaker is enrolled into the system using some known speech materials, and then the speaker is asked to verify his identity by producing a spoken utterance on demand. The main difference to forensic applications is that the speaker hopes to be identified successfully and is therefore willing to co-operate. To ensure that recordings of the speaker cannot be used to fool the system, a speaker verification system will typically request novel phrases to be produced to gain access – random digit strings for example. Most systems exploit low-level, speaker-specific spectral information found in the signal – that relating to pitch, voice quality, vowel quality and vocal tract length. This is because it is harder to extract robust speaker-specific information at higher levels. The restriction to low-level features also enables the possibility of text-independent verification, where speaker identity is verified without knowledge of what they are saying. It is hard to make speaker verification systems particularly accurate: false alarm rates and false rejection rates of 5% or more are common. When a system is modified to accommodate the variability in production that occurs within the true speaker (when they have a cold or are tired, for example), this inevitably increases the success rate of impostors. The possibility that there is unused information present in the speech signal that would improve the performance of such systems is still open to investigation.

Acknowledgments

The IPA chart is used by permission of the International Phonetic Association.

References

1. Flanagan, J.: Speech Analysis, Synthesis and Perception. Springer Verlag (1972).
2. Atkinson, M., McHanwell, S.: Basic Medical Science for Speech and Language Pathology Students. Whurr (2002).
3. International Phonetic Association: Handbook of the international phonetic association. Cambridge University Press (1999).
4. Ashby, M.: Phonetic classification. In Brown, et al (eds), Encyclopedia of Language and Linguistics 2nd Edition. Elsevier, (2005).
5. Banse, R. & Scherer, K.: Acoustic profiles in vocal emotional expression. Journal of Personality and Social Psychology, 70 (1996), pp614-636.
6. Nolan, F.: Speaker identification evidence: its forms, limitations, and roles. Proceedings of the conference 'Law and Language: Prospect and Retrospect', Levi Finland (2001).