The Production and Perception of Speech

Mark Tatham
Katherine Morton

This textbook was first published in 1988. The final edition reproduced here was published in 1997. As a textbook the work is suitable for anyone interested in the relationship between phonetics and speech production and perception.

www.morton-tatham.co.uk – mark.tatham@morton-tatham.co.uk – katherine_morton@morton-tatham.co.uk

Contents

IMPORTANT NOTE

Several of the figures used in Sections 4, 5 and 6 were taken from Borden, G. and Harris, K. (1980 – 1st edition) *Speech Science Primer*. Baltimore/London: Williams and Wilkins. These are much better than we could have drawn for the 1988 edition of this book – hence the borrowing. In addition, the discussion in these Sections was also influenced by the same book, which, now in later editions, remains on our teaching reading lists. The latest edition (as of 2008) is:

Raphael, L.J., Borden, G.J., and Harris, K.S. (2006 – 5th edition) *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Hagerstown MD: Lippincott, Williams and Wilkins. ISBN-10: 078177117X and ISBN-13: 978-0781771177

I. INTRODUCTION

Up to around 1950 phonetics was mainly concerned with modelling surface anatomical and articulatory aspects of speech production. Basic applications of the subject included areas such as developing a transcription system for speech, and training people to make and discriminate perceptually the sounds which occur in languages. This simple theoretical approach, together with a simple view of the acoustics of speech, was the basis of much of the early work in transformational generative phonology, especially in the development of distinctive feature theory, beginning with Trubetskoy in the 1930s.

In the 50s and the first half of the 60s it could be said that acoustics was the dominant area of phonetics. Most of the classical research into the acoustics of speech was done during the period, and the acoustic theory of speech production was being continuously refined. There were a few practical applications: crude efforts at speech synthesis and automatic speech recognition were made, though these were dominated by considerations of the complicated and expensive electronic technology involved. Tape recorders replaced transcription of the acoustic signal for serious work in noting down how people spoke, and as phonology developed it was being realised just how wide the gulf was between our subjective view of speech and the actual facts.

Later in the 60s and throughout the 70s the major concern was articulation and its control. Initial experimental work led to an abstract theory of speech production and articulator control. This work derived much of its impetus from developments in phonological theory within the framework of transformational generative grammar. Lately, proposals in the area of speech production theory have taken a more mechanistic approach following work on movement control in neurophysiology, and have tended to move away from a purely linguistics orientation.

The 80s saw return to studying the acoustics of speech. This time the focus was on applications of the acoustic theory of speech production and acoustic aspects of speech perception. The impetus here was and has remained the enormous pressure to provide practical and reliable systems giving artificial speech production and perception to computers. The work is not so heavily technology oriented as it was in the 60s since the technology itself is no longer seen as the difficult part of making machines talk or respond to speech. In the 60s this field of application was dominated by the technology itself, but more recently it is spoken language engineering which has become central. Spoken language engineering is concerned with how to make the technology produce the right soundwaves in speech synthesis or recognise them correctly in automatic speech recognition. Thus it is concerned with manipulating the technology rather than with the actual design of the technology itself.

Alongside the development of applications of acoustic theory, there have been significant changes in general phonetic theory since 1980. In particular the view promoted by transformational generative phonology, though not so much by earlier phonologists, that phonetics was in some sense an entirely physical component tacked on the end of the phonology for the purposes of realisation of phonological requirements has been shown to be

fundamentally unsound. We have seen a return to the idea that much of speech production at the phonetic level is *cognitively* dominated, as opposed to being *physically* dominated.

Throughout this time (1950 to the present day) work has been continuing in the study of speech perception, and phoneticians have profited from advances made in experimental psychology and psychological theory. Perceptual phonetics has come into focus because of the current need to provide theoretical backup to automatic speech recognition systems. Advances in our understanding of the neurophysiology of hearing, and perception have led to some significant revisions of the theory of speech perception since 1980. These revisions parallel those in the theory of speech production, and to a large extent concern the relative roles of cognitive and physical aspects of production and perception. One particular change has been the introduction of the idea that speech production and perception are not separate activities or behaviours, but are fully integrated and for the most part may well be simply *different operational modalities* of a single system.

All this has taken place against a backdrop of important groundwork in phonetics and phonology since the late 19th century. In no way should this heritage be denied, even though some aspects of it may seem a little strange to us today. Early phoneticians were also the phonologists since within the general field of speech little distinction had been made between these areas. They established a metatheoretical framework for the discussion of observations about speech which could lead to the insights necessary for progress in linguistics. Important classificatory systems were devised as well as several systems for the symbolic representation of speech events, known as transcription systems.

The 20th century has seen the enormous development of the technology permitting laboratory investigations of all aspects of speech from the acoustic waveform to the neurology of the motor control of articulation and complex computer simulations of production and perception. It can reasonably be said that much of the impetus of any particular area of research in the field has come from technological development. An example of this would be the invention in the 40s of the sound spectrograph, a device for analysing the acoustics of speech easily and reliably.

If there is a technological impetus today it comes from the computer. Not only has the computer replaced much early special purpose equipment in the phonetics laboratory where the concern is with speech analysis, but the computer's need for ears and a mouth (so to speak) has pushed phonetics into the areas of artificial intelligence and information technology.

Since the early 50s there have been metatheoretical developments. If language can be regarded as a knowledge based system what is the appropriate representation of that knowledge in the theory? As linguists we have been used to knowledge representation in terms of rules or productions, but ideas are changing because many observations have indicated that rules may be an inadequate mathematical approach to modelling language. Certainly as we pass from modelling competence to modelling performance we see more and more that the use of rules does not enable us to account for many of the properties of language, its acquisition and usage. This parallels the development of computational models which are intended to *simulate* human language behaviour, rather than more simply describe aspects of it. It is interesting that the deficiencies of rule based accounts of language are most apparent in the areas of semantics and phonetics: these are the areas recognised in linguistic circles as being those of the greatest complexity and these promise to be the areas of most intense research activity over the next couple of decades.

II. THE COURSE

In this course you will be studying phonetics and phonology as part of the phenomenon of language, not specifically how to learn or teach the pronunciation of any particular language. How we learn or teach pronunciation comes under the heading of applied linguistics (including applied phonetics in this case) which is a much less formal area of study based on the core disciplines of linguistics and phonetics.

The theory of phonetics and phonology forms part of a complex and multidisciplinary subject area, the range and depth of which goes considerably beyond the scope of this particular course. Phonology and part of the phonetics of speech production involve cognition (and as such call upon psychology as a foundation discipline), but at the periphery of speaking and hearing phonetics also draws on anatomy, neurophysiology, aerodynamics and acoustics. The theory itself is formal and mathematical in nature, and modern models built using the theory are usually *computational*.

The study of speech production is therefore difficult in the sense that some understanding of several neighbouring disciplines is essential. While other areas of linguistics, such as syntax and semantics, draw mainly on logic and psychology in their theories and a small well defined area of mathematics in their modelling, phonology and in particular phonetics go well beyond these areas.

In addition phonetics and phonology have made considerable progress in the area of simulation. Much work has been done in the last twenty-five years or so in the area of computer modelling of the human processes of speech production and perception. Although there has been some work on simulation in linguistics over this period it is only comparatively recently that computational linguistics has begun to mature to the point where computer simulations will contribute to our understanding of the natural processes.

In other words, the study of speech production and perception is vast. In putting together this course we had a choice between skating over as much of the surface as possible in the time available, or choosing a small and firm foundation on which to elaborate in some depth on narrow topics highly relevant to language study, including learning and teaching. We chose the latter.

But then there was another choice: the activity in the discipline over the last quarter century has resulted in dispelling many of our earlier ideas about speech and in the emergence of new ways of looking at the subject. In some ways the new ideas look more difficult to understand (though this is almost wholly an illusion brought about by their newness). Do we talk about the old ideas or the new ones? The answer to this question is not easy. The new ideas are obviously the ones to go for: they will be elaborated in the future and you would have a basis from which to understand future work. But the old ideas are the ones on which many of the ideas in linguistics (especially phonology) are built, and unless you understand something about them you will not understand important areas of contemporary linguistics.

We intend a compromise mix of the old and the new: not to blend them, but to tell you about both when necessary. When it’s not necessary to understand both old and new, we shall deal only with the new ideas. You can help by understanding from the very beginning that the 60s and 70s saw a revolution in our approach to speech, and by keeping straight in your minds which ideas come from before that period and which grew out of it. In the late 80s new computational methods for modelling speech have emerged, together with a shift from description to simulation, and although we shall not be dealing in any detail with these recent changes you should bear in mind that movement in the subject is rapid as we move into the next century.

1. THE EARLY DESCRIPTIVE FRAMEWORK

1.1 Anatomical Considerations

Traditionally an anatomical approach has been taken for the description of articulatory configurations. Phoneticians began by identifying what are called the organs of speech, or the articulators. Typically textbooks list, for example, the lips, the teeth, the palate (often identifying parts of the palate: the alveolar ridge, the dome, etc.), the velum, the uvula, the tongue, the pharynx, the larynx (site of the vocal cords). As well as the anatomy, phoneticians identified the oral cavity (the space manipulated by the organs forming the mouth), the nasal cavity (comprising the nasal passage) and the pharyngeal cavity as those parts of the overall vocal tract resonator which determine the characteristics of the filter applied to source sounds in the acoustics of speech (see *Acoustics*).

The overall descriptive model implied active manipulation of the individual organs of speech to form the various articulatory shapes associated on the one hand with speech sounds and on the other with the discrete segments of phonology. The chain of events was:

1. a cognitive decision to produce a particular sound,
2. direct control of the anatomical system to make the corresponding vocal tract shape, resulting in
3. the correct sound segment as characterised by the acoustic theory of speech production.

1.2 Classifying Consonants

Having identified the speech organs the next stage in the traditional descriptive model involves showing how these combine their individual shapes to form the overall configuration. There are two important points to this stage.

1. Phoneticians establish the primary and secondary articulators used to produce a given sound. Thus for example the primary articulator for the vowel [u] is the tongue [high at the back of the mouth]; the secondary (i.e. less important or critical) articulator is the lips [somewhat rounded]. Since [u] is a vowel it is a given that there is vibration of the vocal cords.
2. The articulations are classified, using some of the named anatomical features, on a grid or matrix. For consonants one axis of this grid names place of articulation (*where* the primary articulation takes place in the system), the other names manner of articulation (or how the articulation takes place).

		place		
		labial	dental	velar
manner	stop	p / b	t / d	k / g
	affricate		ʧ / ʤ	
	fricative	f / v	s / z	

Fig. 1 Fragment of the matrix classifying consonants.

Symbols representing individual phonetic segments are placed within the cells forming the two dimensional matrix. In the early model the notion *phonetic segment* was ambiguous: on the one hand a segment meant the articulatory configuration associated with a phonological unit, and on the other it meant the steady state sound produced by the configuration. The

symbolic representation of the International Phonetic Alphabet was similarly ambiguous – the symbols meant both articulations *and* sounds.

A third dimension to the matrix – voicing – is implied by the frequent placing of *two* symbols in a cell. Thus, [p] is the voiceless counterpart of [b], with the implication that aside from voice [p] and [b] are identical. In the above fragment we have adopted the convention that the rightmost symbol of each pair represents the voiced articulation; this is the usual convention found in textbooks on speech.

Notice that the labelling of the rows tells us *how* the articulation is made, and uses classifiers such as stop, fricative, etc. A stop describes an articulation involving a complete stoppage of airflow by the articulators at the place identified on the other axis. Thus [p] and [b] are articulations involving airflow stoppage at the lip place. A fricative involves articulatory constriction at the identified place to produce frication (not friction): thus [s] and [z] are alveolar fricatives. An affricate is an articulation which begins like a stop, but dissolves into the corresponding fricative: thus [t] and [d] are alveolar affricates which start like the alveolar stops [t] and [d], and end like the fricatives [s] and [z].

There is a similarity here between this place and manner matrix and the more recent distinctive feature matrix in phonology. Both characterise segments in terms of more elemental units or labels; both enable the easy identification of classes or special subsets of the set of segments. So, for example, the place/manner grid identifies [p], [b], [f], [v] as members of a labial subset, or [f], [v], [s], [z] as members of a fricative subset,

1.3 Classifying Vowels

In this early descriptive system vowels are treated differently from the consonants we have been looking at so far. Given that the tongue is the primary articulator in vowels, a map-like chart is set up as a kind of stylised cross-sectional two dimensional view of the oral cavity (Fig.2).

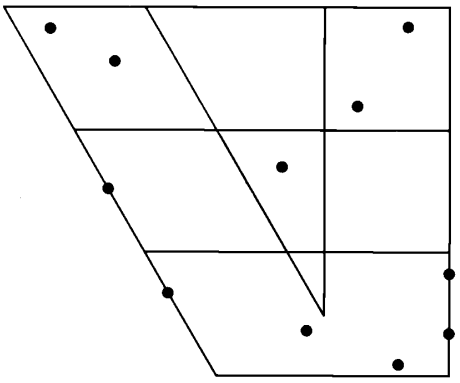


Fig. 2 The standard vowel chart based on a stylised cross section of the vocal tract looking left.

Points are located on this grid and were thought of as corresponding to the *highest* or most significant point of the tongue in the mouth during the articulation of the various vowels. This was subsequently (in the 30s) shown, using x-rays, not to be true, but the diagram persists to this day, and indeed the labels high, mid, low and front, centre, back which were used on the two axes of the grid initially to refer to the actual position of the tongue's highest point in the mouth are now used in distinctive feature theory to refer to an abstract mental map of tongue position. The claim here, of course, is that although it is now recognised that there is no *real world* identity to these labels they nevertheless have some mental or psychological reality.

This concept of psychological reality is an important one for linguistic theory.

Psychological Reality

- A concept is said to be *psychologically real* if it is used by the mind in its processing (of speech, in this case). Sometimes such abstract concepts may not have real world correlates. So, for example, the mind may think of a particular speech sound without having any regard for actually what is involved mechanistically in making that sound. As an illustration of this point take the feature of length. Sometimes it is cognitively useful to have the concept of length, as when vowels become lengthened preceding voiced obstruents which are later devoiced. The only mental consideration is an abstract contrast of long *vs.* short. But at the phonetic level length translates to duration in terms of a certain (and perhaps crucial) number of milliseconds. It is meaningless to refer vaguely to the length of a vowel in phonetics, while at the same time it is both proper and meaningful to do so in phonology.
- The sort of question a linguist might ask can be illustrated by such examples as *Does the mind think of speech as a string of discrete segments?* or *Does the phonologist's segment have psychological reality?*. An important point to remember is that whether or not we are consciously aware of the units and processes in the mind is irrelevant to the notion of psychological reality: to say that something is psychologically real does *not* imply that we are conscious of it.

Diphthongs are regarded as long vowels whose tongue position changes during the diphthong's pronunciation. Thus they are characterised in terms of their start and end points by representation as two symbols, e.g. [ai], [ou]. etc.

1.4 Claims of the Anatomical Model

We are dealing with a descriptive model of vocal tract configurations. The little theory there is to support this makes a few weak claims about articulation:

- The separate treatment of consonants and vowels by the use of a different kind of classification makes the weak claim that there is some special but unspecified *difference* between them.
- It is useful and appropriate to use a feature representation for the classification. This means that it is possible to identify groups of articulations which share features. The groupings are labelled as classes according to the shared features. Psychological reality is implied for these classes.
- The features are *mutually exclusive* (i.e. no symbol may appear in more than one cell of the matrix).
- In the classification of vowels two types of feature are needed: those concerned with the *place* of the primary articulation within the vocal tract, and those concerned with the *manner* of the articulation. The representation is therefore two dimensional. A third dimension consisting of a single binary feature (*voicing*) is brought in to distinguish those segments which share place and manner of articulation but which use phonologically the voiced/voiceless opposition (traditionally equated with presence or absence of vocal cord vibration at the phonetic level).
- The classification of vowels is sufficient on a two dimensional cross-section place through the oral cavity (or some abstraction of it), though, as with consonants and voicing, a third dimension is sometimes brought in to distinguish between *lip-rounded* and *lip-spread* contrasts.
- A defect in the classification is that there is no clear way of stating that certain cells might in fact be *unfillable*. That is, a *formal account* of why it may not be possible for particular cells on the matrices to be filled is missing. Some may not be filled by chance in any particular language, but for others it may be the case that it may simply

not be possible to have a particular segment there. These days we would feel it necessary to seek an *explanation* for the impossibility.

- There is implication in the classification of consonants that there is *independence* between the feature subsets of place and manner. This implication is not substantiated. The lack of a clear statement on this claim leads on the another implied claim: that features are able to be specified *independently* or without reference to one another. At the level of articulator control it implies that in terms of the manipulation of the anatomy independent control of the positioning of the various organs is possible.

[*footnote:* In the section on *Articulator Control* we shall see that this implication is false.]

1.5 Transcription

Phonetic transcription is a means of recording on paper the way people pronounce. Several systems of transcription have been devised, but the most widely accepted one is that proposed, and periodically revised, by the International Phonetic Association. The system is orthographically based, meaning that it uses individual letters, or sometimes a pair of letters, to represent *symbolically* a particular sound. The symbols are strung together, as in normal orthography, representing the stringing together of sounds to form a complete utterance. Some of the symbols bear a resemblance to those used in orthography, others are specially designed for the phonetic alphabet. But it should be remembered that these symbols are a representation of sounds and should be thought of as sounds, *not* as letters used in ordinary orthography.

A major principle of phonetic transcription is that the symbolic representation is intended to be unambiguous: each sound has its unique symbol, and no symbol represents more than one sound. For variations on a particular sound a system of diacritic marks has been devised for placing on the basic symbol. This is intended to indicate that a variant sound is not a completely different sound, and, more practically, to save having to have hundreds of different main symbols.

