Visual Cues in Speech Perception

Mustapha Skhiri
Department of Computer and Information Science, Linköping University.
Swedish National Graduate School of Language Technology.

Abstract

This report presents a summarised view of some research papers on visual cues in speech perception, especially the work of Harry McGurk and the McGurk effect. The report does not present any research done by the author, nor does it pretend to cover the whole research area. It is merely a summary of some well-known science reports in this area. The report includes a broader chapter on research done with consideration to the McGurk effect, i.e. research on speech perception done within the last twenty years. Some of the articles that are summarised in the report were presented quite recently and therefore cover some of the most recent studies done in speech perception.

1 Intorduction

Background

After reading some articles and watching some film-clips about speech perception, I got somewhat interested in the area of visual influence on speech perception. This eventually led me to choose the topic visual cues in speech perception for the term paper.

Purpose

The document is produced in order to fulfil one of the obligatory moments in Speech Technology1 course. In this term paper I'm supposed to study some field in speech science a bit more thorough and gain knowledge about a specific topic.

Disposition

I start with what I call setting the stage, there i address the background of Harry McGurk and early research in the auditorial-visual speech area. Then I present the research that led to the effect nowadays known as the McGurk effect. I wrap it up with a summary of more resent research articles, based upon or done with consideration of the McGurk effect. The final chapter contains a summary of the research presented in this paper. All sources used in order to produce this paper will be mentioned in the reference chapter at the very end of the report.

2 Setting the Stage

The saying - Read my lips - is used to implicate the truth or seriousness of the speaker. Why is that? Probably it means something like: if you can't believe your ears, believe your eyes. This statement clearly implies that a great deal of spoken information is transmitted optically. Is this a fact?

It is well known that lip-reading is necessary for hearing impaired to understand speech. But as early as 1935, Cotton stated "there is an important element of visual hearing in all normal individuals". Even if the auditory modality is the most important for speech perception by normal hearers, the visual modality may allow subjects to better understand speech. Note that visual information cannot in itself provide normal speech intelligibility.

Well, humans produce natural speech through the actions of several articulators e.g. lips tongue and jaw. Of these articulators only some are visible, but natural speech is not continuously audible either. It is made up by both sounds and significant parts of silence, during these silences the speaker gestures in order to anticipate the following sound. To sum it up, some parts of speech movements are only visible, other parts are only audible, and still other parts are not only audible, but also visible.

The Bimodality of Speech

Risberg and Lubcker observed in 1978 that, subjects who only were provided with visual modality of a speaker, perceived only 1% of the words. On the

other hand, subjects that only were provided with the low-pass filtered version of the speech sound got about 6 % of the words right. When presented with the bimodal information of the speech, the subjects got 45 % of the words right. This observation exemplifies the synergy of the modes of speech perception.

Harry McGurk's observations

Human development was the primary field of psychological research for Harry McGurk, a senior developmental psychologist at the university of Surrey in England. He conducted infant perception studies back in the early '70s. His research were to great extent in conflict with the, at the time, authority in this area, Tom Bower. He refuted the research of Aronson and Rosenbloom, claiming intermodal spatial dislocation, and instead went on to investigate the unity of the senses hypothesis by way of intermodal conflict.

As documented in his research paper from '78, McGurk took videotapes of productions of /ba/ and /ga/ and had them dubbed to produce not just the matching ba-voice/ba-lips, and a ga-voice/ga-lips sequence. What is not documented, is his displeasure when the videotapes returned from the auditory-visual centre at Surrey. Harry was confused until he realized that the perceived "da" resulted from a quirk in human perception and not an error on the technician's part. He realised that he was experiencing something quite remarkable.

3 The McGurk Effect

Most studies of the role of visual cues in speech perception from before 1976 have been either addressing vision as an alternative mode to hearing in speech perception, or investigating the compensatory or complementary role of vision upon the auditory perception of speech under conditions of noise. McGurk and his research assistent MacDonald supplied revolutionary new evidence in two articles in the late seventies that the assumption that speech processing is a unimodal process does not hold. An overview of these experiments is given in this section.

