# MULTISENSORY SPEECH PERCEPTION: INTEGRATION OF SPEECH INFORMATION

Michael Oerlemans and Peter Blamey

Australian Bionic Ear and Hearing Research Institute

ABSTRACT - Speech perception usually occurs in the context of more than one source of sensory information. The effectiveness of speech perception in these multisensory situations depends on how successfully two or more sources of information are combined. An analysis of auditory-visual (AV) speech perception suggests a model of multisensory integration which argues that the extent of integration depends on the physical and cognitive characteristics of the signals which are combined. It is proposed that such a model can explain observed differences between AV, auditory-tactile (AT) and visual-tactile (VT) combination.

## INTRODUCTION

Speech perception occurs in the context of multiple sources of information; sensory, syntactic and semantic, and contextual. One of the aims of psycholinguistics is to determine the relative importance of each of these sources of information, the degree to which they interact, and the nature of this interaction. Recent research and models of auditory-visual combination have shown that heard and seen speech are combined early in perceptual processing, at a feature level. In the case of the hearing-impaired, information derived from the auditory channel is degraded or absent, however, they may perceive speech through the use of prosthetic devices, including multi-channel electro-tactile devices. Perceiving speech in face-to-face interactions using such a device will depend on the ability of the device to represent acoustic parameters tactually, *and*, the ability of the perceptual system to combine information with lip-reading and residual hearing. This paper will focus on an examination of the characteristics of auditory and visual speech and investigate the application of a model derived from this to explain differences between auditory-visual, auditory-tactile and visuo-tactile performance.

## AUDITORY-VISUAL SPEECH PERCEPTION

It is clear that in many cases a bisensory perceptual advantage is found over unisensory performance for AV combination (e.g. Jeffers and Barley, 1971), AT combination (e.g. Cowan, Blamey, Galvin et al, 1990) and VT combination (e.g. Osberger, Rines-Weiss and Kalberer, 1986). These multisensory perceptual advantages occur for both normally hearing and hearing impaired subjects (e.g. Cowan, Alcantara, Blamey and Clark, 1988). While it is evident that multisensory speech perception is generally better than perception in single modalities, the basis of this perceptual advantage is still at issue. The crucial question is whether or not this bisensory advantage is simply the result of having access to two sources of information, *used selectively*, or whether the advantage is the result of special information available *in combination* which is not available when combination does not occur.

## THE BASIS FOR INTEGRATION

A number of physical characteristics of heard and lip-read speech, and cognitive characteristics of perceptual processing may set boundaries for understanding when integration or combination will occur including; sensory channel, sensory signal, redundancy relations, spatio-temporal characteristics of the stimuli and cognitive architecture. The arguments outlined below will focus on the representation of segmental features of language. It is clear that suprasegmental aspects of articulation, facial gestures and prosody in sign language are necessarily combined at some stage in information processing to generate meaning, however for the purposes of this discussion they are secondary.

### Sensory channel

Sensory information may be integrated to the extent that it is processed by the same perceptual channel. Fodor (1983) argues that sensory input systems have a number of characteristics which uniquely define them, and separate them from central cognitive systems. These systems are domain specific and correspond to the traditional five senses plus a system for language (Massaro, 1987:2).

Input systems process information in qualitatively different ways. Given this premise one might argue that auditory and visual speech information are processed separately. In contrast, sign language and lip-reading, both operating through visual sensory input, would be expected to interact.

Sensory signal

The structural characteristics of a signal need to be compatible in order for integration to occur. Hearing, lip-reading, electro-tactile and electro-cochlear stimulation perceptually encode a signal derived from articulation. In the case of lip-reading, the signal is undoubtedly visual, but it is a visual representation of articulation. Sign language, on the other hand, encodes gesticulation. Sign language cannot be combined with heard or seen speech at the phoneme or feature level, because it does not represent speech sounds. It does however represent words, and combination of signed and heard speech may occur at the word level. Auditory and visual speech, on the other hand, perceptually represent input derived from articulation. One reason for integration of auditory and visual information may be this structural congruity.