One of the original ideas coming from the distinction between main symbols and diacritic marks to indicate variants was that there could be a number of different levels of transcription for any one utterance. Thus there could be broad and narrow transcriptions which differed in the amount of detail of the sounds being represented. At one extreme a broad transcription would use only the basic symbols with no diacritics; at the other extreme as many details of the individual sounds as could be distinguished by the transcriber would be represented using as many of the diacritics as necessary.

It was often said that a broad transcription represented only the phonemic structure of the utterance, whereas a narrow transcription included allophonic detail. This is not the place to discuss what is meant by the term phoneme, except to say that phonemic transcription for phoneticians usually meant using the symbols as general labels corresponding to classes of variants. Thus the symbol /t/ would be used in a phonemic transcription to represent all variants that could receive this phonemic label. A narrow transcription would substitute symbols with diacritics to show the allophonic variants, such as [tʰ, t̪, t̬] (respectively, these mean: an aspirated [t], a dental [t], a voiced [t].

Although we use phonetic transcription today to jot down a quick reminder of the way something was or might be pronounced, we do so bearing in mind that in the light of modern theory it is difficult to justify doing so.

1. One reason for this is that instrumental work can easily show that our *subjective* assessment of a pronunciation can be very different from the *objective* facts, and that this subjective assessment can vary widely from one transcriber to another, making it difficult for someone to interpret someone else's transcription.

2. A second reason is that fundamental changes in the theory question the claim that speech is a string of readily identified separate sounds.
3. A third reason is that our ideas concerning the relationship between *abstract objects* (like phonemes) and *physical objects* (like the articulators or soundwaves) have changed, and we no longer have the same definitions for terms like phoneme and allophone.

As mentioned above, by all means use phonetic transcription to assist remembering how something is pronounced, or to illustrate to someone else the pronunciation of an utterance, but in doing so remember that there are sound reasons why transcription of the traditional kind no longer enjoys the theoretic support it once had. This booklet is not the place to go into the details of the International Phonetic Alphabet: there are several good textbooks which more than adequately cover transcription and give many examples of its usage.

2. ARTICULATION

2.1 Anatomy and Physiology

Lungs

The major source of energy needed in speech production to produce a soundwave is compressed air coming from the lungs and passing through the vocal tract. Muscles on either side of the lungs compress them, with the result that the air inside rises in pressure. A pressure differential is established – the air pressure within the lungs becomes higher than the air pressure outside the vocal tract. As a result of the pressure differential, air flows from the lungs into the oral cavity and then to the outside world. The air flows from that part of the system with the highest pressure to that part with the lowest pressure. The airflow is said to be egressive when its direction is *from* the lungs and out of the mouth or nose. It is said to be ingressive when the flow is *into* the lungs.

Ingressive airflow, used in breathing in, is also used in speech, though comparatively rarely. The lungs are caused to expand to create an internal air pressure lower than that in the air outside. The result is that air flows inwards from the outside, passing through the vocal tract. Ingressive airflow is used comparatively rarely in speech.

Larynx

The larynx is a cartilage structure within which are situated the vocal cords. The vocal cords are attached to the arytenoid cartilages which are able to rotate under contraction of the attached musculature. It is this rotation which results in tensing or relaxation of the vocal cords. Thus the mechanical system responsible for control of vocal cord tension has three elements:

- a musculature,
- the arytenoid cartilages,
- the vocal cords.

In men the average length of the vocal cords is around 23mm, whereas it women it is around 17mm. Notice that they are not tensed by direct control, but indirectly by means of the leverage of the cartilage system they are attached to.

In the production of vibration (voicing) the vocal cords are first tensed using the control system, and held tense while air from the lungs is forced between them. The theory which describes how the vibration occurs is called the aerodynamic myoelastic theory of vocal cord vibration, and involves three stages:

1. The glottis (the area between the vocal cords) begins to close as the result of muscle contraction which applies a force to the arytenoid cartilages to which the vocal cords are attached.
2. Air is forced through the glottal constriction under pressure from the lungs. Due to the Bernouilli effect (whereby there is a pressure drop between the vocal cords) and the myoelastic forces from the vocal cords (which tend to operate towards pulling the vocal cords together), the glottis closes.
3. Since the vocal cords are closed again and airflow is stopped, the subglottal air pressure increases due to continuing flow from the lungs.

One cycle is now complete and the state of the system is such that it will now repeat*: 'a b c a b c a b ....'* and so on. The cycle continues, each pass causing one vibration, until the balance between myoelastic tension within the vocal cords, supraglottal air pressure and subglottal air pressure is disturbed. *The balance between subglottal air pressure, supraglottal air pressure and vocal card tension is critical.*

On average a male speaker's vocal cords vibrate during voiced sounds at somewhere between 80 and 200 times each second (Hz), and a female speaker's roughly between 140 and

380 times each second. The rate of vibration of a child's vocal cords is much higher on average.

Two further modes of operation are possible with the vocal cords:

- narrowing of the glottis accompanied by forced airflow to produce frication (not friction): whisper. In this mode the vocal cords are held apart, but under enough tension to prevent vibration. The laminar airflow beneath the glottis is transformed into turbulent flow as a result of being forced through the narrow gap into a wider cavity above. During whisper this mode of operation replaces the vocal cord vibration normally required for phonologically voiced sounds. It is also used in normal speech for [h];
- holding the glottis tightly closed with sufficient tension to prevent the air from flowing between the vocal cords. When held a specified amount of time, then relaxed suddenly a glottal stop is produced.

Oro-pharyngeal Cavity

Immediately above the glottis is the pharynx, whose muscular walls are normally in a relatively relaxed state. Some languages (e.g. Arabic) do however use this musculature to help produce stops or fricatives in this region.

A little higher and at the top of the cavity we find the velum (ending in the uvula). The velum is the soft, muscular back portion of the hard palate, hence the term soft palate. It can function as a valve controlling the flow of air to and from the nasal cavity. When closed (that is, with the velum raised) the valve prevents air from entering the nasal passage; when open the air flows freely into the nasal cavity as well as through the oral cavity.

The hard palate forms the top of the oral cavity, with the alveolar ridge immediately behind the upper teeth. Together with the upper teeth they are the only two fixed or immobile organs of speech, all others being mobile either under direct control or passively movable.

Mandible

The mandible, or lower jaw, is the only voluntarily movable bone in the head and is used to control the size of the gap between the upper and lower teeth. In doing so, the vertical movement of the lower jaw also has an effect on lip position and in particular on tongue height. It is possible to alter tongue height just by moving the jaw up and down, though normally the tongue musculature and the musculature to move the jaw *cooperate* in producing vertical tongue movement.

Lips

There are three planes to lip movement: vertical, giving rise to lip positions between closed and fully open; horizontal, enabling rounding and spreading; forward/backward, enabling protrusion. Although the direct control of lip positioning is accomplished by only one sphincter muscle surrounding the opening, movement of the mandible and contraction of attached muscles which run from the corners of the mouth back into the cheeks enable opening and spreading gestures. Rounding is by contraction of the sphincter, and protrusion is possible because the sphincter muscle is arranged in three layers which are able to slide forward differentially. Lip opening following closure, say for a bilabial stop, is accomplished mainly by pulling the lips apart using muscles running from the lip corners down the chin.

Tongue

Five locations on the tongue's surface are identifiable as important in the production of speech sounds: tip (or apex), front, centre, back and root. These locations are general areas rather than points on the surface. Although we identify them in order to provide a descriptive framework for sounds produced involving the tongue it is often the case (with vowels, for example) that the entire shape of the tongue is relevant. The tongue's shape is determined by innervation of a complex musculature within the organ.

1.2 Movement

Aside from gravity and other passive effects on movement, the main means of moving the speech organs to achieve the different vocal tract configurations used in the production of speech sounds is the differential contraction of the various muscles involved. Muscular contraction is controlled by innervatory signals arriving *via* nerve pathways – some of which originate in the brain and others in the spinal cord. The control of speech production is complex (see Articulatory Control). Muscles are not usually controlled independently, but are arranged in coordinated groups. That is, they have no independent representation in the brain for the purposes of speech.

1.3 Articulation

Segmental articulation

Free Airflow

Most speech sounds use a pulmonic egressive airstream passing from the lungs through the larynx. The vocal cords are either apart, allowing free airflow into the supraglottal cavities, or they approximate under tension creating the conditions which give rise to glottal vibration (sometimes called spontaneous voicing). This permits the phonological opposition of voiceless and voiced sounds.

The supraglottal cavities (and to a certain extent the subglottal cavities) act as a complex resonator which has the effect of filtering any excitation source. The excitation source can be sited at the larynx (vocal cord vibration or whisper frication), or elsewhere in the vocal tract (in the case of fricatives), or there may be a combination of different sources (see Acoustics). Soft palate positioning determines whether the nasal cavity is excited, enabling the oral/nasal opposition.

Major alterations of the volume and shape of the supraglottal resonator are made by changing tongue positioning. Under non-speaking conditions the tongue is usually at rest, its position being determined by gravitational force and general background tonic activity in the musculature. Immediately prior to speaking the tongue can be observed to assume what is generally called a basic speech pasture: it is positioned higher than in resting and the musculature is activated ready for speech. It is often said that the basic speech posture varies from language to language (that is, is *language specific*), and is determined by the inventory of vowels within the language. From the basic speech posture roughly in the middle of the cavity, the tongue is sent to the various positions needed to provide the correct resonator shape for the different speech sounds it is involved in. It is the alteration of the resonance characteristics of the cavity which determines, for example, the different qualities of vowels.

Stopped Airflow

During vowels and vowel-like sounds the airflow from the lungs through the system is relatively unrestricted, but in the case of consonants there is impedance to this free flow. In the extreme case the airflow is stopped altogether, giving rise to the stop consonants (or plosives): the airflow is briefly checked at some place in the vocal tract. This is achieved in English by:

- bringing the lips together, as for example in [p, b, m],
- bringing the apex or front of the tongue against the upper teeth, alveolar ridge or frontmost part of the palate, as for example in [t, d, n],
- bringing the back of the tongue against the back of the palate, as for example in [k, g].

There are other possibilities in other languages. During this closure phase of the consonant there is no sound produced in the oral cavity, though there may be a residue of sound from the vibrating vocal cords in the case of voiced consonants.

The stop is released actively and suddenly to produce the burst phase of plosives: the air pressure buildup behind the stop causes turbulence of the airstream in front of the place where the stop occurred. Rapid release is used in the normal stop consonants, but a controlled slow release is possible, giving rise to the longer, less turbulent burst characteristic of the affricate consonants, as in [ts, dz, ʧ, ʤ].

Constricted Airflow

Between free and stopped airflow modes we can identify a mode which is characterised by partial impedance caused by a narrowing or constriction of some part of the vocal tract. Air forced through the constriction gives rise to turbulence for as long as the constriction is held. In the lip/teeth area the fricative sounds [f, v] are made in this way in English; using the tongue and teeth we get [θ, ð]; by placing the front of the tongue close to the front of the palate or alveolar ridge we get [s, z]; and a little further back with a somewhat wider gap we get [ʃ, ʒ].

Vowel Duration and Amplitude

Vowels can be observed to vary in duration. Compare, for example, the words *heed* and *hid* – [i] is said to be longer than [ɪ]. Length is one of the features used *phonologically* to assist in distinguishing between vowel segments. In English vowels can be both long and short, but in French, for example, vowels are all said to be short. Thus English $[i]_{Eng}$ is longer that French $[i]_{Fr}$.

However, although for phonological purposes a simple distinction between long and short is adequate, at a more objective phonetic level we can observe that there are systematic durational differences between vowel sounds even when they are all, as in the case of French, phonologically short. These differences are intrinsic and are caused by non-linguistic factors in the way the different vowel sounds are actually made. Since they have no linguistic function (that is, are not used phonologically), they are linguistically irrelevant and go unnoticed by speakers and listeners. In its phonetic realisation phonologically determined length is said to be *overlaid* on the intrinsic durational characteristics of individual vowels.

Similarly, different vowel sounds have different intrinsic intensity or amplitude. For example, [ɑ] is intrinsically greater in amplitude than [ɪ]. These differences, which are once again determined by physical constraints involved in the different way in which vowel sounds are produced, are not linguistically productive and therefore go unnoticed. But just as length can be used phonologically, so different degrees of amplitude can be overlaid on vowels. So, whatever its intrinsic amplitude, any vowel can (by increasing subglottal air pressure) be made to sound louder or less loud. This actively overlaid change to intrinsic amplitude can be used phonologically as one of the ways of marking stress or prominence.

The terms length, loudness and stress are *subjective*; the terms duration and amplitude refer to physical *objectively measurable* quantities (see Perception). Duration and amplitude can be measured absolutely, whereas the subjective counterparts are determined by the human being on a relative basis. So we might say that a certain vowel has a measured duration of 150ms, whereas another has a duration of 180ms. But we would equally refer at a more abstract level to the fact that the second was simply longer than the first, since that is all that might matter phonologically.

If the measured durations of 150ms and 180ms were found for, say, the same vowel in the same word, but spoken by different speakers, the vowels would have the same phonological length because, despite their different physical durations their phonological length would be functioning in the same way for both speakers.

Nasals

In terms of place of articulation, nasals correspond to the stop consonants, and since they are accompanied in English by vocal cord vibration, each can be thought of as the nasal counterpart of a particular voiced stop. Thus [m] is the nasal counterpart of [b]; [n] is the nasal counterpart of [d]. The difference is that the velum or soft palate is lowered, allowing airflow into the nasal cavity, which in turn causes excitation of the nasal resonance. Unlike the resonant effects of the oral cavity, nasal resonance cannot be made to vary in frequency: there are no moving parts to alter the size and shape of the nasal cavity.

Although during the production of nasals airflow is free through the nasal passage, they are usually grouped with the stop consonants because there is oral stoppage of the airflow – air flows only out of the nose. The nasal is said to be released when the stop in the oral cavity is pulled apart. Sometimes oral stops (like [p, b] and [t, d]) have nasal release – that is, they are released not by pulling apart the primary articulators causing the oral air stoppage, but by lowering the velum and releasing the air pressure into and through the nasal cavity. In some accents of English nasal release of oral stops occurs in words like *button* or *happen*.

Laterals

Laterals are articulated with partial closure of the oral cavity made by raising the tongue. However the tongue shape is such that air can flow relatively freely round its sides – hence the term lateral. The shape is not necessarily symmetrical about a centre line, with the consequence that for some people the airflow is unilateral, round one side only of the tongue.

Because laterals are continuants (that is, they can be produced for as long as a pulmonic airstream can be continued), some phoneticians class them phonetically as vowels. This can be misleading because they function as consonants phonologically.

Approximants

Approximants are sometimes thought of as vowels because their characteristics are quite similar. Phonologically, however, unlike vowels they are unable to be used as syllable nuclei. That is, they function like consonants and can only be used in conjunction with vowels in a syllable.

English Consonants in Articulatory Terms

Consonants are basically obstruent, involving *partial* (in the ease of fricatives) or *total* (in the case of stops) closure of the vocal tract at some point, causing impedance to airflow. The airflow itself is always pulmonic (originating in the lungs) and egressive (the flow is toward the mouth). They can be reliably distinguished phonetically from each other along several independent parameters to provide a quite large phonological inventory of sounds usable in the language.

Place of articulation

The partial or total closure of the vocal tract can be made in a number of places: lips, teeth, alveolar ridge, back of the palate, or at the vocal cords themselves (glottal stop, [h] and whisper).

Manner of articulation

There are three major types of manner used to distinguish between consonants:

- plosive: involving complete closure and the production of a burst immediate after the stop as the pressurised air is rapidly released,
- fricative: involving partial closure at some place along the vocal tract to give rise to turbulence audible as frication,
- affricate: involving complete closure, followed by slow release to give an audible fricative quality to end the consonant rather than the rapid burst associated with

plosives.

Vocal cord vibration

This can be present or absent during consonants. The stops and fricatives are found both with and without vocal cord vibration, but nasals in English are always accompanied by vibration. By definition, the glottal stop (involving stoppage of airflow by the vocal cords themselves) has no vibration, and similarly [h] (involving tensed vocal cords with a narrow gap between) cannot have glottal vibration. The presence or absence of vocal cord vibration (at the phonetic level) permits the phonological opposition of voicing.

> [*footnote:* We shall see later though that the correlation between phonological voicing and phonetic vocal cord vibration is a loose one.]

There is a third possibility for the vocal cord vibration parameter: partial voicing (strictly vibration) for only part of the duration of the consonant. This is usually caused by assimilation with adjacent sounds which may not normally have glottal vibration. This state is often referred to as devoicing (of an otherwise voiced consonant), but beware the usage of the term voice. This is normally reserved for the phonological parameter or feature: the effect here is, of course, phonetic.

Muscular effort

Some phoneticians claim that in the articulation of consonants which are normally accompanied by vocal cord vibration there is a general reduction in muscular effort involved in all parameters. Many experiments have been conducted to show that this claim is probably false. The muscular effort involved in contracting, for example, the sphincter muscle of the lips to achieve closure during both [p] (with no glottal vibration) and [b] (with glottal vibration) is quite similar for most speakers – indeed some speakers regularly produce [b] with more effort than they produce [p]. In this model the voiceless consonants are usually referred to as tense, and the voiced ones as lax. In *Distinctive Feature Theory* consonants which are [-voice] and usually also [+tense], whereas [+voice] consonants are usually [-tense] – this idea is probably carried over from the early phonetic model.

