ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal - Sydney, Australia
December 4-6, 1998

# INTEGRATION OF DICHOTICALLY AND VISUALLY PRESENTED SPEECH STIMULI

*Mikko Sams and Sari Rusanen*

*Laboratory of Computational Engineering, Helsinki University of Technology, P.O. BOX 9400, FIN-02015 HUT, Finland*

## ABSTRACT

In dichotic listening, two competing messages are delivered to the left and right ear, respectively. Right-handed subjects tend to report hearing more frequently the message input to the right ear. This is called Right Ear Advantage (REA). When intensities and other properties of the messages are properly adjusted, subjects may have a single perception localized to the center of the head. Frequently the two stimuli 'fuse' and the resultant perception corresponds to neither of the presented stimuli. In a similar vein, in audiovisual speech perception discordant auditory and visual components of the stimulus may fuse (the McGurk effect). In the present study, we investigated the relationship between dichotic listening and audiovisual speech perception.

Our results demonstrated REA and dominance of acoustical /ta/ stimulus. The influence of visual speech on the perception was clearly stronger than REA. When visual information was concordant with the auditory input to one ear, the perception of that syllable increased strongly, irrespective of the ear of stimulation. Interestingly, REA appeared even when visible speech modified perception.

## 1. INTRODUCTION

In dichotic listening, two competing messages are delivered to the left and right ear, respectively. Right-handed subjects report hearing more frequently the message input to the right ear. This Right Ear Advantage (REA) has been suggested to be due to left hemisphere dominance for language. Simple synthetic or natural consonant-vowel or consonant-vowel-consonant syllables have been used as stimuli in dichotic studies [1]. This allows accurate timing of the two inputs as well as manipulation of the component features of the stimuli. If the intensities and other properties of the messages are properly adjusted, subjects have a single perception that is localized to the center of head. Quite frequently the two syllables 'fuse' and the resultant perception corresponds to neither of the presented stimuli. For example, when the dichotic syllables are /ba/ and /ga/, many subjects report hearing /da/ [2]. When the subject reports hearing the right-ear stimulus, actually he/she might base the perception on an integrated representation in which the features of the right-ear stimulus are dominant.

In face-to-face communication, seeing the articulating face provides important additional information for speech perception. Visual cues are especially important when part of the acoustical message is masked by noise [3,4] but can affect the speech perception even under optimal listening conditions. When the acoustical syllable /ba/ is dubbed to the visual articulation of /ga/ English-speaking subjects usually perceive /da/, less frequently /ga/ but very infrequently the acoustical /ba/. This "McGurk effect"[5] has been usually studied using meaningless syllables but it occurs also when real words are used as stimuli [6,7].

To illuminate the relationship between dichotic listening and audiovisual integration we presented syllables via three channels (left ear, right ear, eyes). We hypthetized that if REA mainly is explained by anatomical connections, it should be apparent even when an additional information channel is used. On the other hand, when visual speech is available, it most probably very strongly catches subject's attention and decreases the possible tendency of the subjects to pay attention to the right perceptual filed, which might explain REA.

## 2. METHODS

The voluntary subjects were 46 right-handed Finnish male students.

The visual stimuli consisted of a single /pa/, /ka/ and /ta/ syllables articulated by a female talker. They began and ended with talker's face in neutral expression with closed lips. Still facial image of the same talker was used in control stimuli.

Auditory stimuli consisted of the same three syllables as in visual tokens and were produced by the same talker. The solitary sounds were paired to produce the nine dichotic combinations: /pa-pa/, /pa-ka/, /pa-ta/, /ka-ka/, /ka-pa/, /ka-ta/, /ta-ta/, /ta-pa/, /ta-ka/ (/left ear– right ear/).