Redundancy

Codes which have the same structural characteristics may be redundant or complementary, when they are combined. Redundant stimuli encode similar information at a given level of representation. Where redundant information is identical, perceptual combination occurs; if an individual hears a particular word in one ear, and hears the same word in another ear, only one percept is perceived. When redundant information is different, two separate percepts usually occur. For example, in dichotic listening different information is presented to either ear. Subjects report one or the other percept, subject to certain stimulus conditions (e.g. Kimura, 1973).

Auditory-visual speech perception, however, provides a case where information may be redundant on some features, but complementary on other features. Auditory information under degraded or noisy conditions tends to result in the loss of *place of articulation* information before manner or voicing (Summerfield, 1987). Lip-reading, on the other hand, represents (front) places of articulation relatively well. Auditory and visual information are *complementary* for manner and voicing in the sense that these are features poorly represented in lip-reading, and well represented in hearing. AV perception therefore results in combination of place information while tending to preserve manner and voicing. The best example of this is the AV fusion illusion (McGurk and MacDonald, 1976). In the AV fusion illusion, combining an auditory stimulus with a discrepant visual stimulus results in perception of a novel syllable. For example, on hearing /ba/ and seeing /ga/ subjects tend to report /da/. If perceptual codes are integrated to the extent that they are complementary on a number of features we would expect integration of auditory and electro-tactile speech with lip-read speech, but not for combinations of these perceptual modalities with each other. Where modalities are highly redundant, perception in one modality may interfere with perception in other modalities.

Spatio-temporal structure

Campbell and colleagues argue that the basis for integration of auditory and visual information is in the structure of the signal, rather than the channel of input (Campbell and Dodd, 1980; Campbell, Dodd and Brasher; 1983). Heard and lip-read speech are both temporally structured, that is, they represent signals occurring across time. In contrast, other visual codes such as writing or sign language are spatially organised structures. Campbell and Dodd argue that integration only occurs when both signals change in state over time. They showed (Campbell, Dodd and Brasher; 1983) that auditory-like recency was greater for moving as opposed to still hand signs. Further, the recall of lists of visually presented digits showed recency effects when the digits were revealed in a changing display over time (e.g Crowder, 1986). The necessity for a common metric for which to encode information calls for equivalent spatial or temporal characteristics between signals.

Cognitive architecture

The extent to which sensory information is combined may be dependent on the 'cognitive architecture' for speech processing; this argument may be expressed in terms of either the 'naturalness' or 'primacy' of the language code. Natural language codes are those which are historically evident. Spoken intercourse has presumably always allowed access to heard and lip-read perception, and in this sense, they are natural. Writing, sign and prosthetic communication are contrived forms of communication. On the basis of this argument hearing and lip-reading are combined and are unable

to be integrated with writing, signing or prosthetic communication. Cognitive architecture, however, may also be structured according to the primacy of the language code used. Normally, hearing and lip-reading are primary language codes; writing is learnt later in life. For some communities, signing may be a primary language code, and it is possible that in the future electro-cochlear and electro-tactile communication may be the primary language code for some individuals. Combination of speech information may occur between primary language sources, and not between primary and secondary language sources.

Shand and Klima (1981) argued that interaction effects in SOR were based on different perceptual mechanisms for storage and representation of primary (heard and seen speech in normally hearing populations) or secondary (written language) language codes. Investigating serial recall of American Sign Language signs, they found strong recency and suffix effects in recall by native signers. This suggests that the type of perceptual encoding may be dependent on early perceptual experience. However, this does not necessarily imply shared processing of sign representations with auditory or visual speech representations (Campbell, Dodd and Brasher, 1983).

## A MODEL OF MULTISENSORY INTEGRATION

The research into AV speech perception suggests a model of multisensory combination which explains the extent of bisensory advantage in terms of dimensions of similarity and complementarity between sensory sources.