Coarticulation

Coarticulation can be roughly defined as the effect of the influence of an articulatory segment on adjacent segments. Two subdivisions of coarticulatory effect are made:

- left-to-right, or carry-over effects, in which properties of a segment carry over to influence the characteristics of the following segments;
- right-to-left, or anticipatory effects, in which some of the characteristics of a segment influence those of earlier segments.

Coarticulation is *universal* in the sense that in all languages neighbouring segments interact phonetically with one another, but the extent of the effect and the balance of direction of the effect vary from language to language. There are considerable coarticulatory effects observable in English, with right-to-left effects being commoner than left-to-right effects.

Some researchers have linked coarticulation with the so-called Principle of Least Effort. The idea here is that speech production at the phonetic level need be only as accurate as is necessary to communicate to a hearer the required segmental, and hence morphemic, contrasts to enable meaning to be transferred. This idea assumes that the most accurate realisation of a phonological string would involve the precise rendering of the articulatory and acoustic features which make up individual segments: they would not blend with each other and each would be fully realised. Because phonological segments and their phonetic correlates are generally over-specified and contain redundancy the information they encode can be communicated even if phonetically segments fall short of full realisation. Since, from

the point of view of the motor control of speech, accuracy and precision are therefore *less* than completely necessary, the principle of least effort holds that they will be relaxed as far as possible whilst maintaining a good level of communication. Relaxation of the precision of motor control results in segments running into one another, and target positioning of the articulator being missed on occasion. We say that a balance is struck between using the least effort possible to render the articulation and the need to realise the articulation sufficiently accurately to prevent loss of communication (see Articulatory Control).

An earlier term, assimilation, was used for the phenomenon, now called coarticulation, at both the phonological and phonetic levels. In general the modern usage is to reserve assimilation to refer to *phonological* influences of one segment on another, and coarticulation to refer to *phonetic* influences on adjacent segments. Phonological assimilation reflects the phonetic tendencies of coarticulation, but is voluntary. Phonetic coarticulation describes effects which are *not* under voluntary control – though the degree of the effect can often be manipulated (see Cognitive Phonetics).

From the theoretical point of view the notions of assimilation and coarticulation are interesting because they rely heavily on the idea that speech at both the phonological and phonetic levels is made up of a string of discrete segments, blended together to produce a relatively continuous articulation and soundwave. In fact there is little evidence of an experimental nature to support the idea that speech is made up of a string of discrete segments which have become blurred together. The main piece of evidence we have is that when questioned about speech people usually refer to it as though they feel it to be made up of individual sounds: those who no nothing of linguistics or phonetics will readily refer to the three sounds in the word *dog* or state that the last two sounds of *dog* are the same as the last two in *fog*. At the cognitive level of speech production the segment appears to have reality. It is not necessarily the case, though, that the segment has reality at the physical level.

The usual model of speech production at the phonetic level does however assume the reality of the segment. Speech is said to consist of strings of gestures of the vocal apparatus which are realisations of canonical targets. In the articulation of isolated, steady state segments these targets are said to be *fully realised*. When the segments are strung together execution of the targets is less than full: targets get missed as assimilatory and coarticulatory effects are introduced. The effects are progressive in the sense that the more we depart from the ideal of isolated steady state segments the more the effects occur.

Phonetically, in coarticulation the predominant influence on the extent to which ideal targets are missed in running speech is time. The greater the rate of utterance, the greater the degree of coarticulation. This suggests that the effects are mainly mechanical, since mechanical systems are particularly sensitive to constraints such as inertia and friction which tend to smooth out and blur the precision of rapid or finely detailed movements. The accuracy of motor control is heavily influenced by rate of utterance. Motor control failure at a higher level than the control of the mechanical system results in the slurring of speech, for example under the effects of alcohol or other drugs which might affect the central nervous system or the response of the musculature to neural impulses.

Prosodic Articulation

So far we have been discussing some of the more important aspects of the articulation of the segments which are strung together phonetically in speech production. There are however additional elements of speech production which *span* segments, and which operate irrespective of the particular segments in anyone utterance. These features are called prosodic or suprasegmental. They are apparent in changes of a speaker's rate of utterance (the varying speed of speaking), the distribution of stressed elements within stretches of speech larger than the segment, the rhythm associated with strings of segments, and, during the course of long stretches (up to sentence length) of the utterance, the changes in rate of vocal cord vibration associated with the intonation pattern.

We shall look at these respectively under their traditional labels: length, stress, rhythm,

and intonation. These labels refer to abstract phonological phenomena, but are sometimes used by phoneticians when they are referring to the linguistic function of the physical effects observed. It will help avoid the confusion of levels, which is more likely to occur when dealing with prosodics than with segments, to focus on the idea that whether we are dealing with segmental or suprasegmental effects we always model speech production as a process involving two distinct stages or levels. We shall call these assignment and interpretation.

Assignment

The assignment of particular segments for the overall sound shape of a word or longer stretch of material, and the assignment of prosodic features to span that string of segments are cognitive processes which arc described abstractly in linguistics within the phonological component of the grammar. As cognitive processes they are free of physical constraints such as the variability inherent in the vocal apparatus and its control mechanism. At the level of assignment such considerations are irrelevant. This is what is meant when a linguist speaks of idealisation: abstraction to a level where variability is not under consideration.

Decisions are taken cognitively as to what segments shall be used eventually (at the lower phonetic level) to produce a soundwave appropriate to encoding, say, a sentence. As a parallel operation, decisions are also taken as to length, stress, rhythm and intonation to be *overlaid* on the chosen string of segments. These decisions are taken in the light of what the speaker knows about the way the language works in general, and what he knows about how to encode some extra-linguistic phenomena such as emotion. Usually core theoretical linguistics accounts only for the knowledge base expressing the way the language works in general. Other more peripheral linguistic models such as psycholinguistics and sociolinguistics account for the extra-linguistic phenomena.

Interpretation

The interpretation of the segmental and suprasegmental features which have been assigned comes next. The speaker has to decide how these abstract markers are to be interpreted *physically* such that the correct impression can be reliably decoded by a listener. As before, the decisions have to be taken in the light of what the speaker knows about such matters, but this time we are at a physical level where milliseconds, decibels and Hertz replace the earlier abstractions.

Understanding the interpretation of the prosodic features of length, stress, rhythm and intonation is difficult. The difficulty lies in the fact that these abstract terms do not have one-to-one correlates in the physical world. Thus it is not the case that length correlates just with duration (expressed in milliseconds); stress does not correlate well with amplitude (expressed in decibels); intonation does not equate well with changes in the fundamental frequency of vocal cord vibration. All the abstract features correlate with all the physical features, but in varying ways. So, the answer to the question *What are the physical correlates of the abstract notion of stress?* is *Duration, amplitude and frequency – all three*.

Length

Phonetic segments can be thought of as having intrinsic duration. That is, all things being equal, each segment is timed in milliseconds. Segments have, for physical reasons, different intrinsic durations. For example, in English the low back vowel [ɑ] may have an intrinsic duration of 200ms, whereas the high front vowel [ɪ] may be only around 100ms or often less. These figures reflect, in addition to the physical reasons mentioned earlier, language-specific reasons: the range of intrinsic durations of vowels in English, for example, is much greater than it is in French. These *language specific* differences are part of the tradition of a language, and are overlaid on the physically determined differences (much smaller) between vowels.

But all things are not equal and no segment exists in the real world divorced from adjacent segments in a string forming the utterance. The overall rate of delivery of the utterance affects the underlying intrinsic durations of segments. And segments are affected

differentially. If, for example, a particular utterance is spoken rapidly not all segments are shortened in duration by the same proportion (vowels are generally shortened more than consonants). Double the rate of utterance and you do not halve the length of every segment in the utterance.

The duration of speech segments generally ranges from around 30ms to around 300ms. The just noticeable differences for segmental duration vary from segment to segment, but are between 10ms and 40ms.

Stressed syllables are normally of greater duration than unstressed syllables, by about 50% of their intrinsic durations, though there is considerable variation among speakers. There is a sense in which stress can be thought of as being one of the factors which govern the final duration of a segment contained within a running utterance. Increases in overall rate for an utterance involve changes to the duration of segments within the utterance. Vowels are the prime candidates for shortening to achieve increased rate, but vowels within unstressed syllables shorten much more than those in stressed syllables (because stress itself has a lengthening effect). This is a good example of how abstract prosodic features cause interaction of the various physical parameters at the phonetic level.

Stress

There is a tendency to think of the physical correlate of the abstract prosodic feature stress as being the amplitude of the soundwave, or alternatively the amount of effort put into the articulation. But we have seen above that a major correlate of stress is in fact an increase in *duration* of the particular segment. The differentiation of stressed and unstressed vowels (and therefore of syllables) is complex. In fact, experiments have shown that manipulation of the duration is sufficient to produce a differentiation between stressed and unstressed vowels. As a result of this finding it is quite common, for example, for synthetic speech to use only the physical parameter of duration to interpret the assignment of stress.

> [*footnote:* In early synthetic speech systems amplitude manipulation was much harder than duration manipulation.]

In addition stress is perceived, or decoded by the listener, when the vowel nucleus of a syllable is given an unexpectedly high fundamental frequency by increasing the rate of vocal cord vibration, or by causing a sudden change in the rate of vocal cord vibration within the duration of the vowel.

In the interpretation of assigned stress in speech production all three acoustic parameters may be brought into play, often in different combinations. Similarly for the listener, stress may be decoded when one, two or all three parameters are adjusted in the way described (greater duration, higher amplitude, change of fundamental frequency). The exact combination and ratio of the parameters has not yet been satisfactorily modelled since there is so far insufficient data to enable an understanding of their relative roles. One reason for this is that the balance between these parameters seems to vary.

Rhythm

The abstract prosodic feature of rhythm cannot be modelled except by incorporating the features stress and length. One reason for this is that rhythm is defined in terms of the regularity or patterning of the occurrence of stressed syllables within the utterance. Do remember, though, that we are at an abstract level: confusion over this point has led several researchers to make mistaken observations concerning rhythm. At this cognitive level we are concerned with what speakers and listeners *feel* about prosodics, not with what they actually do. In terms of the processes involved we are concerned with the *abstract assignment* of rhythm, not its *physical interpretation.*

Native speakers of English *feel* that they assign rhythm such that its interpretation results in stressed syllables falling equidistant from each other in time – they are isochronous. With

respect to rhythm, English is said to be a stress timed language. Some languages, on the other hand assign rhythm with a view to all syllables, whether stressed or not, being equidistant from each other in time. Such languages are said to be syllable timed, and examples would be French and Greek.

Several researchers have been able to show in the laboratory that in fact in the resultant acoustic waveform the isochrony is not there as regularly as was believed: there is quite a lot of variation in the timing, and therefore in the realisation of rhythm of sentences. Some have taken these results to falsify the notion of isochrony. But people's intuitions at the cognitive level are not so easily falsified. There are numerous examples in segmental phonology and phonetics where there is no one-to-one correspondence between phonological assignment and phonetic interpretation, and these are readily accepted as non-anomalous. For example, the distinction between the words *writer* and *rider* in American English. Although phonemically the distinction is in the /t/ *vs.* /d/ opposition, the soundwaves of these words are distinguished not on this consonantal segment, which is identical in both words, but on the soundwaves corresponding to the preceding diphthong which has greater duration in *rider* than in *writer*.

There is no reason to suppose that wide variability and the transfer of correlation between features at different levels should be any different for prosodic features than for segmental features.

Intonation

At the phonetic level intonation is generally interpreted by varying the rate of glottal vibration during an utterance. This is perceived by the listener as a patterned suprasegmental movement of pitch which is linguistically significant. For example, a rising intonation (increasing rate of glottal vibration) signals that the utterance is a question in the absence of subject-verb inversion or a *wh-* word; a falling intonation (decreasing rate of glottal vibration) signals that the same utterance is a statement. Compare the normal pronunciations of *John has gone* and *John has gone?*

The listener is also able to perceive effects which are not linguistic. So for example altering glottal vibration according to one particular pattern will convey that the speaker is surprised, another that they are angry or using irony, and so on.

The phonological assignment of intonation is complex and there are several current models. At the *phonetic* level the complexity is increased by the fact that there are constraints on the range of changes in glottal vibration which are available to interpret the intonational assignments, and the fact that these constraints alter during the course of the phonetic realisation of an utterance. For example, at the beginning of an utterance, because at this point the volume of air in the lungs is at its maximum for the utterance, the upper rate of glottal vibration available is at its highest. As the utterance proceeds and the volume of air available becomes less the available upper rate declines. This means that in an utterance several words long, a high intonational level might well be physically lower towards the end of the utterance than a previous 'low' intonational level. Since the cognitive assignment is the same high at both points, the perception of high and low intonation levels must be relative against the declining physical level.

2.4 Articulator Control

Besides the anatomical viewpoint in articulatory phonetics, we also have to consider articulator control. The anatomical model said nothing about how the articulatory configurations of the vocal tract are achieved, and nothing about the mechanism or functioning of any control system for articulator movement. It seemed enough until relatively recently (the 50s and 60s) to leave the whole matter of articulation at the descriptive anatomical level.

> [footnote: You can readily see by examining the phonological feature labels in Distinctive Feature Theory, for example, how often one discipline or part of a discipline can lag another.

Thus, even in The Sound Pattern of English Chomsky and Halle base their feature set partly on this early anatomical model, although both theoretical and experimental phonetics had already progressed to a more dynamic control model. In fact. in phonology even in the 80s we find little to reflect the progress in phonetics. This is not to decry phonology, for indeed the reverse is also true: too much of 80s phonetics has not taken account of the considerable developments in phonology since 1968.]

Movement is the keyword here. The articulators move – indeed x-ray videos seem to present a picture of almost *continuous* movement, especially of articulators like the tongue and jaw which are involved in articulating almost every segment. We must begin though by being very careful: we may observe visually (perhaps with the help of x-rays or other experimental techniques) movement of, say, the tongue, but in fact the tongue is the name given to an anatomical organ the movement and shape of which are not *directly* under control. Beneath the surface of the tongue and other articulators lies a complex musculature, and it is this which is controlled to produce movement and shape.

Even the contraction or tensing of a single muscle is more complex than it might appear visually. A muscle consists of a sheath or outer covering beneath which are hundreds of individual muscle fibres. It is these which are ultimately under innervatory control from the brain's motor cortex. Muscle fibres are recruited to participate in the overall muscle contraction.

When a muscle fibre receives a neural instruction to contract, three interrelated events occur:

- mechanical contraction,
- chemical reaction,
- electrical discharge (resulting from the chemical reaction).

The mechanical contraction is *all-or-none*. That is, whenever contraction occurs it is total: a muscle fibre cannot contract partially. Usually this contraction results in a shortening of the muscle fibre by around one third its normal length. The apparent paradox of all-or-none contraction of individual fibres and the graded (or analog) contraction of the whole muscle is explained by the operation of two mechanisms:

- There is control of fibre firing rate. That is, the firing rate of individual fibres can be varied from occasional firing up to an upper rate determined by the fibre's speed of recovery from the previous firing. Immediately following firing the recovery period begins during which the muscle fibre returns to its original mechanical, chemical and electrical states. Firing cannot recur (even if an innervatory signal arrives) before near completion of the recovery period.
- There is progressive recruitment of muscle fibres. The number of fibres recruited (or brought into play) for a particular overall muscle contraction can be varied. Thus 50% of the number of fibres available might be recruited to achieve 50% overall contraction, 20% to achieve 20% contraction, and so on.

In practice both mechanisms operate together, though the relationship between them is not fully understood.

The neural signals innervating muscle fibres have an all-or-none character: they take the form of pulsed electro-chemical activity which can be shown graphically in a stylised way:

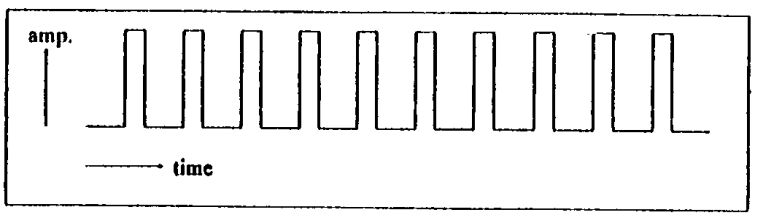


Fig. 4 Stylised graph of neural signals.

These signals have one or two special characteristics:

- the width of duration of each pulse is comparatively short and does not vary,
- the height or amplitude of the pulses does not vary,
- degree of required muscular contraction is coded by how often these signals occur: increased frequency signals more contraction, decreased frequency less. That is, the required amplitude of the contraction is coded as the frequency at which the signals are sent to the muscles.

The signals controlling muscular contraction are said to take a digital or binary format because of their discrete, on/off, all-or-none nature. Likewise the behaviour of muscle fibres as described above is digital in nature. But the behaviour of the *overall* muscle is clearly not pulsed and not binary: smoothness and continuous variation, typical of an analog system, characterise whole muscle behaviour. What has occurred is digital-to-analog conversion (DAC): the digital behaviour of individual muscle fibres has been converted to the analog behaviour of the whole muscle. The DAC is accomplished mechanically by an arrangement of the system which permits asynchronous firing of the muscle fibres. This, coupled with the elasticity of the muscle contents, has the effect of smoothing the abrupt, jerky nature of the firing of individual fibres.

The Control System

Two basic types of general control system are possible contenders for modelling articulation control:

1. There are those systems which assemble very detailed information about how to perform the required effect (in this case articulatory movement), and use this information to send carefully organised and accurately detailed control signals, knowing that these signals will be sufficient to achieve the desired objective. This is referred to as the comb model.
2. The other type of control system involves the sending of coarse signals which are less detailed and which rely on local checking (monitoring) and adjustment by the device itself (in this case the musculature). This latter is referred to as the chain model.

