Tracing Lip Movements: Making Speech Visible

Ruth Campbell

Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OXI 3UD, U.K.

Visible Language XXII, 1
Ruth Campbell, pp. 32-57

Lipreading cannot deliver the phonetic structure of a spoken language very effectively; for no phoneme can be unambiguously identified from lip-pattern alone. Nevertheless, under some circumstances, speech that is not heard, but just seen by lipmovements on a speaker's face, can be understood and recalled verbatim. Moreover, under some conditions, heard speech that is different to that which is seen to be spoken, seems to 'fuse' to produce a different speech percept (The McGurk Effect).

These paradoxical aspects of lipreading and the constraints on the conditions under which lipreading can be helpful or can 'fuse' with heard speech are hard to accomodate within some theories of auditory speech perception. An interactive activation account is offered in which lipreading is considered to provide a phonetic feature - that of seen mouth opening and closing - to the speech analysis system. While such a feature appears to be necessary to account for these effects, it is not yet clear whether such a single seen phonetic feature may be sufficient for effective integration of seen and heard speech in all circumstances.

When people speak, their lips move. Is this natural visible consequence of articulation important in the normal perception of speech, or is it a useless, even disturbing epiphenomenon; best ignored unless we are hard of hearing?

It is clear that lipreading can
A. usefully complement heard speech that has been degraded by hearing loss or noise. It seems to be useful because it can sometimes offer phonetic cues that can be lost in noise. For example, /pa/ and /ka/ differ phonetically in where the plosive part of the sound is made. The first is made by puffing the closed lips open, the other can only be made with the lips apart, the plosion being effected by the back of the tongue hitting the velum. In terms of phonetic gestures needed to make these sounds, this distinction, that of *place* of articulation, is the major way in which they differ, and it is a distinction that has low acoustic energy characteristics and so can be easily lost when listening to speech in noise or with impaired hearing — but of course, it is a distinction that is easily seen.
B. add a comprehension component to *clearly* heard speech where material is difficult to follow because it is conceptually complex or pragmatically unclear (Reisberg et al. 1987).
C. support full comprehension without any heard input at all; for instance where the spoken material is limited by context to a few possibilities that are reasonably visually distinctive (Gailey, 1987). Silently spoken digits are a good example (Cambell & Dodd, 1980).
D. interact with heard speech to give illusory speech perceptions. So, when /pa/ is heard, while /ka/ is seen to be spoken, /ta/ is often reported as the heard sound (McGurk & MacDonald, 1976).

We know, further, that efforts to describe what makes a good lipreader have been unsuccessful in pinpointing any *particular* attribute. On the whole, what makes for good lipreading is what makes for good understanding of heard speech; an awareness of speech structures and their possibilities — and a flexible and powerful intelligence that can back such awareness (Gailey, 1987; Jeffers, 1967; Kitson, 1915).

We also know that infants, from their first days, are biologically predisposed to be sensitive to face movements and imitate them readily (Meltzoff & Moore, 1982). This sensitivity transmutes into more complex perceptual patterns which could come to be intertwined with speech perception (Dodd, 1987; Mills, 1987; Vinter, this volume).

Speech is probably best acquired and utilised *bimodally*. What one sees of the speaker seems to form a natural, sometimes a necessary complement to what is heard.The acoustic quality of heard speech in everyday settings often leaves much to be desired, yet we are able to understand such material with little effort.

Not only are we able to understand spoken speech distorted through various transmission devices (telephones, PA systems, wind, whispers) but usually the auditory attributes of natural speech bear no immediate invariant relationship to the natural 'units of speech' — phonemes, words. Instead, the articulatory and auditory context, as well as the lexical context and linguistic and semantic knowledge of the listener *set* the perception of the parts of an utterance in complex and interactive fashions. It is in this swirl of natural speech sound identification that lipreading can start to be understood and needs to be explained.

Theoretical Approaches to Lipreading

There are a number of plausible theoretical ways to accomodate lipreading within the context of natural speech perception.

The Motor Theory of Speech Perception

The simplest is perhaps the motor theory of speech perception. According to this proposal (Liberman & Mattingley, 1985), the invariance of heard speech perception in highly variable auditory contexts resides in the speech *gesture* — in the invariance of the highest form of speech articulation. The identification of a particular phoneme is a function of the ability to produce the phonetic structure that characterises it. In order to explain how a speech might be perceived as a speech sound, such theorists tend to make use of J. J. Gibson's theories of 'direct perception' (1950; 1966). The natural

invariance of the speech gesture reflects an emergent sensitivity to such meaningful units in the heard/spoken environment; a sensitivity that is genetically given. We hear speech sounds for what they are, whoever says them, however they are said, through a perceptual system which allows the abstraction of such identities as 'higher order perceptual invariants'. It was by such higher order invariance that Gibson claimed that we perceived visual aspects of the environment such as distance and shape. In this context it is natural that lipreading should have a place. If the invariant unit of speech perception is the gesture, not the sound made, then 'naturally', the perceptual analysis of the gesture will be as effective when it is seen as when it is heard. But, almost by definition, the mechanisms of direct perception are hard to instantiate. It becomes difficult to assess the explanatory, rather than descriptive, power of this aspect of motor theory.

Nevertheless, the motor theory of speech perception can direct us to many relatively unexplored aspects of speech perception. What would it predict, for example, to be the *lipreading* skills of the person struck mute by a brain lesion affecting the central cortical sites of speech production? Such central motor disorders do not always dissociate cleanly from perceptual ones; impairment of speech production is often intimately related to speech comprehension impairment. We are working on the neuropsychology of lipreading at present, in our laboratory.