It is necessary that the different sources of information are able to be converted into a common metric. This will be referred to as 'structural congruity'. This does not imply that sensory information should be derived from the same channel (auditory, visual, tactile), but rather that the structure of the signal and its temporal characteristics be able to be represented *at a given level of representation* in a single code. Auditory, visual and electro-tactile stimuli all represent articulatory information, which is temporally ordered. At feature, phoneme, and word levels, information is able to be represented in terms of the same metric: articulatory features, phonemes and words. In contrast, a common metric for heard speech and sign language is only available after lexical categorisation. As a theoretical construct, structural congruity would seem to be a necessary prior condition for integration of sensory information.

Given structural congruity, a second factor influencing multisensory integration is complementarity. When the task involves perception of structurally congruent information arising from different sensory sources, combination of information at a specific level will be more efficient if the inputs are complementary rather than redundant. Redundancy of structurally congruent information from different sensory sources may lead to interference, reducing the amount of information derived from either source.

The third principle, cognitive architecture, is the most tentative. It embraces two issues; naturalness and primacy. If integration only occurs for natural perceptual modes, then efficient combination of auditory and tactile, or visual and tactile, information will never be possible, given that only auditory and visual speech perception develop in the context of normal communication. If, however, it is only essential that a speech mode (which is structurally congruent) is developed as a native mode, that is, used from infancy, then integration may be possible. Evaluation of this hypothesis will rely on investigations of the use of prosthetic sensory devices by pre-linguistic infants.

## TACTILE SPEECH PERCEPTION

Tactile speech perception is defined as the encoding of spoken language by the use of touch. Two main approaches to the tactile representation of acoustic parameters exist; vibro-tactile encoding and electro-tactile encoding. Vibro-tactile encoding represents speech as a system of vibratory pulses, felt through either the sternum, the forehead or the forearm (Lynch, Oller and Eilers, 1989). Electro-tactile encoding represents speech information by the use of electrical pulses on the abdomen (Saunders, 1985) or the fingers (Blamey and Clark, 1985; Cowan, Alcantara, Blamey and Clark, 1988).

The University of Melbourne Tickle Talker uses a feature extraction approach to encode a number of dimensions of the speech signal on electrodes located on the sides of the fingers. Information from eight frequency bands represent acoustic features tactually; amplitude of the signal is coded as intensity of stimulation, fundamental frequency a rate of pulse transmission and second formant information as spatial location of stimulation across the eight electrodes used. Speech is perceived as a an array of rapid pulses changing over time and varying in place on the hand. Studies have indicated relatively good multisensory performance is able to be achieved using the tickle talker in combination with (residual) hearing and lip-reading (e.g Cowan et al, 1990). The following discussion considers tactile speech perception in combination with auditory and visual sensory information.

In terms of the model outlined above, auditory, visual and tactile sensory inputs represent three different *channels* of sensory input which are *structurally congruent*. The *redundancy* relations between the different modalities vary. Auditory and tactile stimuli are redundant at the feature level, given that they attempt to represent a variety of speech features. In contrast, (when the auditory signal is degraded) auditory/lip-read and tactile/lip-read combinations are complementary. It would be expected, therefore, that AV and visual-tactile combination would be superior to AT combination. Finally, lip-reading and hearing are *natural* codes for most listeners. Perception of electro-tactile stimulation is explicitly taught. It would be predicted that combination with tactile information would be poorer than AV combination.

A simple sum of commonalities between sensory sources would therefore predict AV combination of information to be better than AT or VT stimulation. Given that VT stimuli are informationally complementary, we might also expect VT combination to be superior to AT combination. The available data provide preliminary support for these predictions.

## AUDITORY-VISUAL AND AUDITORY-TACTILE COMBINATION

Alcantara (1991) trained two groups of subjects in AT speech perception, using degraded auditory stimuli. Two subjects received training in bisensory AT speech perception, and two subjects received training in auditory-alone and tactile-alone speech perception. He found that groups trained unisensorily (with separate auditory and tactile training) improved more than groups trained bisensorily (AT training) for vowel and consonant perception. In contrast, a study of AV bisensory and unisensory training using a similar methodology found no differences between groups trained bisensorily and unisensorily (Oerlemans and Blamey, in preparation).