In the comb model of control the results of the innervating or control signals are not monitored: the system simply assumes that the calculations which form the basis of the control signals are accurate and that the signals themselves will be accurately obeyed or interpreted by the peripheral device.

[footnote: in sea navigation systems such a system is referred to as dead reckoning.]

In the chain model constant monitoring (feedback) of the results of control signals leads to ongoing correction of any signal or peripheral device errors which may be due to the less detailed commands. Such a system minimises the advance calculation of detailed control signals, but involves the monitoring overhead.

From around 1965 to 1970 there was much discussion among researchers as to which of these two models most appropriately described the behaviour of the speech musculature control system. Ultimately it seemed that the chain model (with its monitoring and feedback systems) was the most appropriate, though some comb model based control was not ruled out.

Feedback

If the control system incorporates some monitoring subsystem then feedback mechanisms must be available. In speech production we can identify three major feedback mechanisms which seem to play some role in governing control:

- auditory feedback,
- tactile feedback,

- intra-muscular feedback.

Auditory feedback consists of detecting how the production system is doing by monitoring the resultant audio waveform. We hear the sound *via* two pathways: it can be either airborne or conducted through the bones of the jaw, etc., to the auditory mechanism. Feedback of this kind is characterised by being very slow and usable over only comparatively long periods of time (i.e. longer than syllables or words). We would predict therefore that any effects based on this mechanism would concern long term aspects of speech above the level of segment. Indeed experiments show that if people are deprived of auditory feedback there is some deterioration of their ability to control *suprasegmental* phenomena like intonation (i.e. deprivation of auditory monitoring encourages *monotone speech*). Long term timing control also suffers, giving rise to loss of rhythm and the correct relationships in the timing of segments.

Tactile feedback is provided in general by pressure sensors. There are nerve endings present on the surface of the speech organs which are sensitive to pressure variations, and which generate signals when pressure changes occur. Such pressure changes result when articulators touch. There are even very sensitive sensors in the oral cavity capable of responding to small changes in air pressure. All this tactile information is continuously fed back to improve effectiveness of control. It is however still comparatively slow (though not as slow as auditory feedback). Experiments depriving subjects of tactile feedback by application of mild surface anaesthetics show a segment-by-segment deterioration of speech resulting in a drunken-like slurring.

Intra-muscular feedback is the fastest of the three types and is potentially usable within the timespan of a single segment, though there has been some argument on this point. This speech is achieved by having sensors within the muscles themselves, and by the fact that the response is reflex or automatic with only a minimal secondary role being played by any cognitive processing of the feedback information. The mechanism for the reflex intra-muscular monitoring and response is the gamma loop.

The Gamma Loop Feedback System

Within a muscle, besides the normal muscle fibres discussed earlier, there are special fibres called muscle spindles. A primary role of these muscle spindles is to sense stretch (actually, rate of stretch) of the muscle. They generate signals proportional to any stretch that occurs, and these are sent from the muscle by specially designated nerve fibres called gamma fibres.

> [*footnote:* The ordinary nerve fibres responsible for the general control described earlier are called alpha fibres.]

Before reaching any area of the brain where cognitive, activity might occur, these signals are turned back automatically to travel down the alpha fibres back to the muscle – thus modifying the normal innervatory signals. The entire loop the feedback signal travels is called the gamma loop, and is an example of what is called a reflex arc. Experiments based on deprivation of gamma feedback in speech are difficult to design and carry out. Results have been relatively inconclusive as to the actual role intra-muscular feedback might be playing in speech production.

3. THE ABSTRACT THEORY OF ARTICULATORY CONTROL

3.1 Introduction

The overall physical model of dynamic articulation involves a range of areas from anatomical and mechanical description of the speech organs themselves through to the computations which must be achieved cognitively to feed a control system accounted for by neuro-physiological evidence. This is a complex system (and system is the word to emphasise) drawing on various disciplines for its characterisation in phonetics. This model is essentially a mechanistic one. That is, the model is constructed around what we can learn by experiment of the nature and functioning of the mechanisms concerned.

But there is another kind of model which can be built. This is an abstract model, focussing much less on the mechanisms, and attempting to arrive at a plausible abstract explanation of the results of the mechanism's function. For the most part linguistics itself is just such an abstract theory: the mechanistic counterpart is neurolinguistics which, for the moment, has not been developed much, but which would seek to describe and explain language from the point of view of the neural mechanisms involved in the brain. For the moment very little is known factually of these mechanisms because of the difficulties of experimental work. But since the early 50s there has been increased activity in the modelling of neural mechanisms in general.

This neural network modelling, as it is known, has been paralleled by important developments in the modelling of cognitive processes using networks. The more usual term in the cognitive sciences is connectionist modelling or parallel distributed processing, and the techniques involved abandon rule based systems in favour of the network paradigm, employing a radically different mathematical approach. The cognitive aspects of linguistics and phonetics are included in the areas of cognitive science that many researchers are investigating, using the new techniques.

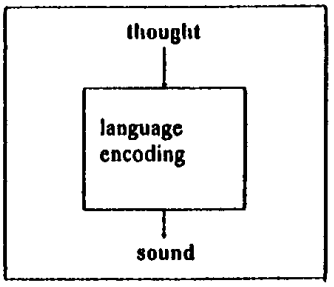


Fig. 5 Thought to sound encoding.

3.2 Translation Theories

General

The abstract model we are now going to examine reflects the essentials of what has come to be known as Translation Theory. The major distinguishing characteristic of translation theory is that information flows from the system's entry point to its exit point. This process can be described in several ways. Thought is translated into sound, or thought is encoded as sound, or thought is mapped onto sound. The encoding is seen as a multi-layer process in which at each layer there is a representation of the information translated from the previous layer's representation and to be translated to the next layer's representation. The original information or thought is thus carried along through the various layers, undergoing translation after translation to give representation after representation as far as the final acoustic signal. This is the general model in linguistics that you are studying.

In the transformational generative model the individual components of the grammar represent what a speaker/hearer knows about the semantics, syntax and phonology of their language. The sets of rules are *not* descriptions of actual procedures during an act of per-formance. They are a descriptive characterisation only of what a speaker of the language must

know to perform language tasks. Though strictly not correct it is nevertheless helpful to imagine an ideal performance grammar which is equivalent to a performed competence grammar, but without any of the special considerations unique to a performance grammar *per se*. The phonological and phonetic processes referred to here are strictly representations of a speaker's knowledge of the regularities in the language though it may be helpful to imagine them as steps in some idealised performance. In real performance other facts outside the scope of linguistics come into play which in some sense degrade this idealised performance to what we can actually observe.

Thus for example the underlying representation of a sentence (or any other string) at the entry point to the phonology is translated into (or mapped onto) the derived representation by rules of the phonology. These rules govern changes to or modifications of the underlying representation but do not essentially add any new information necessary for satisfactory encoding.

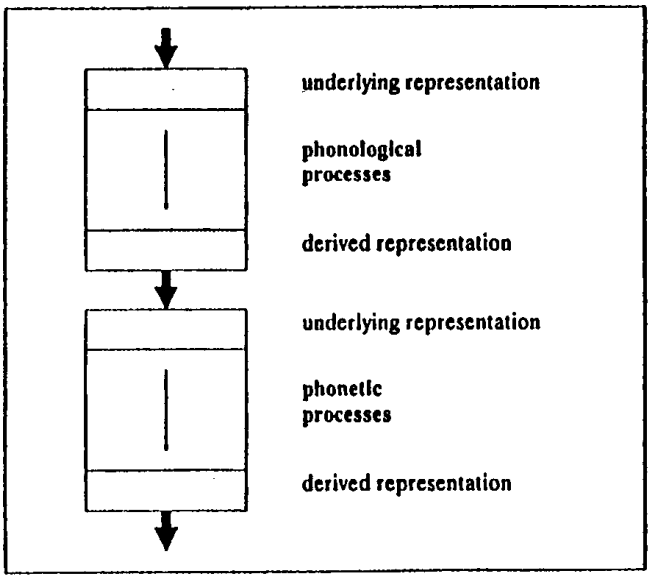


Fig. 6 Phonological and phonetic processes.

The underlying representation of the phonology, as well as the derived representation, comprises a timeless or static description of a string of abstract segments. The derived representation forms the entry level representation of the phonetics.

Phonetic processes characterise the translation of the underlying representation of the phonetics into its derived representation. The phonetics itself is multi-layered (as are other components, of course). Thus we recognise that at the start of phonetics the representation is abstract and mental (or cognitive) and comprises a string of phonologically derived segments.

Now begins a series of processes which account for the translation of this string of discrete abstract segments into a representation of a continuously changing pattern of sound (air pressure variations). The overall process is a complex one.

Notice immediately several major incompatibilities between the input and output representations of the phonetics.

- The input is a representation of something cognitive and the output is a representation of something physical (or, respectively, abstract and concrete).
- Time is introduced somehow during the phonetic encoding.
- The discrete character of individual segments at the input becomes lost to a continuous soundwave at the output in which it is not possible to find any indication of earlier separate segments.

Phonetics has the job of accounting for these discrepancies, and explaining how the various

mechanisms we have been discussing play a role in this translation process.

Time in Translation Theories

The introduction of real time (as opposed to the notional time of phonology) is very controversial: we are by no means sure how to place this in the model. One idea is that incorporated into the production system is a rhythm generator operating on syllable sized portions of the incoming phonological representation to pace the articulations. Syllables are hard to define, but are usually thought of as possessing a [+vocalic] or [+syllabic] nucleus (or focal point) surrounded by a few consonants. The rhythm generator paces the articulation of the focal point nuclei, allowing the consonants to take care of themselves. Such a phonetic process requires the identification of individual syllables within the phonology, some have argued, and those models of phonology which do not say much about the syllable would clearly be of little use feeding such a phonetics. Several comparatively recent phonological models however do characterise in some detail the syllabic structure of the derived level representation.

Coarticulation

The blurring, running together or overlapping of individual segments to produce a smooth continuous sound output from the phonetics has received considerable attention since the early 60s. Various models have been proposed, mostly centering around the phenomenon known as coarticulation. In earlier phonetic theories coarticulation referred to the simultaneous articulation of certain segmental features, but in the modern theory coarticulation is about the overlapping of discrete segments to produce a continuous signal. Notice how any model which incorporates the notion of coarticulation begs the question as to the nature of the input to the phonetics, by simply assuming it is segmental. That is, the decision to account for the coarticulation of segments presupposes that there are segments to be coarticulated. Although there is general agreement that we should model the system around the notion of segment there is little hard evidence to show that this is factually correct.

Much of coarticulation is thought to result from mechanical effects. The organs of speech obviously have mass, and this naturally entails a certain degree of inertia in their movement. A mechanical system like the vocal apparatus is clearly going to be more cumbersome than the physical correlates of the mental processing which controls it. Mental processing occurring within an electro-chemical system is fast and can accommodate rapid and abrupt changes from one segment to another, but the vocal apparatus requires much more time to move from one segment to another because of the mechanical inertia of the system and the mass of its component articulators. It is imagined in the theory that the mental system drives the mechanical system just a little too fast for it to satisfactorily accomplish these abrupt changes from segment to segment, resulting in a blurring together of segments and consequent loss of their boundaries and individual identities. On occasions so much of a mismatch may occur that articulations bear little resemblance to, say, the leisurely (but artificial) articulation of a single isolated segment.

> [*footnote:* The question of cerebral *vs.* mental processing is not being addressed here, though of course it is important. *Cognitive processing* can be understood as an abstract perspective on *neural processing.* One of the advantages of using the new network based models is that the model framework (the mathematical formulation or the net) is common to modelling *both* cognitive and neural processes.]

One obvious question arises here: what is the upper limit of acceptable blurring? This is usually answered by suggesting that in normal speaking the dominant mental control drives the articulatory system as fast as possible, consistent with enabling efficient decoding on the part of any listener. This is quite a claim, for it implies that the speaker is mindful of the listener's perceptual system and takes this into account as an integral part of speaking. In the theory this is referred to as production for perception, implying that no speech is produced without intending it to be perceived, and without adjustment based on the speaker's knowledge of the processing properties of the perceptual system.

Some researchers have attempted to show with more or less success that there are coarticulatory effects taking place at the neuro-physiological level and at the acoustic level, but mechanical and aerodynamic effects continue to be thought of as mostly responsible for the loss of distinct segments somewhere between phonology and sound. In phonetics the term target refers to the articulatory segment which would have been produced (i.e. the one intended) had it not been for coarticulatory effects preventing its full realisation. Those coarticulatory effects which are thought of as mechanical are described in a detailed model using equations derived from mechanics, and are seen therefore as nonlinguistic.

3.3 Action Theory

General

Comparatively recently (since 1975 and gathering momentum since the early 80s) it has been suggested that translation theories as described above are unsatisfactory because they fail to account for some of the observations we have made about speech production. In particular translation theories cannot give a satisfactory account of compensatory articulation, whereby a speaker readjusts his articulation to take account of some external constraint like trying to talk with a pipe held in the mouth. Nor can they take account satisfactorily of cognitively derived intervention in the normal articulatory process.

At a certain point in the development of any science, when enough observations have been made which the extant theory cannot accommodate, the science undergoes a paradigm change. That is, quite suddenly a new theory is proposed which does account for the new observations, and after debate and testing to make sure the new theory is adequate it replaces the old one. Although the proposal of the new theory is sudden the replacement process can be protracted.

A new theory of speech production control was proposed around 1975 by a group of researchers at the *Haskins Laboratories* in New Haven, Connecticut. Acceptance of the new theory is gaining ground as it is modified to the point where it can satisfactorily account not just for the new observations which were earlier unsatisfactorily accounted for, but also everything covered in translation theories.

The new theory, which draws on similar changes in the theory of neuro-physiology, is called Action Theory. And it criticises translation theory on the grounds:

- that speaking does not consist of the handing on of information for re-encoding layer after layer through the process,
- that the amount of information that would have to be added during such a translation process is just counter-intuitively too great, and
- that the neuro-physiological mechanism for action and its functioning in speaking have been misunderstood and wrongly modelled.

These claims form not just a weak departure from established theory, but the basis of a radically new way of looking at speech production.

Action Theory suggest that information processing at the cognitive levels of phonology and early in the phonetics is not in terms of the detailed representations (e.g. bundles of distinctive features) we have been used to in linguistics. Much more it is a comparatively simple handling of broadly based labels (like acoustic targets) describing *gross* effects of articulation. One might imagine instructions like *Do vocal cord vibration!* or *Do vowel-ness!*. A characteristic of such instructions is that they lack detailed information about how to do the actions specified. Action Theorists would claim that this detailed information is itself contained in the way in which the articulatory system itself is structured – so does not need to be specified as part of the higher level instruction.

The articulatory mechanism (that is, the whole neuro-physiology and anatomy of the system) is said to be arranged in structures. These are invoked in the theory as coordinative structures. A coordinative structure is a grouping, say (though not necessarily always), of

muscles which embodies well defined working relationships between its component muscles. In some sense the muscles in a muscular coordinative structure cooperate to fill out and perform the appropriate details of a gross instruction.

How this cooperation or coordination within the structure operates is described in the model by equations governing the working relationships between the component parts (in our example the muscles) of the structure. Using the more usual terminology of computer modelling, we would say that a coordinative structure is *internally* programmed to behave in a particular way. The component parts are not directly or independently controlled. Each operates in conjunction with its colleagues in a well defined way which can be described using an equation.

The speech control system knows that the appropriate detailed contractions, etc., will take place according to the local arrangements as defined by the equations governing the relationships between the structure's components; so it need only issue very gross instructions designed to trigger the coordinative structure's own internal program.

Structures (along with their programmed intra-cooperative abilities) are said to be marshalled by the system to execute the simple linguistic requirements. In addition, structures are nested: that is, one structure may itself, together with other structures, form some super coordinative structure. Thus:

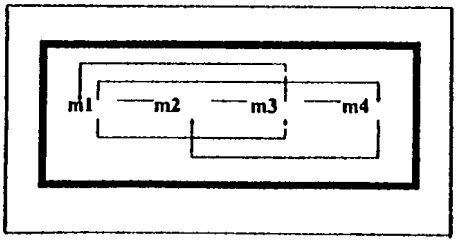


Fig. 7 A coordinative structure.

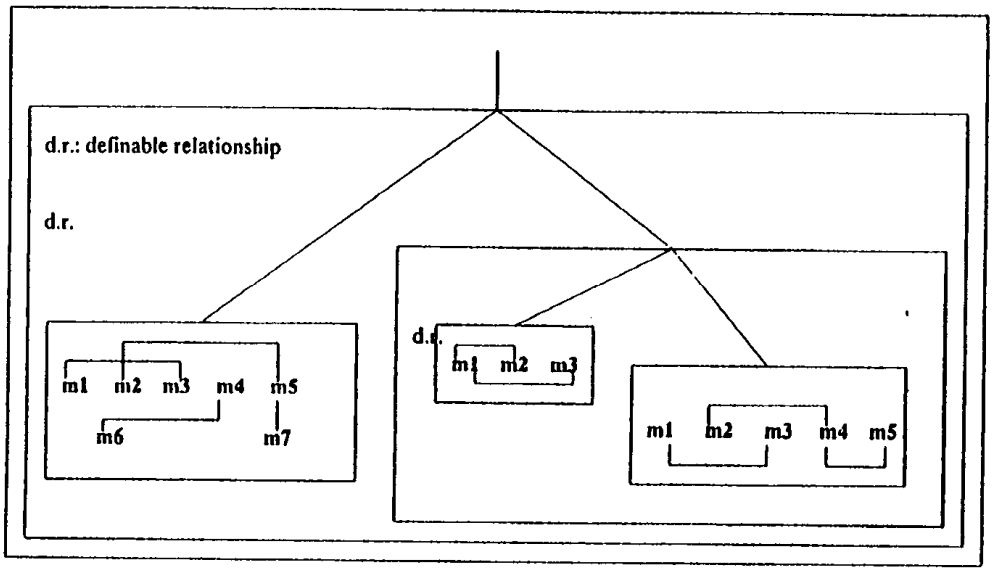


Fig. 8 Nested coordinative structures.