Integration Theories of Speech Perception

Massaro's paper (this volume) suggests a different approach to speech perception, one which he has vigorously pursued for several years. Massaro's achievement is to indicate the *flexibility* of speech integration processes. His goal is to effect a powerful, predictive description of the metric whereby phonetic processing occurs. His solution is the specification of a fuzzy logical integration model which allows sensory information to be handled in a probabilistic and context–sensitive way. This approach invests computational power in the properties of the sensory systems themselves. Lipreading adds yet another component to the speech-integrative system.

One potential problem with this approach is that it may fail to discriminate effectively between natural and unnatural percepts. For example, the methodology that this approach uses requires the perceiver to make sense of a range of unnatural inputs. The extent to which he can do this will certainly tell us about how such decisions can be made — but runs the risk of equating such decisions with natural speech perception. Natural speech perception may well use processes that can be demonstrated by presenting unnatural patterns of stimulation; but the relationship is not transparent.

In fact, one critical test of an adequate theory of lipreading in speech perception must be whether it can distinguish, as normal hearers do, between 'natural' and 'unnatural' (automatic and voluntary, perhaps?) combinations of input. One must also ask, can the effects of reading written material and reading mouth movements be distinguished easily within the theory? If they cannot, then the theory will lack power at one of the most important points at which it should exert it. When one reads the syllable /ga/ while hearing /ba/ one does not experience an illusory, heard /da/; but when the visual stimulus is lipread, one can (McGurk & McDonald, 1976). An integration model (taken at face value) might suggest it might. This coarse example suggests that the highly flexible integration theory exemplified by Massaro's work may, nevertheless, not capture a crucial aspect of lipread speech, but may serve instead as an 'overmodel'; a description of the combinational rules involved in each and every sensory identification process.

Special Purpose Mechanisms: Can seeing the speaker add to hearing?

The particular problem posed by the McGurk illusion, where discrepant heard and seen speech can sometimes give rise to perceptual fusions and blends, has given rise to some thoughtful and provocative speculations by Summerfield (1987).

Rather than develop either a powerful, all-round model of heard and seen speech perception, Summerfield takes a more modest tack and asks what particular 'add-on' components might make an auditory theory of speech

perception lipread? A number of possibilities are considered; none of them completely viable, but all are systematically explored.

The McGurk illusion is the strongest evidence so far that speech that is seen and heard can generate a clear speech percept, different from that predicted from vision alone or from audition alone. The immediacy of the percept suggests that the integration occurs at the most basic level of speech recognition; in the identification of the spoken phoneme. (It is worth noting that this statement remains to be experimentally tested.) Since it is primarily the phonetic feature of *place* of articulation that can be seen on the lips, perhaps vision 'dominates' hearing for the extraction of such phonetic information; the integrated percept reflecting a best fit between the visually and the acoustically specified /ga/ and /ba/. This cannot be the whole story. Seen /ga/ and heard /ba/ fuse to generate / da/ or /ta/. But when seen /ba/ and heard /ga/ are synchronised, the illusory percept varies; /bda/ or /bga/ etc. can be reported (see Massaro, here). So if place of articulation is specified by the lips it is specified in a highly conditional manner.

Indeed, the 'ifs and buts' needed to describe the vagaries of illusory blend percepts lead Summerfield to consider other integration possibilities. Perhaps lexical identification, rather than phoneme identification, is the source of effective lipread-auditory processing? Why propose a word recognition model that sidesteps the phoneme? Because natural speaking causes troublesome problems of phoneme identification, due to coarticulation. Coarticulation describes the necessary way in which fluent speech causes the production of a specific speech sound to vary as a function of what is said before and after it. The movement and inertial constraints on the mouth, vocal cords and tongue are such that the natural production of the sounds of a word cannot be achieved in a simple, serial manner, "like pearls on a string". Consider speaking the two words, "bath" and "both". The only perceived distinction is in the interconsonantal vowel; but this changes the articulatory/acoustic specification of the consonants fore and aft of it. Yet these differences are not *heard* in running speech; the speech identification system has

The movement and inertial constraints on the mouth, vocal chords and tongue are such that the natural production of the sounds of a word cannot be achieved in a simple, serial manner, "like pearls on a string". Consider speaking the two words "bath" and "both". The only perceived distinction is in the interconsonant vowel; but this changes the articulatory/acoustic specification of the consonants fore and aft of it.

managed to absorb and take account of such coarticulatory effects. If you watch yourself in the mirror saying "both" and "bath" you will see that the lip positions of the consonants, too, are quite different in the two words. Coarticulation is as much of a problem for seen as for heard speech. Klatt's (1979) model of lexical identification on the basis of 'delayed commitment' to a particular interpretation of the auditory speech signal, can, Summerfield shows, be extended to include lipreading. Lexical identification, on this model, is achieved independently of phoneme identification– for phonemes are hard to specify*a priori.* Instead, the model allows concatenation of phonetic features into units that will vary with the specificity of the word to be recognized; that is, there can be a variety of 'units' including direct sound spectral specification, that can characterise a particular lexical item. Lipreading could work in such a 'direct access' model through lipreading-specified lexical representations, corresponding to the aural ones. But the problem with this approach is that it fails effectively to account for the gain that lipreading can give to auditory speech understanding, for the word specified by the lips will have similar coarticulation characteristics and problems as the word specified by ear. While this might mean that the invariant speech percept has its source at a level beyond the phoneme, it makes it hard to see just how lipreading can improve listening performance.