The first of these experiment by McGurk and MacDonald (1976) demonstrated a previously unrecognized influence of visual information on speech perception. They tested pre-school children, primary school children and adults by asking what they perceived from a video demonstration of a young woman's talking head, in which repeated utterances of the syllable [ba] had been dubbed on to lip movements for [ga], vice versa and with the syllables [pa] and [ka]. The answers the subjects gave were classifiable into five groups. An example for an auditory ga-ga and visual ba-ba and the combination ka-ka and pa-pa:

Stimuli		Response Categories				
Auditory component	*Visual component*	*Auditory*	*Visual*	*Fused*	*Combination*	*Other*
ga-ga	ba-ba	ga-ga	ba-ba	da-da	gabga	gagla dabda etc
pa-pa	ka-ka	pa-pa	ka-ka	ta-ta	-	tapa pta kafta etc.

Auditory and *visual* responses may speak for themselves. A *fused* response is one where information from the two modalities is transformed into something new with an element not presented in either modality, whereas a *combination* response contains relatively unmodified elements of both modalities. Responses that would not fit into these four categories are classified as *other*.

The results of the experiment were astounding. For the four combinations tested the following answers were given: For the auditory ba-ba visual ga-ga, 98% of all adults and 64 to 81% of all children reported the fused response da-da! The rest reported auditory responses. For the stimulus the other way around, auditory ga-ga visual ba-ba, resulted in 54% combination responses with adults, 57% auditory responses with the youngest children and equally spread responses from the other group of children. With the combination auditory pa-pa visual ka-ka there was a strong McGurk effect again, with 81% fused ta-ta responses. The complement experiment resulted again in mostly auditory responses with children and combination responses with adult subjects.

So these results are generalizable to stop consonants, resulting in fused responses. The results also illustrate that the auditory perception of adults is more influenced by visual input than is that of children. Also, where responses are mainly in a single modality, this tends to be the auditory for children and the visual for adults.

At the moment of appearance of the article reporting these findings, the auditory-based theories of speech perception were unable to explain these new revolutionary observations, in which a strong role for visual information in speech processing is proven. Other experiments show that in ba-voice/ga-lips presentation, there is visual information for [ga] and [da] and auditory information with features common to [da] and [ba]. This experiment shows that a choice in perception is made for the unifying percept [da]. The same holds for a [ta] response to pa-voice/ka-lips presentations. Looking at ga-voice/ba-lips and ka-voice/pa-lips presentations, the modalities provide no overlapping evidence for a certain response, resulting in combined responses

because the listener has no way of deciding between the two sources of information.

Binnie et al. (1974) found that lip movements of da, ga, ta and ka as well as lip movements of ba, pa and ma are visually difficultly discriminable from each other . Labial and non-labial consonants were easily discriminable. So frontal (labial) place of articulation is visually well distinguishable from middle and back while these last two are hard to discriminate from each other. Another finding was that the feature of place of articulation is more efficiently detected by vision than the feature of manner of articulation, while voicing and nasality of consonantal utterances are readily perceived auditorially.

These conclusions led MacDonald and McGurk (1978) to the manner-place hypothesis to explain the observations from 1976, described above. This hypothesis claims that manner of articulation of consonantal utterances is detected auditorially (for example whether the utterance is voiced or voiceless, oral or nasal, stopped or continuant, etc.), while place of articulation is detected visually. The hypothesis argues that at a certain point in processing, information from both modalities results in perception of a best-fit solution.

To test this manner-place hypothesis MacDonald and McGurk (1978) performed another experiment very similar to the first one, but now with every combination of labial consonants p, b and m and non-labials t, d, k, g and n. In the combination of presenting auditory and visually non-labial utterances, in almost 100% of the cases, the auditory stimulus was perceived. This result is in accordance with the manner-place hypothesis, which predicts that these lip movements shouldn't create an illusionary effect, because these non-labials are visually virtually interchangeable (see Binnies results above). The combination of presenting auditory and visually labial utterances also averages to a 96% perception of the auditory stimulus, which is again in harmony with the hypothesis stated.

But of course the other two combinations are equally if not more important to the testing of the hypothesis. First, the labials were presented as sound combined with the non-labial lip movements. Auditory ba mostly led to perception of da, which was strongest with the non-labial ta presented visually. In only 16% of cases the correct answer, the auditory stimulus ba was perceived. An auditory pa sound led, with a very small majority, to perception of pa, together with various fused responses. With the given non-labials presented visually, an auditory ma led in almost all cases to perception of na. Here in only 9% of cases ma was actually perceived.

With the non-labials presented auditory combined with visual labial utterances, the McGurk effect was less strong. An average of three quarters of the stimuli led to correct responses, i.e. perception of the auditory stimulus.