Tuning

The individual components of a structure (and also low level structures which form a super structure), although standing in a well defined relationship in other components of the structure, are capable of being tuned. That is, they are capable of being adjusted if necessary, on an individual basis: their internal programs can be interfered with. However, because of the way in which any one relates to all the others such tuning will results in some correlating

adjustments made automatically among the remaining components of the system. It is a property of each local program that external interference is internally compensated.

Time in the Action Theory Model

The notion of some degree of cooperation between muscle groups or between mechanically linked portions of the vocal apparatus is certainly not new, but has up to now been little more than a relatively vague idea in speech production theory. Action Theory does however add an important new idea: one of the crucial dimensions of a coordinative structure is that of time. In Action Theory much of the timing detail of an articulatory gesture which had hitherto in translation theories been assumed to be calculated (and therefore the result of cognitive activity) is treated as an actual property of the structure itself. Tuning is possible, as with other (spatial) parameters, but the basic timing of the elements within the structure is still a given property of the structure itself and the internal workings of its co-ordinative activity. The notion that time is added at such a comparatively low level in the system is new.

Usefulness of the Action Theory Model

As yet the details (in particular detailed examples of co-ordinative structures in speech) have not been forthcoming. In addition the proponents of the new theory have been somewhat rash in their claims as to the effect it has on phonology. Would it, for example, virtually *eliminate* it? It is understandable that detailed linguistic considerations have not yet been answered by Action Theory since its proponents are for the most part neuro-physiologists and psychologists rather than linguists.

Action Theory is still too new and too tentative at the moment for us to give detailed consideration here to its overall consequences for speech production theory, including phonology. At first glance, however, after recovering from the apparent overthrowing of much of what is crucial to earlier and current speech production and phonological theory the simplicity of the model is attractive, and once we are prepared to allow that detailed information can be added during a process without the need for constant re-representation of information from top to bottom then much of the difficult data can easily be explained.

Arguably Action Theory is essentially a physicalist theory of speech production in that it is attempting to take into account more of the detail of the actual mechanisms involved and show that when this is done it has serious consequences for the way the input to the system (and therefore the higher levels as a whole) is to be specified. There are attempts to partially reinterpret the physical model abstractly to accommodate some of the observations in the area of Cognitive Phonetics – the area of phonetics which is strictly not physical, but also not phonological.

3.4 The Independent Controllability of Features

Translation theories of speech production had not given much thought to the degree of interdependence of features (whether at the phonological or phonetic level). Within the boundaries of some of the more obvious constraints features have been taken as more or less independent in terms of the control of their muscular or articulatory correlates. Action Theory has however brought up this question (though there are no answers yet) rather forcibly, since it implies a great deal of relatively uncontrollable interdependence of phonetic features. There is little to say about this at the moment, but it is important to be aware of the fact that phonological and phonetic features probably cannot be simply manipulated at will without seriously affecting related features.

3.5 Cognitive Phonetics

In the 70s it became clear from careful experimental work that coarticulatory phenomena could not be entirely accounted for using the mechanical inertia model of translation theories. It had been less formally observed earlier that the degree of coarticulation seems to vary from language to language; later it was observed that the degree of coarticulation between

segments varies within a language and even within a sentence. An early and comparatively simple attempt to account for the cross-language observation suggested that coarticulation was *not* in fact a universal phenomenon, but language specific and therefore under voluntary control. This model explicitly denied the universal and automatic explanation of the phenomenon.

A more sophisticated model was developed, however, which accounted for coarticulation as a two layer process. The lower layer was said to be universal and mechanical, but the upper layer somehow overlaid cognitive or voluntary adjustments on these mechanical effects: they constrained them. This model was an important development because it preserved the earlier idea of the universality of coarticulation, but at the same time accounted satisfactorily for the observation that the overall effect (both layers together) is indeed language specific.

It was not until the early 80s, though, that the two layer model addressed the problem of explaining the mechanism by which constraints on universal coarticulation could be applied. It was at this time that the theory became known as Cognitive Phonetics, since it was also being developed to take into account other non-physical phenomena observed at the phonetic level. By this time few researchers were still denying the universality of the lower layer of coarticulation, but the theoretical question which needed to be resolved was whether the upper cognitive layer should in fact be the final part of the phonology. One or two researchers showed conclusively that the cognitive effects being characterised were in no sense phonological within the generally agreed definition of the component.

The solution to explaining how the two layers of coarticulation could interact to enable one to constrain the universal effects of the other way to unite this theory with the general principles expressed in Action Theory. Specifically the tuning mechanism proposed in Action Theory was just the device Cognitive Phonetics had been looking for. Coincidentally the coarticulatory phenomenon provided a perfect example for the Action Theorists of a use of the tuning mechanism which up to that time they had played down a little because of insufficiently convincing examples of its use.

4. ACOUSTIC PHONETICS

4.1 Introduction

Acoustic phonetics deals with the nature of the speech soundwave and how it is produced in terms of the acoustic properties of the vocal apparatus. It overlaps articulatory phonetics (in that how the organs of speech are configured has a direct bearing on the sound produced) and aerodynamics (in that the way the air flows through the vocal tract also affects the sound produced).

Speech as an acoustic event is the end product of a long chain of processes linking meaning and sound in a speaker. To that extent soundwaves are the final encoding of what a person wants to communicate, and could be described as a goal of language.

> [*footnote:* Language can branch to other goals depending on a choice of communication medium, e.g. writing.]

4.2 Speech Sound Production

Sound is nothing more than vibrating particles of air whose movements are oscillations within the range 20-20,000 times each second. This is the range within which a young person's hearing will respond. Although the air particles can and do oscillate at rates beyond this range the term sound is reserved for those which fall within the hearing range. The vibrations per second are referred to as cycles per second, or in more modern terms, Hertz (abbreviated to Hz). Thus we speak of a person's range of hearing as being from 20Hz to 20kHz (where k = kilo = 1,000). The higher the number of Hz for a given sound the higher the perceived pitch of the sound when a person hears it.

> [footnote: Do not confuse k with K. K means 1024 and is used in speaking of computers in such situations as when, for example, it is necessary to refer to a number of bytes. Thus, we might say *This file is 2Kbytes in size* – meaning that it occupies 2048 bytes of storage space.]

The oscillation of air particles is modelled using two dimensions (or parameters):

1. frequency – the *rate* at which the oscillations are occurring;
2. amplitude – the *extent* (or displacement) of each oscillation.

Frequency and amplitude are objective and measurable properties of the soundwave itself, irrespective of our subjective response to the sound. A person is said to hear the intrinsic frequency and amplitude of sound. When however what a person is hearing undergoes any mental processing we speak of perceiving pitch and loudness. Perception of pitch and loudness is a psychological, or cognitive, response to hearing frequency and amplitude. Since the acoustics of speech deals with the characteristics of sound before it reaches the listener the response terms of pitch and loudness should be avoided. We shall be looking closer at these terms when we come to consider the psychoacoustics and perception of speech (see Perception).

Speech sounds are described as a complex waveform. This simply means that they can be *thought of* as being made up of many frequencies with differing amplitudes. It is the particular arrangements of frequencies and amplitudes which give speech sounds the differing qualities which, when perceived, enable us to identify them. The aim of articulation is to produce consistently these different combinations of frequencies and amplitudes such that the corresponding qualities may be perceived as the intended speech sounds. Thus, for example, if you want (a mental process) to communicate the word *I* to another person you direct your vocal apparatus to assume the configuration which results in that sound quality which a listener can perceive as being associated with the word *I*. The speaker is said to encode the word as a particular soundwave, and the listener is said to decode this soundwave back to the

word intended to be communicated. Clearly compatibility between encoding and decoding is essential, with the complementary processes meeting at the soundwave produced by the speaker.

To produce the sounds of speech a speaker uses his vocal apparatus. The shapes or configurations this must assume to produce sounds of particular qualities, together with how to control the system, are the subject matter of Articulatory Phonetics (see Articulation).

All speech sounds can be best regarded as being produced in two stages. That is, what we hear is the result of a two stage operation:

1. stage one in the process involves generating or producing some basic sound, called the excitation source,
2. stage two involves the manipulation (or transformation) of that basic sound into the recognisable qualities of individual speech sounds.

The theory which describes speech production as a two-stage process is called the Source-Filter Theory of Speech Production.

> [*footnote:* Here *Theory of Speech Production* refers to the production of the *acoustic signal* only – it does not refer to other aspects of speech production like motor control.]

To understand what is involved in the two stages we shall examine what happens during the production of vowel sounds. The process is similar for consonantal sounds, though a little more complex.

Vowels

Excitation Source

In the production of the speech sounds corresponding to vowels the excitation source is produced in the larynx by vibration of the vocal cords. This vibration causes sound of a non-continuous pulsating nature (termed periodic or quasi-periodic sound), consisting of a train of sound bursts. The rate at which these pulses are produced by the vibrating vocal cords is called the fundamental frequency (symbolised as f0 and referred to as *f zero*), and, when perceived by a listener, is called the sound's pitch (see Perception). As an example of how fast the vocal cords vibrate to produce the pulse train a male voice averages an f0 of some 120Hz, and an f0 of some 180Hz is the average for a female voice. The difference in f0 between men and women results from a difference in the length and mass of their vocal cords.

Each burst of sound in the pulse train is itself a complex wave. It can be analysed into a fundamental frequency (f0) together with harmonics which are common multiples of the fundamental (that is, are frequencies above f0 determined by multiplying f0 by a whole number: 2, 3, 4, 5, etc.). So the spectrum of a pulse whose fundamental frequency is 100Hz (i.e. the vocal cords are vibrating 100 times each second) consists of a f0 at 100Hz and harmonics at 200Hz, 300Hz, 400Hz, 500Hz, etc. Similarly the spectra of pulses with f0s of 120Hz and 130Hz have harmonics at 240Hz, 360Hz, 480Hz, 600Hz, etc., and 260Hz, 390Hz, 520Hz, 650Hz. etc., respectively.

Each of the component frequencies of the complex wave associated with the pulsed sound from the larynx has, of course, amplitude. The fundamental frequency often (but not always) has the greatest amplitude, with successive harmonics having progressively lower and lower amplitude such that harmonics above 5 or 6kHz are usually inaudible. It is this pulsing spectrum of harmonics up to about 4kHz which is to be manipulated or transformed in the second stage of the process of producing speech sounds.

Notice that the quality of the sound produced at the vocal cords does not vary under normal circumstances. Quality (a perceived phenomenon) is determined by the way in which frequency and amplitude relate: this relationship is constant in stage one of the process, irrespective of what might happen later in stage two. Thus, as an example, when we hear

qualitative differences between the sounds corresponding to differing vowels like [i], [ɑ] or [u], these differences are almost entirely attributable to stage two of the process.

Although under normal circumstances we do not change the quality of the larynx source sound we do change its amplitude by increasing the airflow through the larynx (so that a listener may perceive greater or less *loudness*), and we do change its fundamental frequency by allowing the vocal cords to vibrate faster or slower (so that a listener may perceive higher or lower *pitch*).

> [*footnote:* Remember that physical amplitude corresponds to perceived loudness and that physical fundamental frequency corresponds to perceived pitch. Remember also that the correlation between these physical and abstract properties is not straightforward.]

Filtering

Stage two of the process of producing the final sounds corresponding to vowels consists of manipulating the source sound as it travels from the larynx, through the oral cavity and into the outside world *via* the lips. The oral cavity acts as an acoustic resonator when driven by the source sound, and resonators have the important property of being able to transform (or filter) the sound driving them.

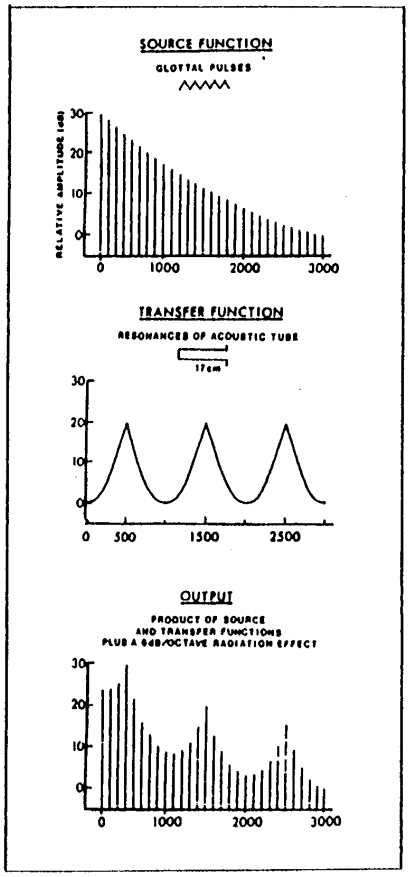


Fig. 9 Effect of the vocal tract transfer function.

This transformation takes the form of altering the amplitude relationship between the harmonics of the source sound by a process of filtering, just as, for example, the tone controls on an audio amplifier alter the quality of music sounds. In the case of the oral cavity the source spectrum is manipulated to produce three important amplitude peaks, corresponding to peaks in the response of the resonator. The relationship between these amplitude peaks in the

frequency domain characterises the recognisable qualities of the sounds used to represent different vowels.

Thus a relationship between harmonics in the source, characterised by each harmonic having a lower amplitude than the next lowest on the frequency scale, is transformed by the oral cavity resonator into a relationship showing three peaks of amplitude among the harmonics.

The three amplitude peaks are called formants, and labelled Fl, F2 and F3, where Fl has the lowest frequency. Do not confuse Fl, F2 and F3 with f0 (the fundamental frequency); note the usage of capital and small letters.

As the resonator responsible for producing this harmonic amplitude peaking by transforming the source sound from the larynx, the oral cavity can be altered in volume and shape. These changes are brought about principally by movement of the tongue to various positions within the cavity, assisted by alterations in jaw height. Thus, as we learn from articulatory phonetics, for the sound symbolised by [i] the tongue bunches relatively high toward the front of the mouth; for [ɑ] it is low toward the back of the mouth; for [u] high toward the back, and so on. Changes in the tongue's positioning alter the filtering effect of the oral cavity's resonating properties. There are still three peaks of amplitude in the resonance, but the relative positions of these peaks on the frequency scale alter. Notice that for [i] F2 is close to F3, and that for [ɑ] F2 is close to F1.

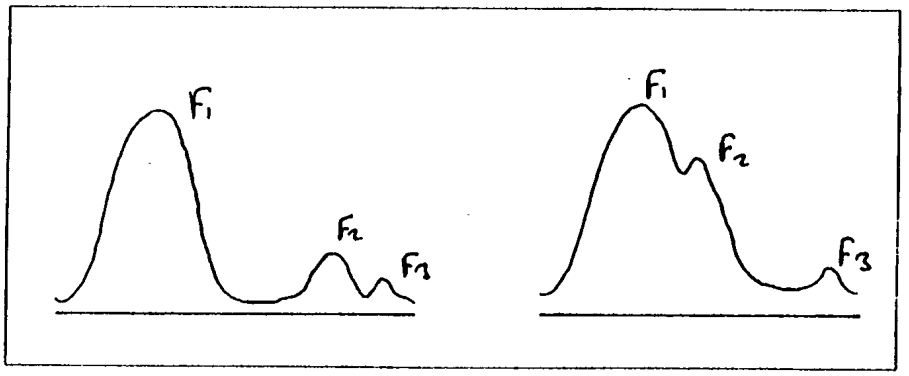


Fig. 10 Envelopes of the spectra of [i] and [ɑ].

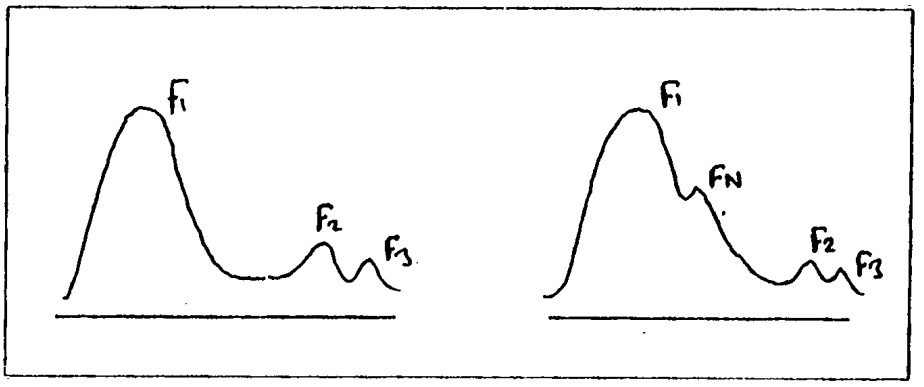


Fig. 11 Envelopes of the spectra of [i] and [ĩ] – notice the presence of the nasal formant caused by resonance of the air passing through the nasal cavity.

The qualitative differences we perceive among the sounds representing vowels are mainly due to these changes in the relative placing of F2 with respect to Fl and F3. And this spectral change is in turn due to alteration in the oral cavity's resonance properties by movement of the tongue within the cavity, and a change in jaw position.

Note, once again, that adjustment of the oral cavity's resonance properties (stage 2) is independent of the source sound (stage 1) driving the resonance. The final effect of the whole system (stages 1 and 2 together) is what determines the overall sound a listener hears.