The four visual and nine auditory tokens were paired to create 36 different audiovisual stimuli which were presented in random order. Each visual token started with a 160 ms fade up from grey background to the face and ended with a similar-length fade away. Black, clearly identifiable numbers on grey videoscreen were presented for 1.5 s before the fade up started. Location of the number on the screen corresponded the location of the talker's mouth. The constant interval between the trials was 7.5 s. Two stimulus sequences with different randomizations were constructed. They were output on a videotape for presentation. Intensity of the auditory stimuli was about 50 db SPL.

Subjects responded to each stimulus by choosing the alternative which best matched their perception of the stimulus. For each item there were nine possible response alternatives: pa, ka, ta, pka, pta, kpa, kta, tpa, tka. For data analysis, four response classes were formed: 1) pa, 2) ta, 3) combinations (pta and tpa) and 4) others (ka, kta, tka, kpa, pka).

## 3. RESULTS

Both diotic /pa-pa/ and /ta-ta/ auditory stimuli were well identified even without supporting visual speech. However, the auditory identification of /ka-ka/ syllable was much inferior and was perceived as /pa/ by 38% of the subjects. Identification was so much inferior to the two other diotic syllables that it was not included in the data analyses. The following data analyses were based on the first stimulus sequence.

When acoustical /pa/ and /ta/ were presented together with a still face, more of the subjects perceived /ta/ than /pa/ irresepective of the ear of presentation (Fig. 1). The proportion of /ta/ identifications was statistically significantly higher when it was presented to the right ear [$Q(3,42 = 4.77$, $p < 0.05$]. In a similar vein, the proportion of /pa/ identifications was higher when /pa/ syllable was presented to the right ear, but the increase of 7% was not statistically significant. These results demonstrate two effects: REA and dominance of acoustical /ta/ stimulus. When the effect of stimulus dominance was taken into account, the REA for stimuli presented with still faces was 14%. When /pa/ was presented to the right ear, the number of combination perceptions was 14% - but not significantly - higher than when it was presented to the left ear.

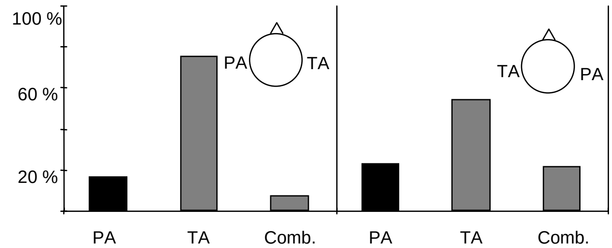


**Figure 1:** The proportions of different identifications, indicated below x-axis, when /pa/ and /ta/ syllables where presented dichotically together with the still face.

Visual /ta/ shown together with dichotically presented /pa/ and /ta/ practically abolished the /pa/ perceptions (Fig. 2) and clearly increased the proportion of /ta/ perceptions, the increase being relatively larger when acoustical /ta/ was presented to the left ear (21%). In comparison to still-face presentation, the REA for /ta/ decreased from 21% to 11%. Note also that there were slightly more combination perceptions when /pa/ was presented to the right ear.

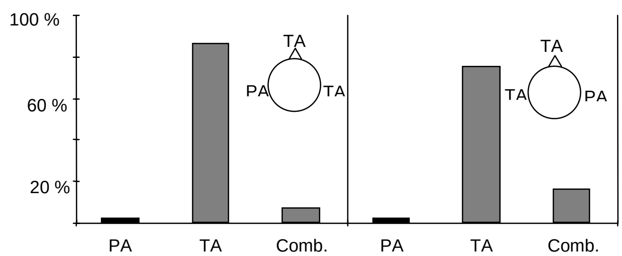


**Figure 2:** The proportions of different identifications, indicated below x-axis, when /pa/ and /ta/ syllables where presented dichotically and the subjects saw the talker to articulate /ta/.

When the visual /pa/ syllable was combined with the dichotically presented /pa/ and /ta/, the pattern of results was rather different (Fig. 3). The number of /ta/ identifications decreased substantially but remained at a rather high level, consistent with the dominance of /ta/ syllable. In comparison to the still-face condition, the proportion of /pa/ identifications increased from 16% to 40% (/pa/ to the left) and from 23% to 53% (/pa/ to the right). This increased the REA for /pa/ from 7% to 13%.