Thus, these observations clearly illustrate a general effect of vision upon speech perception in face-to-face situations. The manner-place hypothesis certainly holds for labial-voice/nonlabial-lips utterances, although it is less applicable to the nonlabial-voice/labial-lips situations.

None of the thusfar developed speech perception theories take the role of visual stimuli into account. Is it then possible to fit these findings into the

existing theories? There are two main streams in these theories, namely passive and active models of speech perception. A passive one based on the notion of automatically registering feature detectors assumes that individual cortical detector cells respond to complex multidimensional features of a specific acoustic waveform. But how could the same waveform cause different detectors to fire in case of different waveform-visual stimulus combinations? So this theory is certainly not enough to explain how speech perception takes place, notwithstanding that these feature detectors could well exist.

An 'active' model is the motor theory, which roughly states that the neural signals produced by the sensory input are associated with internally generated neural commands which would have constructed the articulatory gestures leading to that sensory input. Speech perception and production are thus closely related. The analysis by synthesis model is highly similar, but adds that some phonetic information is directly decoded via acoustic features, allowing a hypothesis about the whole acoustic signal to be generated and tested. Because both active variants already suppose intermodal combination (muscle coordination and audition), MacDonald and McGurk propose that either variant could have room for the role vision plays in speech perception; that the generation of the internally generated articulatory gesture neural commands, the production mediating the perception, is influenced by information visually available from watching the speaker.

4 Consequences of McGurk effect

In this chapter I present some research about Visual cues in speech perception done with consideration to the McGurk effect. All studies, which have been done on Visual Cues, since Harry McGurk discovered the McGurk effect, have taken the McGurk effect into consideration. I present some well-known articles rather briefly in order to give the reader a feeling for this research area, and hopefully the readers' curiosity will inspire them to read the articles themselves.

Evaluation and Integration of Visual and Auditory Information in Speech Perception (*by Dominic W.Massaro and Michael M.Cohen).*
They performed three experiments in order to investigate the evaluation and integration of visual and auditorial information in speech perception.

In the first two experiments, subjects identified /ba/ or /da/ speech events consisting of high-quality synthetic syllables ranging from /ba/ to /da/ combined with a videotaped /ba/ or /da/ or neutral articulation. Although subjects were specifically instructed to report what they heard, visual articulation made a large contribution to identification. The tests of quantitative models provide evidence for the integration of continuous and independent, sources of information. The reaction times for identification were primarily correlated with the perceived ambiguity of the speech event. In the third experiment, the speech events were identified with an unconstrained set of response alternatives. In addition to /ba/ and /da/ responses, the /bda/ and /tha/ responses were well described by a combination of continuous and independent features. This body of results provides strong evidence for a fuzzy logical model of perceptual recognition.

Illusions and issues in Bimodal Speech Perception *(by Dominic W. Massar)*
" The addition of just one more modality has enabled the discovery of several new phenomena, new theoretical endeavours, and a closer link between research and application. The goal of the paper is to review a series of relevant issues in the authers search for an understanding of speech perception by ear and by eye. The issues include a discussion of viable explanations of the McGurk effect, the time course of auditory/visual processing, neural processing, the role of dynamic information, the information in visual speech, the fusion of written language and auditory speech, and the issue of generalising from studies of syllables to words and larger segments".

The Effect of Tonal Information on Auditory Reliance in the McGurk Effect *(by Dennis Burnham and Susanna Lau).*
They test if the incidence of the McGurk effect is greater for speakers of English than for Japanese, and in turn for speakers of Japanese than Cantonese, as Sekiyama have concluded. Sekiyama postulates that this is because speakers of tonal languages rely more upon auditory than visual information in speech perception.

The testing is performed in the following way, subject are presented with both tonal (Cantonese) and non-tonal (English) language speakers with McGurk stimuli in which the tone on syllables. The most interesting result were that, apparently /ba/ is more visually distinct than /ga/. However it could just be that the variability on /ba/ actually accentuates visual information for /ba/, but that on the visually less distinct /ga/ the tonal variation is not so salient.