- Stage 1 gives rise to the perception that the sound is the representation of vowelness (that is, any vowel).
- Stage 2 gives rise to the perception of a particular vowel among the range of vowels possible for the particular language.

Finally on the sounds for representing vowels, we have, in articulatory terms, the ability to cause air entering the cavity above the larynx to flow not only through the oral cavity but also through the nasal cavity by the action of allowing the velum to lower. The nasal cavity introduces another resonator into the system, though the volume and shape of this one cannot be altered because there arc no mobile articulators bounding the cavity. The distinctive quality of sound associated with nasal vowels is caused by the addition of a further formant – the nasal formant (symbolised as F_N) – usually just above Fl on the frequency scale.

Sounds representing many of the consonants (*not* [w, j, l, r], which are more vowel-like in this respect) are produced somewhat differently, though the same idea of a two stage operation involving source and resonator/filter still holds. For many consonantal sounds, though, stage 2 has relatively less importance than it has for vowels.

Besides the pulsed harmonically structured source sound we associate with vowels and vibrating vocal cords, we can make a second type of sound with our vocal apparatus. This second source involves setting up a narrow constriction somewhere in the vocal tract, and forcing air under pressure from the lungs through it – this source is used in whisper and in some consonants.

Whisper

Before passing onto consonants we should note that the lowest part of the vocal tract in which we can cause such a constriction is in the larynx itself. By holding the vocal cords slightly apart but very tense, and forcing air between them vibration is prevented, and instead a hissing sound is produced by the turbulence generated on the side of the constriction away from the airflow source. Since this hissing source travels the whole length of the vocal tract it is subject to the same resonator filtering as the more usual pulsing source we associate with phonological voicing. The result is the familiar formant structure of the final sound, but with a major difference: hissing sources, produced by forcing air through a narrow constriction, neither come as a train of pulses nor have the harmonic structure we saw earlier. Instead the sound is continuous (i.e. not pulsed) and consists of random frequencies (i.e. not harmonically structured in an orderly fashion). When the source is in the larynx we have, of course, whisper. In this case the alternative source has been used for *both* vowel and consonant representation.

Consonants

Fricatives

For the exclusive production of consonant sounds, hissing sources can be produced *within* the oral cavity itself. Thus part of the front of the tongue, for example, can be brought close to the front part of the palate right behind the upper teeth (the alveolar ridge). Air forced through the narrow gap will produce the hissing source which, having travelled (and been slightly modified by a stage 2 effect) between the teeth and lips, is recognisable as the sound perceived as [s]. Similarly, providing the source by bringing the back of the tongue close to the soft back of the palate (the velum) and providing a resonating stage 2 in the oral cavity in front of the constriction produces a sound perceived as [x] (the orthographic *ch* in Scottish *loch*, or German *Bach*), or placing the tongue to form a constriction with the middle of the palate produces the [ç] sound in Welsh *Llangollen* or Yorkshire dialect *Keighley*.

The hissing sound results from random motion of air particles due to turbulence on the

side of the constriction furthest from the air source. The *perceived pitch* of the hissing (its predominant frequency) and its *amplitude* depend on the relative excitement of the air particles: a narrower constriction results in a higher frequency and greater amplitude. For example, the sound [ʃ] is the same as the sound [s], except that the constriction giving rise to air turbulence is somewhat wider. There are important limitations on gap width: too wide and there will be too little turbulence to be audible; too narrow and there will be no gap at all – the airflow will be stopped. You can imagine that articulator placement can be quite critical in achieving the correct fricative sound on demand. It may be for this reason that fricatives are among the last sounds a child learns to produce, and among the first to fail in some disorders of speech or when the functioning of the nervous system is impaired in some way.

As in whisper it is possible to produce frication at the vocal cords by holding them tense, though apart. One consonant in English, [h], has its source at the vocal cords.

Stops

Voiceless slops

The limitation on the production of turbulence or frication that there must be a constriction in the vocal tract which is neither too wide nor too narrow is an interesting one because it is actually used in the production of some consonant sounds: the stops (or plosives), such as [p], [t] and [k] in English.

Take [t] as an example. Here the tongue is brought firmly against the alveolar ridge for a few tens of milliseconds (1ms = 0.001s) with more than enough force to hold back the airstream. At the appropriate moment the tongue is brought suddenly away from the point of contact allowing the builtup air pressure to explode (hence the alternative term to stop: plosive) through the gap created. The result is extremely brief hissing akin to [s] followed by [ʃ] as the tongue is brought further and further away from the palate. [t] is, in fact, silence followed by high amplitude but short-lived hissing or frication.

Likewise, [k] is formed by stopping the airflow by holding the back of the tongue against the velum, and [p] by holding the lips together to produce the stop.

The release or explosion of [p], [t] and [k] is performed very rapidly, and any following frication phase of the consonant sound is very brief indeed (5ms (0.005s) or less). Slower, more controlled, pulling apart of the articulators results in the more prolonged release characteristic of the *affricates*. In these sounds (e.g. [ts] or [ʧ]) greater acoustic prominence is given to the hissing source following the stop of the airflow by moving the mobile articulator away from the stop position more slowly than in the case of stops. In [ts] (as in *Tsar*) there is a slight pause in a narrow gap position, and we hear a brief [s]. Or in [ʧ] (as in *match*) the release pauses on a wider gap and we hear a brief [ʃ]. These frication or hissing sources following stopped airflow are of greater significance in the affricates than the stops.

There is one further voiceless stop, [ʔ], which is produced by halting the airflow in the glottis. The vocal cords are brought together under tension sufficiently strong to stop airflow. In some dialects of English this sound – the glottal stop – is used as an extrinsic allophone, replacing any of [p, t, k]. In some dialects (e.g. in the Tyne-Tees region of England) it is used to reinforce the usual articulation of the three voiceless stops, occurring simultaneously with them. In some languages (e.g. Danish) the glottal stop functions phonologically as a phoneme or underlying segment to distinguish morphemes: it is *not* used this way in English.* Even in those dialects of English where the glottal stop does not occur in speech, it is frequently used in singing at the start of a word beginning with a stressed vowel: this has the effect of producing an abrupt onset to the vowel and gives better or clearer synchronisation with the

beat of the music.

*[*footnote:* In the English dialects where the glottal stop is found it functions as an allophone of one or more stop consonants – it is not a phoneme in its own right.]

Voiced consonants

So far we have looked at two possible sources for stage 1 of the production of speech sounds: pulsed, harmonically structured sound and continuous, randomly structured sound. But there is yet another possibility: the combination of both sources. Many consonant sounds show a contrast depending on whether the single hissing source is used or whether it is combined with the larynx pulsating source. Thus a single-sourced [s] is contrasted with the double-sourced [z], both being identical in articulator placement, but with the addition of vocal cord vibration for the latter. Similarly the pair [f] and [v], etc. Even stops may employ both sources. Hence the pairs [p] and [b], [t] and [d], [k] and [g]. Affricates also: [ʧ] and [ʤ] (as in *judge*). Phonologically these double sourced consonants are classified as [+consonantal, –vocalic, +voice].

Notice that it is not possible for the sounds [ʔ] and [h], produced at the vocal cords themselves, to have double source counterparts. To produce [ʔ] the vocal cords themselves form the stop – so cannot be vibrating. To produce [h] the vocal cords have to be held tense to create the narrow gap essential for generating turbulence – so, once again, cannot be vibrating.

4.3 Summary of the Acoustic Theory of Speech Production

Speech sounds should be regarded as being composed of a source sound and a resonating modification of that source.

Two types of source are available:

1. pulsed, harmonically structured sound produced by the vibrating vocal cords (the periodic source),
2. continuous, randomly structured sound produced (except in a whisper) by narrow constriction elsewhere in the vocal tract. Width of constriction determines frequency and amplitude of the hiss (the a-periodic source).

Vowel sounds use the pulsed source, most consonants use either hissing or hissing and pulsed sources together, and [l, r, w, j and the nasal consonants] use the pulsed source.

Resonance filtering or transformation of the pulsed source provides the differentiating qualities associated with vowel sounds. Three amplitude peaks are imposed on the harmonic structure of the source, and the *relative* positions of the three peaks or formants on the frequency scale are responsible for the perceived quality.

Aside from whisper (where full resonance is available because the hissing source is at the far end of the vocal tract) resonance in consonants, though present, is not as complete as with vowel sounds – that is, less than the entire oral cavity resonator is brought into play. It does however provide one feature enabling us to distinguish, for example, between [t] and [k].

In addition, for consonants, the hissing sound may be long (as in the fricatives [s] or [f], for example) or, following a total silence, very brief (as in the stops/plosives [t] or [k]), or more prolonged (as in the affricates [ʦ] or [ʧ]).

For consonants both sources may be used simultaneously in fricatives (such as [z]), or stops (such as [d]), or affricates (such as [ʤ]).

4.4 Spectrograms

Because sound is a transitory phenomenon and only able to be perceived subjectively by listeners, experimental phoneticians have several methods of making the sound permanent by displaying it in visual form. This makes objective measurement of phenomena possible, and removes the possibility of subjective error while making auditory judgements.

The commonest visual display we have is called the sound spectrogram. The display takes the form of a graph on a computer screen; for a more permanent record this graph can printed onto paper. It is not necessary to understand the details of how the computer program transforms sound into a visual presentation to be able to understand what the graph is showing. Basically all the computer does is

- firstly divide the acoustic signal temporally into slices of quite brief duration (less than 10ms),
- secondly examine the spectrum of each slice, and
- lastly display these individual spectra in successive columns on the screen.

The amplitudes of frequencies within the spectra are presented either with different colours or with varying shades of grey to indicate different levels of amplitude.

Thus a complete picture is built up of the spectral content of the original soundwave. On the graph

- the x-axis (horizontal) represents time, running from left to right;
- the y-axis (vertical) represents the frequencies of the spectral components the computer identifies;
- the different colours or levels of grey scale represent the changing amplitudes of frequency components in the signal (this is called the z-axis).

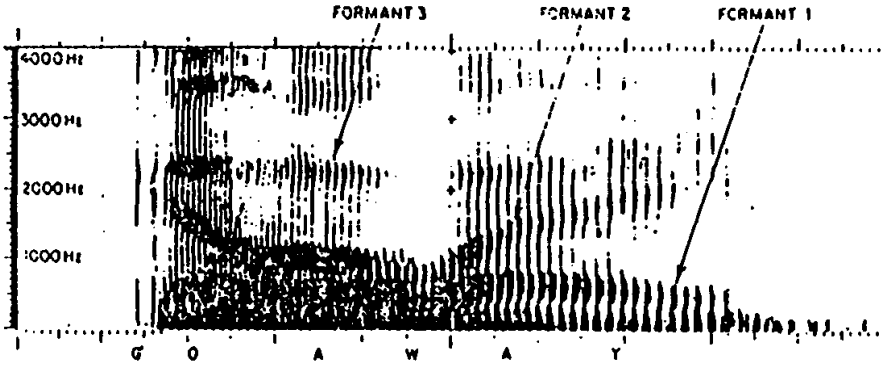


Fig. 12 Spectrogram of the utterance *Go away!*

In vowel sounds the formants are areas within the frequency domain of relatively high energy or amplitude. In the accompanying spectrogram the formants can be seen clearly as prominent bands running horizontally during the utterance. Notice that they do not hold steady in time (left to right) at any particular frequency: their frequency varies over time. The utterance in the illustration is the phrase *Go away!* By checking against the frequency scale to the left of the graph the formant frequencies in the mid-central first vowel of *away* can be seen to pass through 560Hz, 1500Hz and 2500Hz.

The vertical markings on the graph show single pulses of sound coming from the glottis as the vocal cords open and close. Notice that the spacing between these vertical marks changes during the utterance: spacing of vocal cord vibrations gives us the variation in fundamental frequency which is detectable by the listener as changes in pitch.

In the diphthong at the end of the utterance you can see how the second formant moves in

time from a relatively low frequency through to a relatively high frequency. This reflects the movement of the tongue, and hence changes in the resonances of the oral cavity that take place during a typical diphthong.

Sound can be described using the parameters (or features) of amplitude, frequency and time – three dimensions which, when fully specified, entirely and uniquely characterise a sound. Sound spectrograms provide a display graph indicating all three dimensions. Such a display provides all the information about a sound. Occasionally, however, all is too much, and researchers use other programs to provide display graphs of, say, just frequency or amplitude.

There is, in fact, a fourth dimension to sound – phase, which characterises how much the various simple sounds which make up a complex sound like speech are in step with each other. Phase is usually taken as being irrelevant in speech. In perception studies in general, however it *is* important. Detection of phase is one of the ways in which we are able to decide the direction from which a sound is coming toward us. Because of the difference in arrival time of a sound to our two separate ears we are able to compute the direction of the sound from the degree to which the two different signals (one from each ear) are out of phase (i.e. out of step with each other). It is, among other things, artificial manipulation of this parameter which makes stereophonic or binaural sound reproduction possible.

5. HEARING

5.1 Introduction

It is important in the study of auditory phonetics to distinguish between hearing and perceiving. Hearing is an automatic process. By this we mean that no thought processes are involved; that is, there is no cognitive intervention in the hearing process itself. Hearing is thought of as a function of the design of the hearing apparatus. By contrast, perceiving is generally thought of as an active process (but see Direct Perception), involving cognitive intervention.

5.2 Some General Facts about Hearing

- The range of hearing of a 20yr old person is around 20Hz to 20kHz. The upper limit declines as the person gets older. It is quite common in someone aged around 60yrs for the upper limit to have declined to about 12kHz, though there is enormous variability between people.
- The frequencies of components in the complex speech waveform rarely exceed 8kHz-10kHz. Such high frequencies occur only in the fricatives – specifically the voiceless alveolar or dental fricative [s]. The frequency components of vowel sounds which are important for distinguishing between them lie between 500Hz and 3kHz (the second formant).
- The lowest frequency usually encountered in speech is around 80Hz – the lower limit of a deep male voice. The average fundamental frequency (the rate of vibration of the vocal cords: f0) of a man's voice is around 120Hz.
- The sensitivity of the ear is not constant over the entire range of frequencies. Optimum sensitivity lies between 1kHz and 3kHz, meaning that to be clearly heard sounds within this range need less intensity than those outside this range.
- If we take as a reference point for sensitivity a tone at 1kHz which has an intensity such that it is only just audible, then a tone at 100Hz would need an intensity 100 times greater to be just audible. At 18kHz the intensity would need to be 1,000 times greater to be just audible.
- The overall range of intensity which can be detected by the ear is very large. If we define the lower limit as the threshold at which a person can just hear a sound, and the upper limit at the threshold at which pain in the ear is just felt, then the overall ratio at 3kHz (that is, in the most sensitive frequency band) is around 1 to l,000,000,000,000 (one to one trillion).

5.3 The Ear

Hearing is a fairly well understood passive process involving the transducing of air pressure changes (sound) into neural signals propagated along the auditory nerve to the brain's auditory cortex. The processes involved are quite complex, but can be fairly simply described. There are still one or two puzzles concerning exactly how the cochlea works – but this need not bother us here.

The ear is anatomically divided into three parts: the outer, middle and inner ears. Each of these parts performs a different function in a cascaded sequence of processes. The object is clearly to convert the original speech waveform in the air into a form suitable for processing by the brain.

The overall system is sensitive to *changes* in air pressure, but not sensitive to steady state pressures. This is true of all the systems a human being has for sensing information from the outside world. The sensors we have available (like the eyes or the touch sensors in the finger tips, for example) detect change rather than lack of it. In the case of the ear, the air pressure must be oscillating at least twenty times per second (20Hz) before anything can be detected at

all. This defines the lower limit of human hearing: any signal below this frequency is not heard or can be heard only with greatly increased amplitude. Above the 20Hz threshold the ear is sensitive to pressure oscillations ranging as far as 20kHz, or twenty thousand vibrations per second. Notice though that the ear's sensitivity within this 20Hz-20kHz range is not even. The ear responds better to certain areas with the range than to others. It is possible to graph this variation in sensitivity to produce what is known as an audiogram.

Air pressure can oscillate of course at rates lower than 20Hz and at rates greater than 20kHz, but because these are the average lower and upper frequency limits of human hearing only pressure variations within this range are referred to as sounds. Notice, though, that speech does not occupy all of the available frequency range of sound: speech sounds fall within the range 80Hz to around 10kHz. One reason for this may be the fact that as we grow older our ears' sensitivity to high frequency sounds declines. It is unusual for people over 60 years old to be able to hear sounds above about 12kHz with the same efficiency as a 20-yr old.. Because speech keeps within the hearing range of most people this gradual fall off in the ear's response as a person gets older is unimportant in speech communication. We can say that the acoustic system of speech is designed down to the available hearing range of the significant majority of the population (there are bound to be exceptions). There would be no point in having an acoustic system for encoding language which had sounds within it which could only be detected if the listener were less than 25yrs old!

The speech waveform arriving at the listener consists of acoustic energy which is characterised by having relatively high amplitude, but relatively low pressure. This acoustic energy is transduced in the middle ear into mechanical energy; then into hydraulic energy in the inner ear. When the change to hydraulic energy in the inner ear is complete we have relatively *low* amplitude, but *high* pressure. It is here in the inner ear that the signal is transduced into electrochemical energy, and then sent *via* the auditory nerve to the auditory cortex in the brain.

The Outer Ear

The visible part of the ear, the auricle or pinna, captures acoustic energy (hence its shape) and directs it into the ear canal. Its shape is clearly complex and is such that it is directionally sensitive to the higher frequencies present in the acoustic signal. By contrast the auricle is much less sensitive to the directionality of low frequencies. Together with the relative timing of a signal at both ears this directional sensitivity enables us to locate spatially the source of the sound we hear.