Discordant visual information influenced perception even when the same syllable was input to both ears, producing the McGurk effect. When diotic /pa-pa/ was presented together with visual /ta/, the proportion of /pa/ identifications dropped to 28%. The proportion of /ka/ or /ta/ identifications was 42% and that of combinations 30%. When diotic /ta/ was presented together with visual /pa/, the proportion of /ta/ identifications was 60%, that of combinations 33% and that of /pa/ 7%.

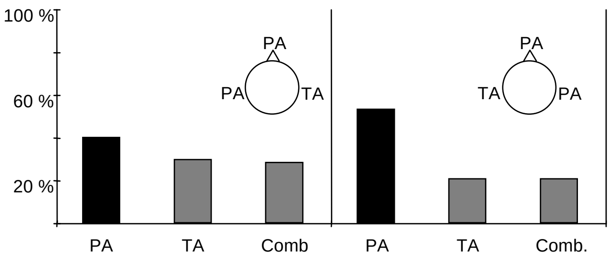


**Figure 3:** The proportions of different identifications, indicated below x-axis, when /pa/ and /ta/ syllables where presented dichotically and the subjects saw the talker to articulate /pa/.

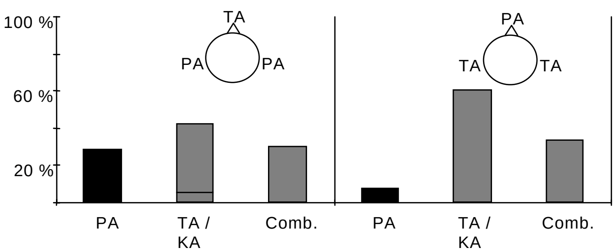


**Figure 4:** The proportions of different identifications, indicated below x-axis, when diotic /pa-pa/ and /ta-ta/ syllables where presented together with the discordant visual syllable /ta/ or /pa/, respectively.

## 4. DISCUSSION

The present results demonstrated REA for dichotically presented /pa/ and /ta/ syllables. Visual speech also influenced perception. When visual information was concordant with the auditory input to one ear, the perception of that syllable increased strongly, irrespective of the ear of stimulation. When the auditory syllable was presented diotically, discordant visual syllable strongly decreased the perception of the auditory stimuli.

It is interesting that the dichotic /ta-pa/ syllable produced more combination responses than /pa-ta/ stimulus. We tentatively suggest that the perception of dichotic stimuli is based on a trace where the features of the inputs to the two ears are integrated. Because of anatomical reasons, features of the right-ear stimulus contribute slightly more to this trace. With the present stimuli, the features of /ta/ syllable dominate the trace, especially so when it is presented to the right ear. However, when /pa/ is presented to the right ear, the relative amount of /pa/ features increase. In addition to the slight increase in /pa/ perceptions, this is also reflected in the increase of combination perceptions. Therefore, in calculating the amount of REA, the number of combination responses perhaps should also be considered.

Kimura [8] has suggested that REA is a result of stronger contralateral than ipsilateral connections to the auditory cortex. When two different stimuli are simultaneously presented to the left and right ears, the contralateral signal inhibits the ipsilateral one. Right-ear input projects stronger to the language-specific left hemisphere, and therefore dominates the perception. Left-ear input, projecting dominantly to the right hemisphere, has access to the left hemisphere via corpus callosum. In concordance with this view, the neuromagnetic response peaking at about 100 ms (M100) is larger and of shorter latency to contralateral than ipsilateral tones in both hemispheres [9,10]. In the auditory cortex there are neurons, probably getting callosal input, that show summation responses to binaural stimuli [11]. Population of such neurons might underlie the trace used in perceiving dichotic stimuli. Problems in phoneme discrimination and identification, sensory aphasia, that may result from a damage to the secondary auditory areas of the left hemisphere support the idea that the trace is located in these cortical areas.