Training is limited by the amount of perceptual information available. Subjects can lip-read a maximum of nine places of articulation (Summerfield, 1987), some manner contrasts and little voicing. Degraded auditory input tends to result in a loss of place information. In the case of complementary modalities, it is only in the combined situation that a relatively complete signal is able to be reconstructed. We would therefore expect unisensory training to result in poorer or equal post-training performance than bisensory training. In contrast, auditory and tactile information represent similar signals. There is, as a result, more advantage to unisensory training, given more scope for unisensory improvement and more opportunity for interference of one modality on the other in bisensory training.

Combination of AV information at the feature level appears to be better than combination of AT information. Using a mathematical model of multisensory combination derived from Blamey (1990), Alcantara found that a model of multisensory combination of information at the feature level was a poor predictor of AT performance; differences between observed and predicted scores were significant for a number of vowel and consonant features. In contrast, the earlier study of Blamey (1990) compared the model of combination of multisensory information at feature, phoneme and word levels for AV stimuli'. He found no difference between observed and predicted scores for feature level combination. Significant differences for models postulating phoneme and word level integration were found. These data suggest that while feature level integration of auditory and visual information occurs, integration for auditory and tactile information is less effective.

## AUDITORY-TACTILE AND VISUO-TACTILE COMBINATION

There is no direct evidence comparing integration of AT and VT information, however data presented by Blamey and Clark (1988) are suggestive. The authors compared the percentage of featural information transmitted by auditory, visual and tactile information alone or in combination. Using the combinatorial model reported earlier (Blamey, 1990), they compared observed feature scores with predicted feature scores. A summary of their results are shown in Table 1. It appears that observed vowel and vowel feature scores are closer to predicted scores for VT combinations than for AT combinations. This suggests that integration of information from complementary sources is better than integration of redundant sources.

| | Observed | | | Observed - predicted | | |
|---|---|---|---|---|---|---|
| | Auditory | Visual | Tactile | AV | AT | VT |
| ***Vowels*** | | | | | | |
| Total | 43 | 75 | 43 | 0 | -12 | -4 |
| Duration | 91 | 71 | 49 | 3 | 1 | -4 |
| F1 | 33 | 76 | 28 | 1 | -12 | -2 |
| F2 | 24 | 75 | 51 | 1 | -13 | -8 |
| **Consonants** | | | | | | |
| Total | 54 | 53 | 29 | 13 | 11 | 7 |
| Voicing | 79 | 7 | 11 | 11 | -3 | -2 |
| Nasality | 91 | 27 | 43 | 7 | 5 | -10 |
| Affric. | 66 | 70 | 29 | 5 | -6 | 8 |
| Duration | 38 | 58 | 80 | 20 | -14 | -11 |
| Place | 15 | 80 | 20 | 1 | -9 | -4 |
| Visib. | 24 | 100 | 12 | 0 | -4 | 0 |
| $A_0$ | 84 | 24 | 25 | 8 | -4 | -7 |
| High F2 | 27 | 74 | 59 | 3 | -18 | -6 |

Table 1: Percentage of information transmitted in closed set vowel and consonant tasks (Blamey and Clark, 1988)

The data for consonants, however, is not able to be described as simply. It is surprising that large positive differences are apparent between AV observed and predicted scores. Particularly, perception of consonant voicing and duration is much better than would be predicted by a feature integration model. Comparing tactile combinations, VT combination appears to be predicted by an integration model better than AT combination for consonant duration, place, visibility and F2 frequency. The reverse is true only for nasality and amplitude envelope information. The results of these studies generally support the model of integration outlined in this paper, and further investigation of the applicability of such a model would be worthwhile.