The ear canal which opened out to the air at the pinna, is closed off at its other end by the eardrum (or tympanic membrane). Soundwaves entering the open end *via* the pinna travel along toward the tympanic membrane. As a tube containing air the ear canal is itself a resonator. As with all resonators the ear canal amplifies the acoustic energy travelling along it. The amplification peaks around 3–4Hz; as a consequence hearing is most sensitive in this frequency range.

The tympanic membrane divides the ear canal from the middle car. The membrane is roughly round in shape, and is also thin and very elastic. This means that it vibrates along with the air pressure oscillations reaching it along the ear canal. This is where the sound wave's acoustic is transduced into mechanical energy: oscillating air particles cause mechanical oscillation of the tympanic membrane – we have an acoustic to mechanical conversion system. Later we shall find, in the inner ear, a mechanical to electro-chemical conversion system.

The Middle Ear

The middle ear which runs from the tympanic membrane and to the oval window is filled with air. There is a tube (called the eustachian tube) which runs from the middle ear to the pharynx and whose function is to equalise any air pressure variation in the middle ear.

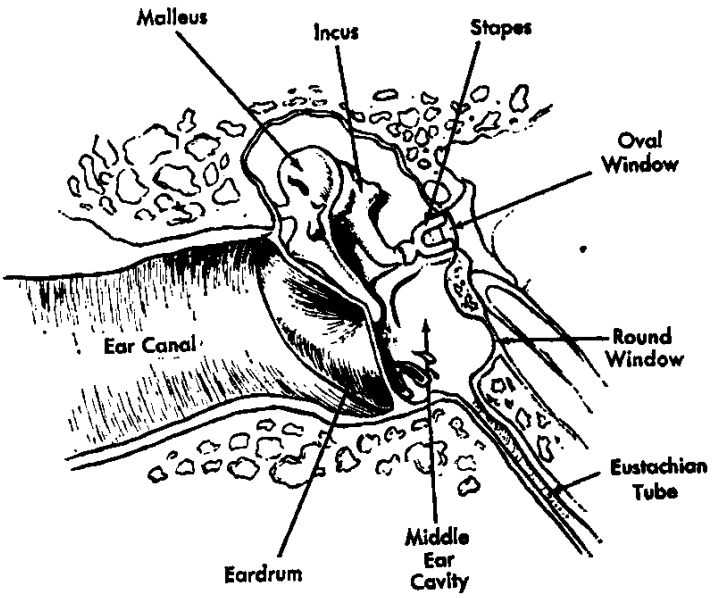
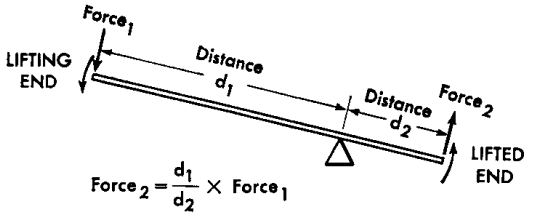


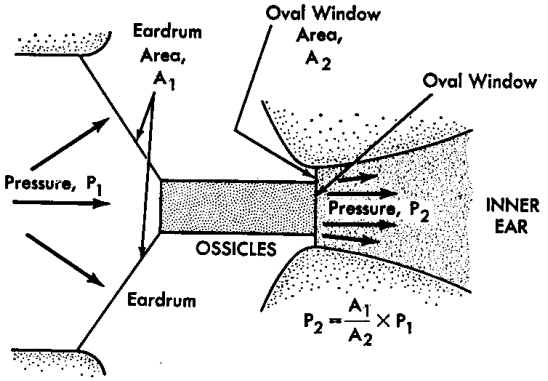
Fig 13 The middle ear

The tympanic membrane is connected by a tiny bone, called the malleus to a muscle, the tensor tympani, which when contracted tightening the membrane by leverage. This stiffens the eardrum, altering its sensitivity to external air pressure variations. The mechanism serves to protect the eardrum in the event of very loud sounds. The malleus is the first bone in a lever system involving two others; the three are known together as the ossicles: incus and the stapes. As we have seen the malleus is attached to the eardrum, and this enables vibrations to be conducted down the lever system. The malleus undergoes a pumping action, whereas the incus has a rotating oscillating movement, followed by a pumping action again, this time by the stapes. The stapes itself is attached to the oval window which is a membrane separating the middle and inner ears. As the stapes pumps the oval window vibrates in sympathy. The stapes is also capable of rotating action. Pumping takes place if the sound intensity is fairly low, whereas the rotating action seems to occur with higher intensity sounds. For more detail here see: Raphael, Borden and Harris.

Fig. 14 The middle ear leverage system.



(a)



The bony set of levers in the middle ear serves as a kind of mechanical amplifier as mechanical vibrations pass from the eardrum to the oval window. The window is about 5% the size of the eardrum – resulting, with the lever system, in a considerable pressure increase at the window at the end of the level system. It is said that the pressure increase at this point is some 30dB..

The rotating action which the stapes undergoes when there are high intensity sounds reduces vibration of the oval window and seems to be another protective mechanism. The stapedial muscle acts like the tensor tympani, drawing the stapes away from the oval window. These two protection mechanisms together form an automatic gain control: as the *amplitude* of sound reaching the ear increases they are progressively brought into play to dampen the increase thus preventing the occurrence of damage. This gain control mechanism is automatic – it is not something we consciously bring into play. The system is not entirely foolproof, however, because like any feedback or feedforward system, whether mechanical (as this is) or electronic, there is always a delay before the control system completely responds. Under certain circumstances the system can be beaten – as when, for example, the ear encounters a loud transient sound, or one with a very sharp onset (that is, one with a very fast amplitude rise time). Sounds like gunshots can potentially beat the system and could result in damage.

As we shall see, the inner ear is filled with fluid. This is why the amplification of the signal, provided by the leverage system of the middle ear, is necessary. To set up vibration in the inner ear fluid requires more energy that contained within the original soundwave. Without the mechanical amplification vibrations would not be set up in the inner ear.

The Inner Ear

The inner ear comprises the cochlea – a fluid filled coiled tube about 35mm long. At one end of the tube we find the oval window. It is the vibrating movements of the oval window which transduce the mechanical vibrations from the middle ear into hydraulic vibrations.

The cochlea is in two longitudinal sections separated by the cochlear membrane. The ‘chamber’ terminated by the oval window is called the scala vestibuli, and the chamber terminated by the round window is called the scala tympani. Each chamber is filled with a viscous fluid called the perilymph. The perilymph can flow between the chambers since the cochlear partition has an opening at the apical end called the helicotrema.

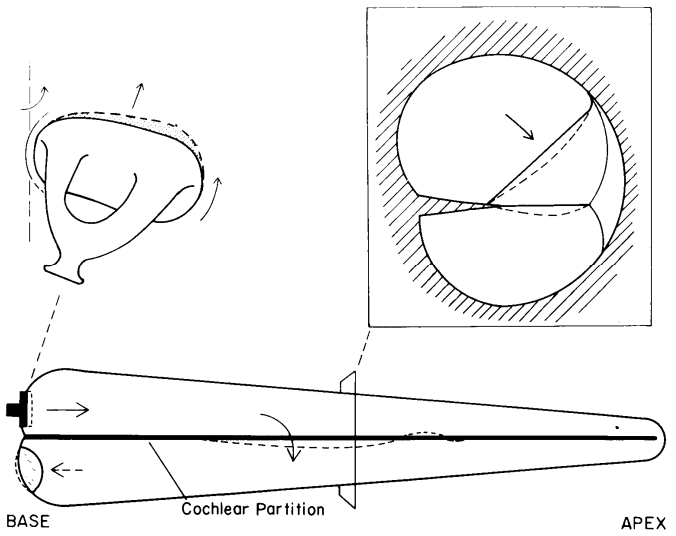


Fig. 15 The cochlea.

Attached along the length of the basilar membrane is the organ of Corti. And Attached to this are some 5,000 ‘inner’ and some 25,000 ‘outer’ hair cells, their upper ends being attached to the tectorial membrane. The hair cells are able to resonate to vibration in the basilar membrane, exciting nerve cells at their ends. The outer hair cells detect any *transverse* bending motion across the basilar membrane, and the inner hair cells react to vibration travelling *along* the basilar membrane.

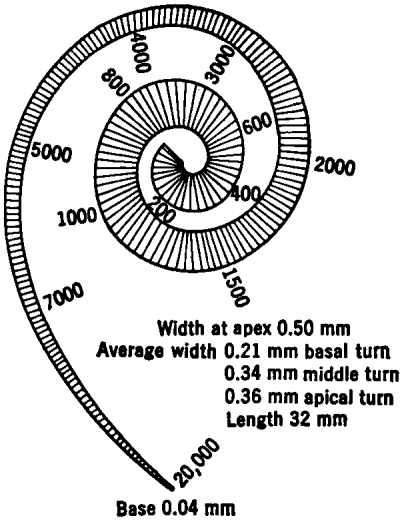


Fig. 16 Cochlea response.

If vibration below 20Hz arrives at the inner ear the perilymph flows from the scala vestibuli to the scala tympani *via* the helicotrema. The pressure created in the scala tympani is lessened by the expansion of the round window membrane; this effectively reduces our ability to hear very low frequency sounds.

At frequencies greater than 20Hz approximately, the pressure difference across the cochlear partition causes a corresponding vibration in the basilar membrane, and as the basilar membrane gets narrower towards the apex its resonance properties change along its length. As the frequency of vibration rises the area of the basilar membrane resonating is closer to the base of the cochlea, toward the oval window where it is narrowest and stiffest. But as the vibration gets lower in frequency the resonant area of the basilar membrane is nearer the helicotrema where it is wider and more elastic.

Oscillation of the basilar membrane causes the hairs of the sensory cells to 'wave', exciting the nerve endings attached. This nerve endings excitation results an electrochemical signal which is sent *via* the auditory nerve (a bundle of some 30,000 nerve fibres) to the brain's auditory cortex.

5.4 Complex Wave Analysis

The mechanism of the inner ear is in effect a mechanical spectrum analyser – that is, it is able to separate out the component frequencies in a complex wave and ensure that they are delivered along separate nerve fibres in the auditory nerve for projection onto the auditory cortex. Besides time, there are two parameters which characterise sound: frequency and amplitude. Within the cochlea frequency and amplitude remain as they are, except that they occur within a fluid rather than a gas (the air which was the original medium for the sound). After the processes which take place in the cochlea, however, these two parameters have been transformed or re-encoded. In the auditory nerve frequency is now encoded by the identity of the neural channel which is excited, corresponding to which hair cell group was displaced, and amplitude by the rate or frequency of impulses propagated along the excited channel, corresponding to the amount by which the hair cells were displaced.

Notice also that whereas sound pressure waves are characterised by *smooth* variation in time (i.e. are analog waveforms), the signals in the auditory nerve have a *spiky*, all-or-none nature (i.e. are digital waveforms). Within the mechanism of the cochlea an analog-to-digital converter has made the transformation from analog to digital signals.

Thus

- the analog waveform is digitised within the cochlea – that is, the cochlea acts as an analog-to-digital converter (ADC);
- the frequency components of the waveform are encoded by channel identity in the auditory nerve;
- the amplitude of each frequency of the analysed waveform is encoded by the frequency of the neural impulses in the auditory nerve.

5.5 The Role of the Brain in Hearing

The brain has about 100 billion nerve cells which are connected together as a network. The network is used to enable cells to communicate with each other using an electrochemical messaging system. Each cell can make simple computations, although these are neither very accurate nor fast. There is a kind of paradox here because the brain is generally considered to have enormous computational power and speed – and this is indeed true of the brain *as a whole*. The overall performance of the brain is down to the fact that the particular network arrangement of the nerve cells permits simultaneous or parallel operation. Man-made computers rarely have this property of multiple processors working in parallel; generally our computers consist of one or two processors performing operations serially, very fast and very accurately – the opposite in fact to the way the brain works. It is possible, however, to draw a better analogy between the brain and the parallel processing computers which began to emerge in significant numbers in the early 80s. These computers contain several or many relatively low powered processors working in parallel on portions of an overall computational task.

Nerve cells within the brain each have a body containing the so-called nucleus. A nerve

fibre called an axon extends out from the cell and terminates in a fine network of filaments called the terminal branches. Nerves within the body are no more than bundles of these axons, with some extending from the brain throughout the length of the entire body. The axons and their branches which extend beyond the brain itself are called the peripheral nervous system.

There are also further fine extensions from a nerve cell's body which are called dendrites. The endings of axon terminal branches may come into contact with dendrites or with the cell body itself of a second neuron. Contact is made by so-called synapses. By means of axon to dendrite connection at the synapses messages are passed from one neuron to another.

Messages between nerve cells take the form of pulses of electrochemical activity which propagate or spread through the network. Cells are said to fire or operate on receipt of a pulse or pulses which exceed a certain threshold of activity. Pulses have a duration of only about one millisecond, or one one-thousandth of a second. All message transmission within the network is done this way: the pulsing format of the messages and the fact that cells either fire completely or not at all is why the system is termed a digital system (as opposed to analog system).

The synapse transmits a message in one of two ways:

1. excitation: the synapse can pass the pulse on by making the next neurone fire. The firing neurone sends pulses to futher neurones *via* the network.
2. inhibition: the arrival of a pulse *prevents* the firing of the connected neurone.

Using these two possibilities – excitation and inhibition – synapses direct messages throughout the network. Like logical gates they can open or close systematically, passing on, stopping or sometimes redirecting messages throughout the system. It is normal for messages to arrive together at a single cell – cells react not just to receiving a single messages but also to receiving multiple messages. A system of thresholds on the synaptic junctions determines whether the cell receiving messages (the receptor cell) will fire or not. Cell *C*, for example, might fire only if signals over a certain threshold reach it simultaneously from cells *X, Y* and *Z*: if one of these is not firing, or not firing sufficiently, then *C* itself will not respond, that is, it will not itself fire. It takes just the right combination of messages, timed correctly and of the correct strength to cause a cell to fire.

Axon width or thickness is what determines the speed at which a message will travel in the network. Axons vary in thickness, so that messages are travelling around the network at different speeds. Sometimes this phenomenon enables messages to *catch up* or be *delayed* as they pass between the various nodes or cells. This is a useful method of arranging the temporal ordering of messages.

Although it is possible for a single nerve fibre to extend all the way between the extreme periphery of the system and the centre it is more often than not the case that the pathway consists of a series or chain of nerve fibres. In cases like this the fibres in the chain are linked by synapses (with their attendant thresholding properties).

In conveying signals from the ear's organ of Corti to the auditory cortex, synaptically linked nerve fibres are connected to *both* hemispheres of the brain. The pathways followed are called ipsilateral (left ear links to left brain, right ear to right brain) and contralateral (left ear links to right brain and right ear to left brain). Researchers agree that it is the *contralateral* pathways which dominate the system, meaning that the *contralateral* connections conveying auditory signals to opposite sides of the brain from the source ears are the more important.

A Note on Spectral Channel Encoding

Because of the fact that the auditory nerve is a bundle of a large number of nerve fibres, each of which can be thought of as conveying *separately* information from individual receptors attached to the hair cells of the basilar membrane, the analysed acoustic parameter of frequency is said to be channel encoded.

[*footnote:* In fact the spatial representation of frequency found in the inner ear is reproduced at

the auditory cortex so well that the analysis is literally projected as a kind of three dimensional display in the cortex. This display is very similar to a real time spectrogram of the kind phoneticians use in their laboratory analyses of speech waveforms.]

So in summary the spectral analysis of the speech waveform is accomplished as

- a frequency analysis distributed spatially the length of the basilar membrane as resonant properties of the membrane vary along its length,
- individual groupings of hair cells along the basilar membrane respond to their tuned frequencies,
- particular nerve fibres are associated with particular groups of hair cells and therefore particular frequencies, accomplishing channel encoding of the spectrum.

6. PERCEPTION

6.1 Introduction

Perception is a complex active process involving cognitive processing of the data or signal received at the auditory cortex. We shall look at several theories of speech perception, devoting most attention to the Associative Store Theory. Theories divide basically into two categories: Active Theories and Passive Theories. Active Theories stress the idea of cognitive intervention in the perceptual process, whereas Passive Theories minimise the role of cognitive intervention.

The task of speech perception is for the listener to accept as input to the perceptual system the speech waveform coming from another speaker, and to decode that waveform somehow or other into a sequence of phonological labels which identify the sequence of phonological elements used by the speaker to create the waveform. The task is one of labelling the acoustic signal with appropriate phonological symbols.

In the part of this book concerned with Speech Production we have seen that prior to Action Theory the commonest way of modelling speech production was to assume that phonologically intended speech segments become blurred together as part of the production process. This blurring was called coarticulation. Theories of Speech Perception are concerned with how the listener in some sense reverses the coarticulatory process to recover the intended sequence of phonological elements.

6.2 Active Theories

The Motor Theory of Speech Perception

The Motor Theory of Speech Perception was developed by Alvin Liberman and other researchers at the Haskins Laboratories. It is based on the claim that speech perception is an active process involving cognition and direct reference to the listener's speech production processes.

The Motor Theory tackles the problem of unravelling coarticulation by proposing that the listener has knowledge of the way to produce isolated speech segments, and of the rules governing how they become coarticulated in normal running speech. Listeners have this knowledge because they do it themselves, though they are not, of course, consciously aware of what they are doing. The knowledge of the motor and coarticulation properties of speech called upon as part of an active process of decoding the waveform into appropriate phonological labels.

The reasoning goes like this:

1. I hear an acoustic signal,
2. the signal is continuous and blurred with no phonetic segments obviously demarcated,
3. by reference to my own speech production I know that if I had produced that signal it would have been because of the coarticulatory blurring of such-and-such a string of segments,
4. so the speaker I hear also themself intended such-and-such a string.