How can the lipread phoneme uttered in running speech achieve phonemic invariance? How do we know that /t/ was spoken when 'tooth' or 'teeth' is the word uttered? A glance in the mirror should confirm that the lip patterns of the two 't' sounds are utterly unlike each other because of the articulatory pattern of the aftercoming vowel. Here Summerfield suggests an articulatory theory may be useful. One of the skills of the listener-speaker may be the knowledge of vocal tract configurations, for this is reflected in the speaker's own ability to produce invariant speech sounds under different articulatory conditions. From this 'deep' dynamic knowledge will come awareness of the seen as well as the heard kinematic reflections of this set of movements. Very similar theories have been expounded to explain the developing child's perceptual sensitivity to the dynamics of hand and arm actions in

One of the skills of the listener- speaker may be the knowledge of vocal tract configurations, for this is reflected in the speaker's ability to produce invariant speech sounds under different articulatory conditions.

holding, catching, reaching, lifting (e. g. Mounoud & Hauert, 1982). An explication of this 'deep motor' approach to the recognition of auditory-visual vowels can be found in Summerfield and McGrath (1984).

Summerfield, then, suggests a range of 'add-on' components to account for lipreading in normal speech perception. These components are 'customized' to fit the particular model of auditory speech perception. The implication of this work is that each and every extension of each model may play a part in auditory-visual speech perception. However, as Summerfield's chapter title makes clear—these are *preliminaries* to a theory of audio-visual speech perception. Could a more integrated theory of vision in speech perception be attempted?

Lipreading in an interactive theory of speech perception: Can TRACE lipread?

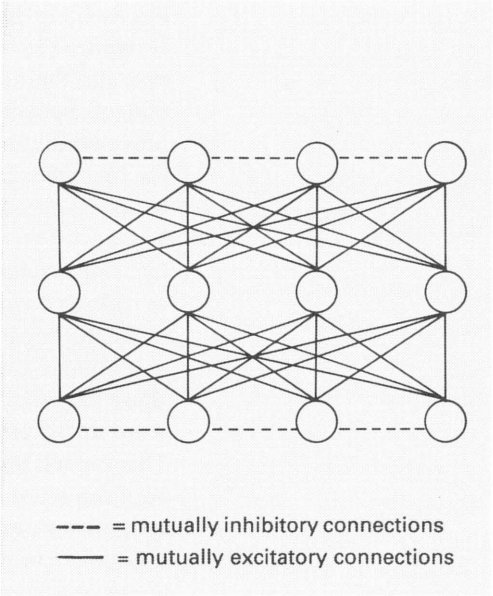
Interactive Activation Models

Twenty years of research on the recognition of written words have made clear that although the identification of letters and letter features is important, words can often be identified when the letters in them cannot. Text comprehension can set word identification, too. Letters in words are better identified than letters in non-word letter strings. Thus not only are bottom-up and top-down processes important in word recognition but these processes interact systematically. McClelland and Rumelhart (1981) and Rumelhart & McClelland, (1982) proposed an explicit interactive activation model to account for these phenomena. This model was the mother of a series of recent explanatory and predictive devices with similar formal properties. These have been used to explain, among other things, speech production (Stemberger, 1985; Dell, 1985) and auditory speech perception (McClelland and Elman's TRACE model of Speech Perception, 1986).

What are the formal properties of these models? Horizontal levels of representation are posited at increasingly molar levels from the 'purely sensory' to the cognitive. Thus for auditory speech perception a *phonetic*, a *phonemic* and a *lexical* level of representation are proposed. Connectivity between and across item representations is organised as follows: lateral inhibition is the dominant

A fully connected network with excitatory and inhibitory connections

Figure 1



connection type between represented items within each horizontal layer, that is, mutually inconsistent units inhibit each other. Across levels, however, the pattern of interaction is different. Here, mutually consistent units can excite each other (the original model of visual word recognition includes both excitatory and inhibitory cross-unit connections, but such inhibitory cross-level connections tend to make the model too inflexible in dealing with poorly and partly specified information). Finally, patterns of excitation and inhibition across and within layers work in cascade; there is temporal recruitment of these processes. Thus a particular stimulus array generates a temporary and highly dynamic pattern of excitation across the network. This then settles to a stable distributed pattern of activity. It is this stable pattern of excitation — rather than the firing of a particular 'node' in the array — that is the concomitant of categorisation or identification of a particular stimulus.

Such highly interative, distributed models, might, at first sight, appear to be *too* interactive to generate anything but noise, but, as recent theoretical and simulation developments show, they can be both precise and practical in implementation. They behave like human beings (see McClelland and Rumelhart, 1986).

For auditory speech perception an important component is required in addition to the three levels of representation and their interconnections. This is the TRACE component. Spoken words take time to say and this is reflected in their characteristic recognition time. One of the first speech recognition models to take seriously the temporal characteristic of spoken language was Marslen-Wilson & Tyler's COHORT model which showed how decisions concerning word identification start to operate as soon as a word starts to be spoken, rapidly and automatically constraining the identification of a possible word depending on the size of the cohort of words that share the same phonemic features up to a critical, unique word decision point (see Marslen-Wilson & Tyler, 1981). This model, however, cannot backtrack; it is a feature of auditory word recognition that we are easily able to recognise words whose initial phonemes may have been misspoken or misheard. Then, if co-articulation effects are to be accomodated in a sensory, rather than a motor theory, it is necessary for the speech recognition process to take account of aftercoming speech context, as well as prior context, in order to achieve invariant phoneme perception. An auditory theory of speech perception should incorporate a delay window to account for such right-context as well as mispronunciation effects. The TRACE serves this purpose. TRACE is not an acronym; it is a literal description of the sustained state of activation of the system; the TRACE, that is the pattern of activation corresponding to a not-yet resolved stimulus pattern, persists until categorical recognition of the speech sound is achieved. In this crucial sense, this model is not an 'on-line' model of speech recognition, in the way that the COHORT model is; rather it suggests, along with others (e.g. Crowder, 1983), that the distinction between perception and immediate memory in speech sound processing is a blurred one.