The Influence of Quality of Information on the McGurk Effect *(by Eric Fixmer and Sarah Hawkins).*
Recent studies by Fixmer and Hawkins (1998) investigated the influence of quality of information on the McGurk effect. They found that extraneous noise changes the reliance subjects place on auditory versus visual sensation. The proportion of McGurk responses decreases when the extraneous noise is visual, whereas it increases when the noise is auditory. Still these changes in McGurk responses arise directly from changes in the quality of the physical stimuli themselves or through some decision process that is mediated by conscious awareness of the extraneous noise. On one hand, because the subjects didn't know of differences in intelligibility of the spoken syllables in their experiment, it seems reasonable to suggest that, in this case, the quality of the physical stimuli affects the final sensory percepts rather than conscious judgements about their quality. On the other hand, subjects were presumably aware of the degrading of visual or auditory information by extraneous noise, so the number of McGurk responses could be affected by provoking subjects to change a decision criterion about the relative reliability of the two modalities rather than by affecting the sensory percept itself. Current research has not investigated that yet. Fixmer and Hawkins' data are in accordance with models that claim that speech perception involves activation and decay in stimulation of perceptual units, and that perceivers of speech use all available sources of information, proportional to their informational load. This involves again, as explained above, the physical quality, relative quality to other sources and the belief of quality the

perceiver has in these sources. Other knowledge-based and decision making sources and processes also contribute to the final percept. Further investigations are needed to discover the specific influences of these factors.

Gender Incongruity and the McGurk Effect *(by Kerry P.Green,Patricia K.Kuhl, Andrew N.Meltzoff and Erica.B.Stevens))*
The purpose of the research conducted by Green, Kuhl, Meltzoff and Stevens (1991) was to manipulate the cognitive congruence between the auditory and visual signals. They introduced a gender discrepancy, a male talker's voice with a female talker's face, and vice versa. Their experiments show that the McGurk effect is not influenced by this gender discrepancy; the effect is equally strong, regardless of whether face-voice stimuli are gender-compatible or gender-incompatible. Also, the magnitude of the McGurk effect is not influenced by reduction in the congruency between the auditory and visual signals. They conclude that auditory and visual phonetic information are integrated after the auditory signal has been normalized with respect to talker differences, so the auditory information is already talker-neutral at the time of integration of the phonetic information. Another interesting fact in the research of Green, Kuhl and Meltzoff was that an i vowel context produced the strongest McGurk effect, an a context produced a moderate effect and an u context almost none.

5 Summary

The McGurk effect occurs when discrepant auditory and visual information results in an emergent percept, e.g. auditory [ba] and visual [ga] result in the perception of [da]. This discovery has had a huge impact on the view at that time existing theories of speech perception, because these theories saw visual information only as complementary or alternative to the auditory information. McGurk and MacDonald's papers were a big eye-opener on the importance of visual information and since then a lot of research has been conducted on visual cues in speech perception. Theories have been retested, modified and created, and a whole new area of research has emerged.

6 References

Benoit, C. (1992) *The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking Faces.* Journal on Communications, Budapest, Hungary, Sept. 1992.

Burnham, D. (1998) *Harry McGurk and the McGurk Effect*. In Burnham, D., Robert-Ribes, J. and Vatikiotis-Bateson, E. (eds.), Proceedings of the international conference on auditory-visual speech processing, p. 1-2.

McGurk, H. and MacDonald, J. (1976). *Hearing lips and seeing voices.* Nature, 264, p. 746-748.

MacDonald, J. and McGurk, H. (1978). *Visual influences on speech perception processes.* Perception & Psychophysics, 24 (3), p. 253-257.

Massaro, D. W. & Cohen, M. M. (1983) *Evaluation and Integration of Visual and Auditorial Information in Speech Perception*. Journal of Experimental Psychology: Human Perception and Performance, 9 (5) p. 753-771.

Massaro, D. W. (1998) *Illusions and Issues in Bimodal Speech Perception*. In Burnham, D., Robert-Ribes, J. and Vatikiotis-Bateson, E. (eds.), Proceedings of the international conference on auditory-visual speech processing, p. 21-26.

Fixmer, E. and Hawkins, S. (1998). *The influence of quality of information on the McGurk effect*. In Burnham, D., Robert-Ribes, J. and Vatikiotis-Bateson, E. (eds.), Proceedings of the international conference on auditory-visual speech processing, p. 27-32.

Green, K.P., Kuhl, P.K., Meltzoff, A.N, and Stevens, E.B (1991). *Integrating speech information across talkers, gender and sensory modality: Female faces and male voices in the McGurk effect*. Perception & Psychophysics, 50 (6), p. 524-536.