This active processing involves having knowledge of the phonology of the language – its elements and rules. It also involves a kind of non-linguistic knowledge – how the motor control, mechanics and aerodynamics of the vocal tract work. This latter is often referred to as vocal tract dynamics (especially in modern speech production theories like Task Dynamics – a 90s refinement of Action Theory).

The Analysis by Synthesis Theory

The Analysis by Synthesis Theory is, to put it simply, the acoustic equivalent of the Motor

Theory. It was devised by Kenneth Stevens and Morris Halle at MIT in an attempt to reconcile the apparent mismatch between the speech waveform the listener hears and the phonological labels which have to be applied to it. The difference between the two theories is that the Analysis by Synthesis Theory is concerned with bring *acoustic knowledge* to the decoding process, where as the Motor Theory is concerned with using *articulatory knowledge*.

The decoding process runs like this:

1. I hear an acoustic signal which I know to be speech;
2. I consult an auditory model of my own acoustic production of a signal like this;
3. I use phonological knowledge to deduce the speaker's intended phonological utterance.

6.3 Passive Theories

General

Central to most passive theories of speech perception is the idea that the incoming signal is processed through fixed passive filters. That is, the acoustic signal is filtered in a way which does *not* require active cognitive processing. Proponents of passive theories do not generally deny that might on some occasions be active decoding of the waveform – but they reserve this for extending the basic capabilities of a prior passive system. What they are saying is that active cognitively determined processing occurs as an extension of passive filtering if the incoming signal is very blurred or degraded in some way; normally cognitive processing is not involved.

As one (along with Roman Jakobson and Morris Halle) of the original proposers of the Distinctive Feature Theory, Gunnar Fant (formerly of the Royal Institute of Technology, KTH, Stockholm) suggests that the normal way a listener perceives is to apply passive filtering to the signal based on the idea of distinctive features and their acoustic correlates. Remember that distinctive features are elements within abstract phonological theory, but acoustic phoneticians like Fant are keen to stress that very often they can be readily correlated with identifiable properties of the acoustic signal.

Fant's proposals are interesting because he makes the claim that basically production and perception are one and the same thing – they are simply alternative modalities providing for encoding soundwaves using the vocal tract at the output device and decoding soundwaves using the ear. In the brain (some would say *mind*) speech production and perception become one and the same thing.

Although distinctive features are used in classical phonological theory to characterise abstract phonological segments many researchers, including Fant, suggest that most of the time there are reliable acoustic correlates for the features. This idea attempts to bridge the gap between the abstract representation of speech and its physical representation. In technical language, a speaker is said to map the phonological features *onto* the correlating acoustic features in the speech waveform, whereas the listener maps the acoustic features they hear back onto underlying abstract phonological features. Critical to this idea, of course, is the nature of the two mapping processes, and whether they are mirror images of each other. Proponents of the theory refer to speakers' and listeners' sensitivity to the correlations. Sensitivity means simply that these correlations are readily accessed associations in the speaker / listener's mind.

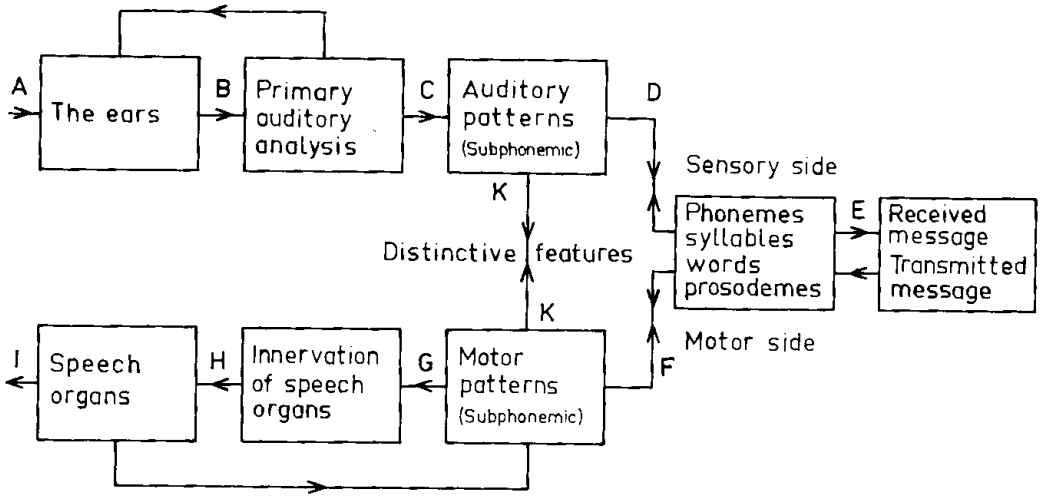


Fig. 17 Fant's model of speech perception / production.

We can borrow a couple of terms from research in the field of automatic speech recognition (the simulation of the human perceptual process by computer). Researchers like Fant refer to detection and matching – detection of the appropriate acoustic features within the waveform, and the matching of these with the correlating phonological features. In most cases, they claim, detection and matching can be carried out passively.

Direct Perception

An extreme form of the passive approach is seen in the theory of Direct Perception (Gibson and others) of speech. Though not new (it was first discussed widely in the 50s) this theory may be a viable alternative to active theories. Presentation of the theory has been somewhat informal and confused, generating much misunderstanding and argument, and awaits more formal, careful presentation before we can take it as a serious challenge to active theories. The theory of Direct Perception parallels Action Theory in speech production, which is also characterised by minimising the role of cognitive processing in *producing* the soundwave.

The essential characteristic of Direct Perception Theory is, as the name suggests, the claim that perception is direct, involving little or no cognitive processing. The spectral and temporal analysis of the soundwave by the ear itself, and the characteristics of the spectral array representation of sound in the auditory cortex (see Hearing) constitute a sufficient passive analysis to enable awareness of the stimulus and its meaning. Cognitive interpretation is considered unnecessary. It is important to stress here the idea of awareness of a signal's meaning – and the idea that meaning is embodied in the acoustic signal.

Understanding such a claim involves adopting a philosophical stance which is a little strange or unfamiliar to us. We might feel that it is true that soundwaves (or light rays, for Direct Perception is about vision as well) have certain properties: the properties that can be analysed out in a laboratory. For us those properties do *not* include or characterise *meaning*. For those with the more usual way of looking at the world, a spectrogram, for example, is a picture of the sound or utterance but *not* of its meaning: meaning is something minds interpret from physical representations. Interpretation, of course, involves cognitive processing.

For Direct Perceptionists, however, meaning is in some sense a property of the soundwaves or light rays themselves. The perception is direct without cognitive intervention or processing. In fact Direct Perceptionists, like Action Theorists, are not as extreme as that: they do allow for some cognitive processing, though this does not (as it does for translation theories) form the main perceptual mechanism or strategy.

Many questions are being asked and sometimes answered as the cognitive and direct theories of perception are discussed. You may like to try to think of some crucial questions

the answer to which would reveal the viability of any theory of Direct Perception.

> [*footnote:* Action Theorists, working in speech production (see Articulatory Control), have explicitly identified themselves with the ideas of the Theory of Direct Perception. Both minimise cognitive intervention in the processes of production and perception. Both give higher status to the passive properties of physical mechanisms than the translationists or proponents of the active theories of perception. For the Action Theorist the *intrinsic* properties of a coordinative structure render it unnecessary to posit very much cognitive activity in phonology or phonetics. For the Direct Perceptionists the analysing properties of the car and the format of signal presentation to the brain in the auditory cortex (both passive processes essentially) make it *unnecessary* to invoke cognitive intervention.]

6.4 The Problem of Speech Perception

The main problem in arriving at a satisfactory theory of speech perception is accounting for the fact that speech soundwaves are not a one-to-one encoding of phonological segments. The latter are abstract cognitive concepts anyway, and we might expect their physical coding to be complex.

Some researchers suggest that in some way when the soundwave is produced the various features of the segments are spread around, merging segment with segment at the physical acoustic level. If this idea is accepted then in principle perception (or decoding) has simply to recover the features from the soundwave and re-assemble them to identify the original phonological segments. The passive theories of perception each try to devise an automatic filtering type of procedure to achieve this. Unfortunately hard work over three decades, and recently with elaborate computing facilities available, has failed to come up with any suitable decoding procedure. In other words the hypothesis that the acoustic waveform is an encoding, however complex, of phonological segments and that those segments are consequently in principle recoverable by passive decoding, has so far defeated empirical verification. This does not mean to say that it will not eventually be possible to find segments in soundwaves, but for the moment alternative models are more viable.

The alternative hypothesis is that the soundwave is *not* a direct encoding of phonological segments. The segments are *not* in the acoustic signal and are *not* therefore recoverable from it – how could they be if they are not there to begin with? The segments are abstract and the signal is physical. Phonological segments are simply an abstract descriptive device devised by phonologists. A quite different perceptual strategy is therefore needed. In active theories of speech perception the incoming soundwave is used to enable the recovery of appropriate phonological segments not from the acoustic signal, but from within the listener's mind.

6.5 The Associative Store Theory of Speech Perception

This is the extreme form of an active theory of speech perception. It is mentalistic in as much as no concrete mechanisms are described in the theory, and as such is in the spirit of Chomskyan linguistics. It would seem to be an attractive model for phonologists to espouse.

If the model is viewed as a series of processes responsible for decoding meaning from the speech soundwave (i.e. it is a translation theory), the start of the overall process is some abstract mental representation of the incoming soundwave. This representation may be de-graded in the sense that it reflects errors which may have crept into the speech sounds by reason of poor production, poor transmission between the speaker and listener or poor hearing on the part of the listener.

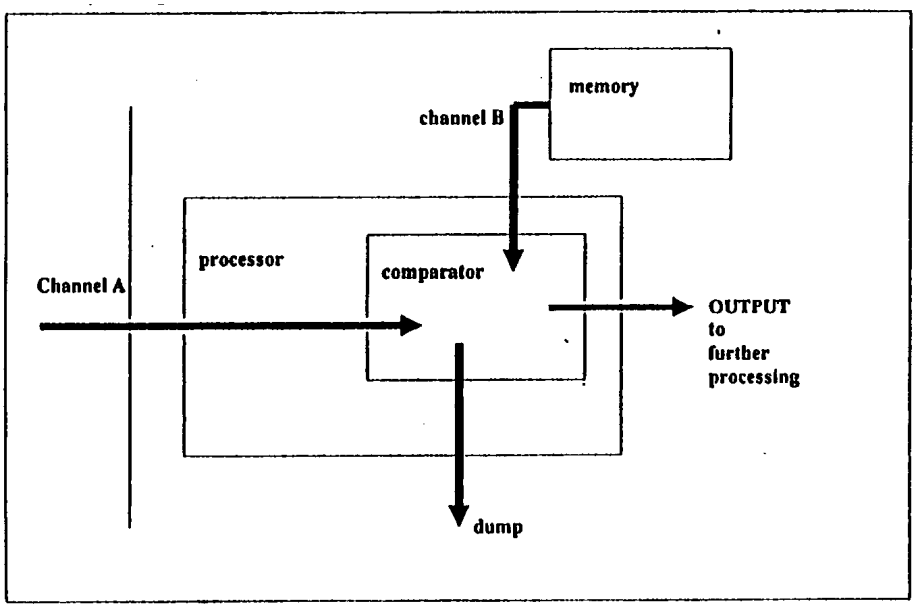


Fig. 18 The Associative Store Model of Speech Perception.

The abstract representation of the incoming soundwave is presented to a decoding processor along what we shall call channel *A* (see the Diagram). Part of the processor is a comparator whose job is to scan (or inspect) the incoming signal and compare it with information along a second pathway – channel *B*. Channel *B* enables the comparator to access memory in which is stored everything that the listener knows about speech in this particular language (i.e. it contains phonetic knowledge and the phonology of the language), including some lexicon.

Sooner or later as a result of the scan the comparator finds in the memory the item which most closely matches the signal input from the real world along channel *A*. Sometimes a near perfect match will be found, on other occasions the comparator will be less certain of the match and may select an item from the memory on some kind of a statistical basis – *It's most likely this one*.

When a match has been found a copy of the item is pulled from the memory for handing on for further (syntactic and semantic) processing. At the same time the original stimulus which came along channel *A* is discarded or dumped.

The essential properties of this theory are:

- it is a mentalistic, translation theory;
- *two* input channels are available to the phonetics/phonology processor;
- potentially degraded input triggers a memory scan to retrieve a best-match stored item;
- a copy of the stored item is passed for further processing;
- the original trigger signal is discarded;
- in some sense perceptual awareness is of the item retrieved from store, *not* of the stimulus trigger item.

The model presented so far does not explain all the data, however. We can observe that when encountering very degraded stimuli (e.g. under conditions of high ambient noise or over a noisy phone link) a listener seems automatically to opt for an attempt to decode even when there may be a significant probability of failure. We notice in addition that if subsequent syntactic and/or semantic processing reveals that the phonological item (say, a word) triggered from the memory is wrong (say, does not make semantic sense), then the system goes back to have another shot at scanning the phonological memory for a different item.

Now, in order to do this the original stimulus has to be evoked. But in the basic model

described above that stimulus has been long since dumped. So a way has to be found of holding onto incoming stimuli for a reasonable time just in case a rerun of the phonological decoding process turns out to be needed.

After the first run at decoding, the stimulus item is put into a first-in-first-out holding stack. A stack can be thought of as a temporary store comprising a vertical arrangement of a number of pigeon holes. Our stimulus, *S1*, is placed in the top pigeon hole. Along comes our second stimulus, *S2* which is placed in the top pigeon hole displacing *S1* down one, *S3* displaces *S2* down one and *S2* in turn displaces *S1* – and so on. There are around seven pigeon holes (the number is highly significant in cognitive science), and when *S8* ultimately causes *S1* to fall out of the bottom pigeon hole then the first item is finally lost.

This first-in-first-out stacking device is also known as a buffer. It buffers the loss of stimuli for a short period of time, just in case they are needed again in the event of an error being detected. The buffer is finite in length (i.e. there is a fixed number of pigeon holes) and therefore a limit on the number of stimuli that can be temporarily stored. Such a buffer is also known as short-term memory in the field of psychology.

So we add to our characterisation of the Associative Store Model:

- the detection of an error during later processing triggers a rescan by the comparator;
- the rescan is enabled by the holding of channel A stimuli in a buffer before dumping;
- the buffer is finite in length, holding around seven items.

The Associative Store Model of speech perception. or some variant of it, is currently the most favoured model of speech perception, though the re-emergence of the Theory of Direct Perception has caused some reassessment of it.

6.6 Some of the Facts about Speech Perception

Speech perception theories, like *all* theories, as based on observations which are often the result of careful experimental work. It is the interpretation of these facts that enables a coherent theory to be built. Unfortunately, perhaps because of incomplete data and varying interpretations of what data is available, theories can differ considerably and often make opposite claims. Here is an outline of some of what we think are the facts surrounding speech perception.

As with speech production, there are two aspects to the perception of speech:

- the physical aspect – the acoustic signal itself in the air outside the listener; and
- the psychological aspect – the interpretation of that signal by the listener.

The study of the relationship between these two aspects falls within the domain of psycho-physics, specifically in psycho-acoustics.

Consonants *vs.* Vowels

There is some feeling among researchers that the perception of consonants and vowels draws on *different* processes. Consonants and vowels can be classified separately by cues such as the presence or absence of periodicity (vocal cord vibration) in the signal and the temporal distribution of the acoustic energy (silent periods followed by bursts of aperiodic sounds in stops, for example).

The identification of consonants may primarily depend on relatively rapid transitions in the nature of the spectrum – that is, changes over fairly short time periods of the frequency distribution within the signal. They can be classified by detecting periods of silence, bursts of signal with an aperiodic source, changes in the formant transitions in adjacent vowels, and so on. The main component of many consonants is aperiodic sound, and some researchers feel that the right hemisphere of the brain (contralaterally connected to the left ear) dominates in the identification of consonants because it is more ‘sensitive’ to acoustic activity of this kind than the left hemisphere.

Vowels on the other hand have a more structured periodic waveform. The associated

acoustic signal is normally continuous in nature (unlike, say, the abrupt or discontinuous signal associated with stops), with a preference for left hemispheric sensitivity (contralaterally connected to the right ear). Whereas consonants could be said to assist in the demarcation of syllable sized chunks of the signal, vowels carry the more dynamic and prosodic features of speech such as stress, rhythm and intonation spanning more than one phonological segment.

Variability

People speak differently – we call this phenomenon inter-speaker variability. It often results in speech signals which differ significantly, when considered objectively, from one person to another. Even within the speech of a single speaker there are variations – this is called intra-speaker variability. Variability of the speech waveform makes both speech production and speech perception difficult.

There are many sources of inter-speaker variation, but we can list some of the main ones:

- differences in the child's surroundings during the learning process (dialect, accent, idiosyncrasies learned from parents, etc.),
- anatomical differences (e.g. resonating cavity differences, vocal cord length and mass variations),
- tempo (rate of delivery modifies the soundwave),
- degree of coarticulation (some people speak more precisely than others).

Notice that some of these will be learned, whereas others may have an intrinsic basis.

It is a remarkable property of the human perceptual system that despite these sources of variability correct identification and classification is possible. Human beings are said to be pattern seeking devices – this means that they cannot help but seek patterns in incoming stimuli (not just speech) and are particularly good at assigning patterns to the stimuli. One of the biggest problems in computer simulation of human behaviour is getting the machines to behave in this way – normal computers are very good at arithmetic and logic but very poor at pattern seeking and matching.