To summarize the general history of this type of model: a visual word recognition model that has different levels of representation all of which are fully interactive each with the other, has been extended into the time dimension, both in its representational and activation components, in order to process heard speech. *Representations* need to be temporally organised (for example, in the specification of

the phonemes required to distinguish 'god' and 'dog'), but also the *state of activation* of each unit at each level persists as long as and until a final, categorical, decision can be made.

Lipreading by TRACE?

Now, how could a TRACE model accomodate lipreading? Let us remind ourselves what such a model should achieve. It should

1. explain how lipreading can aid noisy speech perception; it should also suggest how clear speech can be helped by lipreading and how silent lipreading can be achieved.

2. predict the specific patterns of interaction of the blend and fusion illusions. In particular, it must distinguish between the effect of a seen /ba/ and a heard /ga/ (/bga/, /bda/, etc. . .) and the opposite conjunction (which leads to a fusion like /da/).

Additionally, if TRACE works for lipreading, we might expect some similarity between heard and lipread material in terms of a close relationship between perception and immediate memory processes for lipreading and for hearing.

Can TRACE do this? Clearly in such fully interactive models it would be possible to introduce one or several lipreading features. So let us start with the most basic proposal of all; that the only visual (lipread) feature that the model permits is that of mouth opening and closure. Where should this feature be introduced? Let us, again for simplicity, introduce it at the bottom; as an additional *phonetic feature*. Note that because TRACE is a theory in which identification is a function of distributed processes at different levels of representation, this will mean that at all other levels of representation than the phonetic feature level, the lipread information will have some effect; it will leave a trace.

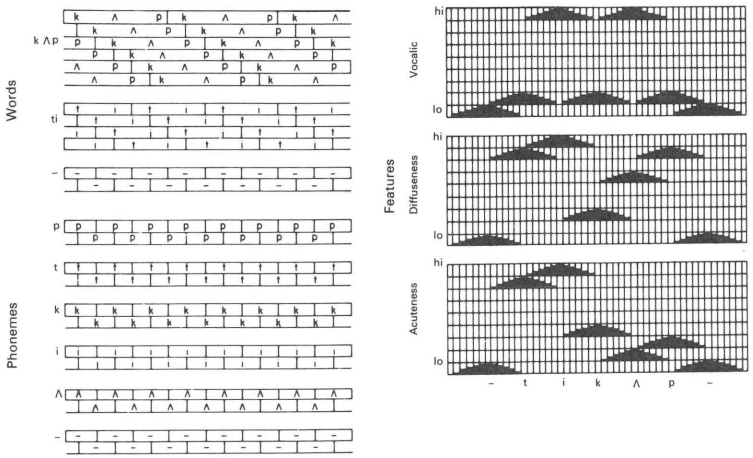
General Effects of Introducing a Visible Phonetic Feature: Mouth Closure

Simply by virtue of the addition of a seen component at the feature level, the state of stable activation corresponding to identification of a word will be more efficiently achieved when one can see as well as hear the

speaker; an additional feature, relevant to speech perception, is added to the system. In this way, the gain of lipreading a clearly heard speaker can be indicated. Where speech is noisy, the phonetic feature specification of the *acoustic* phonetic features will lead to unstable, persistent activation of many possible categorical speech sounds; under these conditions a clearly seen face will improve recognition for those sounds that can be distinguished visually.

How would this work in detail?

Figure 2



A subset of the units in TRACE II. Each triangle represents a different unit. The labels indicate the item for which the unit stands, and the horizontal edges of the rectangle indicate the portion of the Trace spanned by each unit. The input feature specifications for the phrase "tea cup", preceded and followed by silence, are indicated for the three illustrated dimensions by the blackening of the corresponding feature units.

Figure 2, from McClelland & Elman, shows how the auditory input "teacup" corresponds with three levels of unit representation in TRACE. Only the pattern of excitation at the phonetic level is indicated for clarity. Furthermore, only three acoustic feature dimensions are indicated for simplicity. The full simulation model used seven acoustic dimensions including voicing, consonantality, power and burst, as well as the three indicated here.

Consider the phonemes /k/ and /p/ in the figure. In terms of the three acoustic phonetic feature dimensions shown they are similar, differing only in a small change in the level of activation for diffuseness and acuteness. What would the pattern of activation be like for a seen and heard 'noisy' "cup"? Remember the principles of

Remember the principles of activation are these; within a level, units that are inconsistent have mutually inhibitory connections. Across levels, units that are mutually consistent have mutually excitatory connections.

activation are these; within a level, units that are inconsistent have mutually inhibitory connections. Across levels, units that are mutually consistent have mutually excitatory connections. In noise, the acoustic feature dimensions for identifying /k/ and /p/ would not be inconsistent for some would be too poorly specified, and there could not be mutual inhibition between these feature patterns on this basis. So if the recognition of the spoken word were dependent solely on acoustic inputs, the failure of inhibition at the phonetic feature level would mean that a large range of possible phonemes and phoneme combinations would all be relatively excited, and the identification decision on the word would be delayed, or even wrong. Typically, this characterises the identification of heard words in noise.