## CONCLUSIONS

Multisensory speech perception is superior to unisensory speech perception. More information is available when input is derived from two or more sources, than from a single source alone. The extent of multisensory advantage, however, is dependent on the degree to which information from these different sources may be combined. An analysis of AV perception suggests that some of the factors influencing this integration are structural congruity of the signals, redundancy or complementary between modalities, the spatial or temporal characteristics of the stimulus and the development of a cognitive architecture to support integration. On this basis, it appears unlikely that tactile information will be integrated with auditory or visual information to the extent observed in AV combination. It is expected, however, that tactile-visual perceptual combination should result in better performance than tactile-auditory perception. Development of devices for tactile speech perception by the hearing-impaired should consider that such devices may more effectively complement visual speech perception than residual or assisted hearing.

References

Alcantara, J.I. (1991). *The Effect of Training on Tactile Speech Perception.* (Unpublished thesis. University of Melbourne)

Blamey, P.J. (1990). Multimodal stimulation for speech perception. In M.J. Rowe and L.M. Aitkin (Eds.), *Information Processing in Mammalian Auditory and Tactile Systems.* New York: Wiley-Liss, 267-280

Blamey, P.J. and Clark, G.M (1985) *A wearable multiple-electrode electrotactile speech processor for the profoundly deaf,* J. Acoust. Soc. Am. 77, 1619-1620.

Blamey, P.J. and Clark, G.M (1988) . Paper presented to SST-88.

Campbell, R. and Dodd, B. (1980) *Hearing by eye,* Quart. J. Exp. Psych. 32, 85-99.

Campbell, R., Dodd, B., and Brasher, (1983) *The sources of visual recency,* Quart. J. Exp. Psych. 35a, 581-587.

Cowan, R.S.C., Alcantara, J.I., Blamey, P.J., and Clark, G.M. (1988) *Preliminary evaluation of a multichannel electrotactile speech processor* J. Acoust. Soc. Am. 83(6), 2328-2338.

Cowan, R.S.C., Blamey, P.J., Galvin, K.L., Sarant, J.Z., Alcantara, J.I., and Clark, G.M. (1990) *Perception of sentences, words, and speech features by profoundly hearing-impaired children using a multichannel electrotactile speech processor* J. Acoust. Soc. Am. 88(3), 1374-1384.

Crowder, R.G. (1986) *Auditory and temporal factors in the modality effect* J. Exp. Psych.: Learn., Mem. Cog., 12, 269-279.

Fodor, J.A. (1983). *The Modularity of Mind* (MIT Press:Cambridge).

Gathercole, S.E. (1987) Lip-reading: Implications for theories of short-term memory In B. Dodd and R. Campbell (Eds.), *Hearing by Eye: The psychology of lip-reading,* pp.228-2440 (Erlbaum:London).

Jeffers, J. and Barley, M. (1971). *Speechreading (Lip-reading)* (Charles C. Thomas.:Springfield, IL).

Kimura, D. (1973). *The asymmetry of the human brain* Scientific Am. 228, 70-78.

Lynch, M.P., Oller, D.K., and Eilers, R.E. (1989) *Portable tactile aids for speech perception* Volta Review Monograph, 91(5), 113-126.

Massaro, D.W. (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Erlbaum:Hillsdale, NJ).

McGurk, H. and MacDonald, J. (1976) *Hearing lips and seeing voices,* Nature, 265, 746-748.

Oerlemans, M.J. and Blamey, P.J. (1994) *Auditory-visual training of speech perception,* In preparation.

Osberger, M.J., Rines-Weiss, D., Kalberer, A. (1986) *Speech tracking performance of deaf adults using a vibrotactile aid* Conv. Am. Speech-Lang. Hear. Ass. November, 1986, Detroit.

Shand, M.A. and Klima, E.S. (1981) *Nonauditory suffix effects in congenitally deaf signers of ASL* J. Exp. Psych.: Human Learn. Mem. 7, 464-474.

Summerfield, Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception In B. Dodd and R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-reading* pp.3-51 (Erlbaum: London).