But the visual feature of mouth-opening is consistent with a range of phonemes (e.g., k, g, t, d), and inconsistent with others (b, th, m, p). Thus, at the phonetic feature level the *seen* sounds /p/ and /k/ will exert maximal mutual interference. So 'pup', 'puck', or 'cuck', cannot be activated by lipread "cup". In this way, *some* distinctive feature information can travel up the system when lipreading noisy speech. This has been refined by lateral inhibitory mechanisms at the phonetic level *without* necessarily being sufficient to enable particular phoneme identification to take place. If, at the lexical level, the general pattern of activation of phonetic features and of possible phonemes is more consistent with 'cup' than with any other word, the principle of mutually consistent units causing excitation across different levels enables the phonemic and phonetic aspects of that word to become further activated. Thus, top-down information affects the categorical perception of lower level units. Lexical superiority effects should arise in lipreading noisy speech, and common sense suggests that they do. But it is important to remember that TRACE allows a relatively stable pattern of activation to be established sufficiently to identify a word without necessarily being sufficient to identify the constituent phonemes in the word. Klatt's (1979) solution to the problems of coarticulation (and Summerfield's extention of this direct lexical access theory to the domain of lipreading) can be reinterpreted. Interactive activation theory allows words

to be identified without full identification of their constituents. Because lipreading can add a featural component to the phonetic activation pattern, it will aid hearing under many conditions, but particularly when auditorily-based phonetic feature components become less discriminating, through noise or disease.

Improving Lipreading Comprehension

If there is only one seen phonetic feature but several acoustic ones (which may or may not be orthogonal to each other), then the conditions under which lipreading could affect heard speech will be more constrained than the conditions under which hearing could affect lipreading. And they are (Easton & Basala, 1982). It is conceivable that all or any acoustic inputs, if congruent with defined speech percepts, will help the lipreader. One example of this might be the effect of lipreading with an auditory pulse train (Rosen, Fourcin & Moore, 1979). When an acoustic signal, corresponding to the activity of a speaker's vocal cords, is heard in synchrony with the speaker's lip movements, there is a marked improvement in lipreading comprehension for connected discourse over silent lipreading alone. Under these conditions only two potential phonetic feature dimensions are available to the perceiver; mouth-opening and voicing, yet, not only are these feature dimensions useable, they often seem to generate the illusion of hearing noisy speech, which persists until one loses sight of the speaker (the auditory pulse train alone cannot support speech comprehension).

Interactive activation theory allows words to be identified without full identification of their constituents.

The relatively greater effect of heard inputs on speech identification than of seen (lipread) ones can, in an interactive activation model, help to explain why the best lipreaders seem to be people who are generally good at language skills. They could use information at all levels of representation more efficiently to improve the value of lipreading. There may still be individual differences, of course, in the precise level of efficient unit representation (see Gailey, 1987).

It should also be clear how it is possible to lipread rather well when context constrains the number of possible words that could be spoken; in other words, how top-

down processing improves lipreading. Under these conditions only a subset of potential words are possible targets for identification and hence activation. Thus digits can be easily lipread from silent speech, as can footballer's (English) expletives (as seen on TV). Such phenomena suggest that a lipread phonetic feature detector can sometimes provide sufficient information for effective speech recognition by sight. Whether born-deaf and post-lingually deafened people show similar activation patterns for words that both groups lipread well is a subject for experimental research.

Audio-Visual Fusion and Blend Illusions

The simple principle of mutual inhibition between inconsistent feature units can also help to explain the contingencies of the audio-visual (McGurk) illusions. An open mouth is consistent with a range of consonantal speech sounds including /k/, /g/, /t/, /d/. It is also consistent with all vowels. A closed mouth is not consistent with any of these but with the consonants /p/, /m/, /b/. Both lip positions are consistent with no sound at all — the sound of silence cannot be *unambiguously* seen.

When /pa/ is heard and /ka/ is seen, the 'classical' fusion reported is that of a heard "ta" (McGurk & MacDonald, 1976). Those authors suggested that place of articulation of the illusory consonant was calculated by a compromise between the seen (front of mouth) and the heard (back of mouth) place of articulation of the tongue to a more intermediate one. /ta/ is an alveolar sound, made by placing the tongue just behind the dental ridge, at the front of the hard palate — though there are allophones of this sound with more variable places of articulation. But it may not be correct to suggest that /ta/ and /ka/ are systematically distinguished from each other solely by position of the tongue with respect to the front-to-back of mouth dimension (jaw drop may be a feature). In any case, in terms of an interactive model, the first stage of the processing of the dubbed seen and heard syllables is that of phonetic feature unit activation.

Figure 3

Phoneme values for /p/, /t/ and /k/ as used in TRACE II (McClelland & Elman, 1986) with mouth-opening added.

Feature	Phoneme		
	p	t	k
power	4	4	4
vocalic	1	1	1
diffuse	7	7	2
acute	2	7	3
consonantal	8	8	8
voice	1	1	1
burst	8	6	4
seen feature			
mouth open	1	8	8

What are the acoustic activation patterns of the consonants, 'p', 'k' and 't' in terms of such phonetic feature activation? Figure 3 shows the full values used in McClelland & Elman's TRACE simulation of auditory speech discrimination (McClelland & Elman, 1986, p. 15). Now we can ask how, when 'pa' is heard and 'ka' is lipread, does 'ta' seem to have been spoken? How, in other words, can the phonetic activation pattern needed for a /t/ decision be mimicked by a combination of /p/ and /k/ features?As we saw in the exampleon the opposite page, the activation patterns for acoustic phonetic features are quite similar. The differences (the inconsistent and therefore inhibitory features) are, first, acuteness; 'ta' is inconsistent with 'pa' and 'ka' and second, diffuseness; 'ka' is inconsistent with 'pa' and 'ta'. Note also, that on one dimension, burst, 'ta' is intermediate in value between 'pa' and 'ka'. In terms of the seen feature of mouth opening, 'pa' would be inconsistent with 'ka' and 'ta'. Presumed values are shown in the table.

Superficially, the inconsistency on two acoustic dimensions might suggest that these syllables are unlikely to be confused with each other. But the acoustic correlate of acuteness is high frequency spectral energy; that is, the energy that is most likely to be lost with small amounts of noise (put another way, "pa", "ka", and "ta" are all likely to be confused with each other when even small amounts of white noise are added to the acoustic

signal). So, with a little noise, acuteness is unlikely to be a reliable dimension for identifying these phonemes.

Now if 'pa' is heard, while 'ka' is seen, 'ta' could be the predicted percept if:
1. a small amount of white noise accompanies the acoustic signal, rendering the potentially distinctive feature of acuteness nondiscriminating.
2. the *visible* feature of 'open-mouth' is activated. This will *inhibit* /p/ activation by the principle of mutual inhibition between mutually inconsistent units.
3. the remaining feature that distinguishes /k/ and /t/ — that of burst quality — reflects relatively more inhibition of the low value /k/ by high value /p/ than of medium value /t/ by /p/.

This point is worth stressing; the apparent 'compromise' decision on place of articulation results from relatively greater inhibition between the more extreme feature values and less between each of these extreme values and a middle one. Lateral inhibition at the feature level achieves apparent compromise at the phoneme level. It may be worth noting that the same percept — 'ta' from 'pa' and 'ka' has been reported when 'pa' and 'ka' are heard in each ear. Presumably the 'open mouth' feature is not the only one that can produce a 'ta' percept. The *converse* pattern of input; seen 'pa' synchronised with heard 'ka' cannot give rise to the same, illusory 'ta'. Closed mouth detection generates inhibition between 'pa' and 'ka' or 'ta'. The perceptual system resolves this, usually, by ascribing the discrepant inputs to different times slots; 'pka', 'pta', being common reports of this stimulus configuration. Because vision and hearing are mutually inhibitory for this stimulus configuration at the feature level, such blends, (i.e., more than one perceived sound, rather than the unitary fusion illusion, where only one sound, usually different from the one that was acoustically present, is reported) are the *only* permissible percepts for a lipreading TRACE model.

Place of articulation is not determined more by vision than audition, but vision contributes to the pattern of interactive activation at the phonetic level. This allows place of articulation to emerge as a seen feature in a systematic and predictable way.

Place of articulation is not determined more by vision than audition, but vision contributes to the pattern of interactive activation at the phonetic level. This allows place of articulation to emerge as a seen feature in a systematic and predictable way.

Rate of Articulation

Seeing the mouth open and close can inform the speech processing system about another phonetic dimension. The identification of voiceless plosives, like /pa/, is contingent on the perceived auditory rate of speech (Summerfield, 1981). Green & Miller (1986) have shown that the perception of speech rate does not have to be auditory to affect the categorisation of a heard speech sound as /pi:/ or /bi:/. The rate of seen lip-movement can shift categorical perception of the voiceless plosive in a similar manner to the heard rate of speech. This highly context-contingent phonetic effect, can, of course, be accomodated by TRACE as comfortably as other effects of coarticulation, through the delayed commitment principle that TRACE embodies.

Place of articulation is not determined more by vision than audition, but vision contributes to the pattern of interactive activation at the phonetic level. This allows place of articulation to emerge as a seen feature in a systematic and predictable way.

Immediate Memory Processes: TRACE, PAS and Other Phenomena

If TRACE is a good model in which to accomodate lipreading in speech perception, it should indicate ways in which lipread and heard material might show similar immediate memory characteristics. In particular, if it is a powerful model, it might distinguish between such effects and those for written material. What distinguishes the immediate memory processing of written and heard material? While several lines of investigation are being pursued (see, for example, Dodd, Oerlemans & Robinson, this volume), one route has been extensively travelled. This is the investigation of immediate list recall, which shows a very robust effect of auditory recency. The last item of a heard list is easier to remember than earlier list items; this recency effect is less marked, or absent altogether, when the list items are read, rather than heard. Auditory recency can be eliminated by a heard speech sound after the end of the list. This makes little difference to written list recall. Because this suffix effect is not dependent on lexical status of list and suffix, it can be considered precategorical. The combination of auditory recency and suffix effects led Crowder & Morton (1969) to suggest that auditory lists leave a record of their *acoustic* properties in the cognitive system — a precategorical acoustic store (PAS). If this were a sensory-acoustic trace then lipreading should not produce recency and suffix

effects similar to those for hearing spoken lists. But it does (Campbell & Dodd, 1980, 1982, 1984; Greene & Crowder, 1984). Moreover lipread lists are affected by heard suffixes, and heard lists by lipread suffixes; and these effects are highly specific to those input combinations (Gathercole, 1987; Campbell, 1987).

TRACE suggests that, because of deferred commitment to phonetic decisions, just such a record can persist, which will be instigated by phonetic rather than acoustic activation. If the lipread feature of mouth-opening has access to the TRACE system then similarity and interactivity between heard and lipread recency and suffix effects are to be predicted. They will occur because, in recall, TRACE activitation allows a 'second look' at the stimulus array, and the most recently presented material will be relatively more accessible in this phonetically active state. Since reading does not activate such a feature-based, persistent trace, written lists do not *usually* show recency (but see Campbell, 1987 and Massaro, this volume).

As an aside, it may be worth noting that the principle of deferred commitment embodied in TRACE, might lead to recency effects not only for heard lists, but also for other material that demands that categorisation wait on serial presentation of featural information. Campbell, Dodd & Brasher (1983) showed that recall of serially presented lists of unnameable arrow shapes showed recency that depended on the order of display of the discriminating arrow fleche compared with the non-discriminating arrow shaft for each item as it was shown. When the fleche was presented before the shaft, (no deferred commitment to a categorical decision was needed to identify and recall this type of item), no recency was observed. When the shaft was presented before the discriminating fleche, recency occurred.

The deferred commitment principle embodied in TRACE would suggest, moreover, that auditory/lipread recency and suffix effects do not reflect fixed-size storage capacities for phonetic material but, rather, that recency and suffix effects will vary with the extent to which deferred commitment characterises the operating system. The recognition of spoken words, because it is so automatised in skilled hearers, may appear inflexible, giving the

impression that fixed storage size is a characteristic of the immediate memory system that gives rise to recency/suffix effects.

Is Mouth Opening Enough?

The simple detection of mouth opening and closing was proposed as a sufficient feature for a TRACE model that lipreads. But what is the evidence that mouth closure, rather than some other aspect of lip-movement, is the critical feature? Could mouth closure be one of several features that the lips can offer to speech recognition? Summerfield (1979) examined a range of visual manipulations of seen mouth movement to see whether they could be distinguished in clarifying noisy heard speech. The control condition, which provided a significant lipreading advantage over unseen speech, was a synchronised videotape of the lower (full) face. Significantly useful, also, were disembodied lips—that is the lips painted with ultraviolet reflecting paint, videotaped under ultraviolet light. The tongue and teeth are *not* visible in this condition. There was a significant difference between lipreading gain for the control and the disembodied lips condition; *something* is missing from the disembodied lips that can help in the visual clarification of heard speech. However, two conditions gave *no* lipreading advantage. These were the movement of the illuminated four corners of the mouth and the presentation of a visible annulus whose inner diameter varied with vocalisation of the stimulus. Both these conditions display a visual stimulus that corresponds, formally, with properties of the heard speech. But neither of these displays indicated lip closure.

So it seems that lip-closure is a necessary feature of effective lipreading. But is it sufficient? Are more features needed? Kuhl & Meltzoff (1984) examined 19 week-old infants' sensitivity to face-voice synchrony. The infants were shown sequences of the vowel sounds 'ah' or 'ee' being spoken. These were synchronised to the same or the other vowel being produced on the auditory channel. The infants looked significantly longer at the correctly synchronised than the wrongly dubbed face and voice. The investigators confirmed that the identity of the spoken vowel was important, rather than just the temporal

It would seem that lip-shape, as well as lip-opening can be important in extracting speech from faces.

synchronisation of seen and heard sounds, by replacing the vowel with a pure tone signal that maintained the duration of the spoken vowel, its amplitude envelope over time and its synchronisation to the visual signal. Now the infants showed no preference for one face over the other, but responded in the arbitrary way that they looked at wrongly dubbed faces.

Since only spoken vowel shape was manipulated in this experiment, it would seem that *lip-shape*, as well as *lip-opening* can be important in extracting speech from faces. A further piece of evidence suggests that, indeed, lip-shape can be a useful and important feature of phonetic perception. Summerfield & McGrath (1984) examined the integration of discrepant seen and heard vowel sounds. If lip shape were relatively uninformative about vowel identity—if it simply signalled a possible vowel,—then one would not expect that auditory- visual fusions or blends should occur when viewing one lip shape and hearing another. Yet such blends do occur when viewing one lip-shape and hearing another, though not always in the clearcut, categorical fashion that they do for some stop consonants. Storey & Roberts' simulation (this volume) assumes lip-rounding has a role to play, too.

Should lip-shape be considered as a phonetic feature orthogonal to lip-closure in a speech recognition system? Or are lip-closure and lip-rounding nonindependent aspects of the same, essential feature? It is possible to think of vowel movement from 'ee' through 'ah' to 'oo' as following the movement from lip closure to lip-opening. But this question is worth careful experimental investigation. So too are questions concerning the necessity of tongue and teeth visibility. In Big Nambas, it seems, the bilabial – /pa/ is distinguished from a very similar sound produced by smacking the tongue, voice-lessly, against the outer part of the upper lip; a highly visible, but hardly hearable distinction (Ladefoged, 1985).

Here, then, are fruitful areas for further research. An interactive theory of the role of lipreading in auditory speech perception can focus and drive the search for an understanding of the processes involved in speech perception, whether it is seen or heard. TRACE theory seems,

at present, to be the most powerful and accomodating of all theories of speech perception. With a visual (lipreading) component it can be usefully extended and can provide an answer to some of the puzzles that face us when we listen with our eyes

About the author

Ruth Campbell is a University Lecturer in Experimental Psychology at the University of Oxford, England. Her research has been on the neuropsychology of lipreading and, with Barbara Dodd, on immediate memory for lip-read lists. While these seem typically obscure and trivial subjects for psychological research they have turned out to be useful in indicating just where the conceptual line needs to be drawn when considering what underlies the perception of auditory and of visual material. Her present research interests also include deafness and cognition as well as aspects of the processing of facial information (other than reading speech from them) and also reading and writing. Her own hearing problems might have contributed to these research interests, though she is a bad lipreader!

References

Campbell, R. 1987. Common processes in intermediate memory. In Allport, D. A., MacKay, D., Prinz, W. & Scheerer, E. (Eds.) *Language Perception and Production: Common Mechanisms in Listening, Speaking, Reading and Writing.* NY: Academic Press, 131–149.

Campbell, R. & Dodd, B. 1980. Hearing by eye. *Quarterly Journal of Experimental Psychology,* 32, 85–99.

Campbell, R. & Dodd, B. 1982. Some suffix effects on lipread lists. *Canadian Journal of Psychology,* 36, 509–515.

Campbell, R. & Dodd, B. 1984. Aspects of hearing by eye. In Bouma, H. & Bouwhuis, D. G. (Eds.) *Attention & Performance,* 10, L.E.A., Hillsdale, 300–311.

Campbell, R., Dodd, B. & Brasher, J. 1983. The sources of visible recency; movement and language in immediate serial recall. *Quarterly Journal of Experimental Psychology,* 35A, 571–587.

Crowder, R. G. 1983. The purity of auditory memory. *Philosophical Transactions of the Royal Society of London, B,* 302, 251–265.

Crowder, R. G. & Morton, J. 1969. Precategorical Acoustic Storage (PAS). *Perception & Psychophysics,* 5, 365–373.

Dell, G. 1985. Positive feedback in hierarchical connexionist models: applications to language production. *Cognitive Science,* 9, 3–23.

Dodd, B. 1987. The acquisition of lipreading skills by normally hearing children. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates

Dodd, B. & Campbell, R. 1984. Non-modality specific speech coding. *Australian Journal of Psychology,* 36, 171–184.

Easton, R. D. & Basala, M. 1982. Perceptual dominance during lipreading. *Perception & Psychophysics,* 32, 562–570.

Gailey, L. 1987. Psychological Parameters of Lipreading skill. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates. 115–137.

Gathercole, S. 1987. Lipreading: Implications for short-term memory. In Dodd, B. & Campbell, R. (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates. 227–242.

Gibson, J. J. 1950. *The Perception of the Visual World.* Boston: Houghton.

Gibson, J. J. 1966. *The Senses Considered as Perceptual Systems.* Boston: Houghton.

Green, K. P. & Miller, J. L. 1985. On the role of visual rate information in phonetic perception. *Perception & Psychophysics,* 38, 269–276.

Greene, R. L. & Crowder, R. G. 1984. Modality and suffix effects in the absence of auditory stimulation. *Journal of Verbal Learning & Verbal Behavior,* 23, 371–382.

Jeffers, J. 1967. The process of speech reading. *Conference on Oral Education for the Deaf,* 1530–1561.

Kitson, H. O. 1915. Psychological Tests for Lipreading Ability, *Volta Review.*

Klatt, D. 1979. Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics,* 7, 279–302.

Kuhl, P. K. & Meltzoff, A. N. 1984. The intermodal representation of speech in infants. *Infant Behavior & Development,* 7, 361–381.

Ladefoged, P. 1985. *Unpublished lecture to the Laboratory of Phonetics,* University of Oxford.

Liberman, A. & Mattingley, I. 1985. The motor theory of speech perception revisited. *Cognition,* 21, 1–33.

Marslen-Wilson, W. & Tyler, L. 1981. Central Processes in speech understanding. *Philosophical Transactions of the Royal Society, London, B,* 295, 317–332.

McClelland, J. L. 1979. On the time relations of mental processes; an examination of processes in cascade . *Psychological Review,* 86, 287–330.

McClelland, J. L. & Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology,* 18, 1–86.

McClelland, J. L. & Rumelhart, D. E. 1981. An interactive activation model of context effects in letter perception. *Psychological Review,* 88, 375–407.

McClelland, J. L. & Rumelhart, D. E. (Eds.) 1986. *Parallel Distributed Processing.* Cambridge, Mass: MIT Press.

McGurk, H. & MacDonald, J. 1976. Hearing lips and seeing voices. *Nature,* 264, 746–748.

Meltzoff, A. & Moore, K. M. 1982. The origins of imitation in infancy: paradigm, phenomena & theories. In L. P. Lipsett & C. K. Rovee- Collier (Eds.) *Advances in Infancy Research.* Norwood, Ablex, 263–299.

Mounoud, P. & Hauert, C. A. 1982. Development of sensorimotor organisation in young children. In G. Forman (Ed.) *Action and Thought: from sensori-motor schemes to symbol operations.* New York: Academic Press.

Reisberg, D., McLean, J. & Goldfield, A. 1987. Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates, 97–114.

Rosen, S. M., Fourcin, A. J. & Moore, B. C. J. 1979. Voice pitch as an aid to lipreading. *Nature,* 291, 174–177.

Rumelhart, D. E. & McClelland, J. L. 1982. An interactive activation model of the effect of context on perception (part 2), *Psychological Review,* 89, 60–94.

Stemberger, J. P. 1985. An interactive activation model of language production. In A. W. Ellis (Ed.) *Progress in the Psychology of Language, 2.* London: Lawrence Erlbaum Associates.

Summerfield, Q. 1979. Use of visual information for phonetic perception. *Phonetica,* 36, 314–331.

Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: *Human Perception & Performance,* 7, 1074–1095.

Summerfield, Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates, 3–52.

Summerfield, Q. & McGrath, M. 1984. Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology,* 36A, 51–74.