When visual syllable was concordant with one of the dichotic stimuli, the perception of that syllable was enhanced. This effect is rather similar to the beneficial effect of visual speech when auditory stimuli are presented in noise. Discordant input to the other ear may be regarded as masking noise that blurs the neural representation of the stimulus. One possibility to explain the increase in the perception of those syllables, which are supported by visual stimulation, is to assume that there is a common representation for the auditory and visual speech. After the inputs from the two ears are combined, this information would be fed forward to such an audiovisual trace, where the visual input can have its influence.

Our previous magnetoencephalographic (MEG) studies [12,13,14] have demonstrated a specific brain response related to audiovisual speech integration. It is generated by concordant as well as discordant audiovisual stimuli. When the source of this response was estimated, it turned out to be generated at or in close vicinity of the supratemporal auditory cortex. This accords with the perceptual/phonetical nature of the integration and with the perceptual ‘auditoriness’ of the result of the McGurk integration. The integration response tended to be more prominent over the right than over the left hemisphere.

REA appeared even when visible speech modified perception. When visual stimulus was /ta/, the amount of REA for /ta/ was 11% and there was not REA for /pa/. On the other hand, when visual stimulus was /pa/ the amount of REA for /pa/ was 13% and even for /ta/ 9%. The persistence of REA supports the idea that it is due to hard-wired neural connections. Attention has also been shown to affect REA and there is evidence of inherent tendency of humans to attend to the right ear [15]. However, in the present experiment the subjects most probably attended to the lips of the talker, not to one of the ears.

When diotic /ta/ was presented together with visual /pa/, the amount of auditory perceptions decreased from 92% to 60%. This was mainly due to the increase of combination perceptions to 33%, an expected result when auditory non-bilabial and visual bilabial are presented together. The large number of auditory /ta/ perceptions is consistent with the dominance of /ta/ syllable when the auditory stimuli were presented dichotically with a still face. When visual /ta/ was presented together with diotic /pa/, the amount of auditory perceptions decreased from 92% to 28%. The relatively high amount of combination perceptions was not expected, because the auditory bilabials combined with visual nonbilabials typically produce fusion perceptions.

The integration effects obtained using dichotic presentation of auditory stimuli and presenting discordant auditory and visual stimuli resemble each other. For example, both may result in a unified auditory speech perception and the resulting perception can be a blend (/apa/ + /aka/ = /ata/) or a combination. There are, however, also differences. The integration window for audiovisual stimuli is some hundreds of milliseconds but for dichotic stimuli much shorter. The two types of integration even might happen in different cortical areas. Therefore, the neurocognitive mechanisms of dichotic and audiovisual speech integration are most probably different.

## 5. REFERENCES


1. Studderd-Kennedy, M. and Schankweiler, D. "Hemispheric specialization for speech perception," *J. Acoust. Soc. Am.* 48: 579-594, 1970.
2. Cutting, J.E. "Two left-hemisphere mechanisms in speech perception," *Percept. Pscyhophy.* 16: 601-612, 1974.
3. Sumby, W. and Pollack, I. "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.* 26: 212-215, 1954.
4. O'Neill, J.J. "Contributions of the visual components of oral symbols to speech comprehension," *J. Speech Hear. Disord.* 19: 429-439, 1954.
5. McGurk, H. and MacDonald, J. "Hearing lips and seeing voices," *Nature* 264: 746-748, 1976.
6. Dekle, D.J., Fowler, C.A. and Funnell, M.G. "Audiovisual integration in perception of real words," *Percept. Psychophys.* 51: 355-362, 1992.
7. Sams, M., Manninen, P., Surakka, V., Helin, P., and Kättö, R. “McGurk effect in Finnish syllables, isolated words, and words in sentences: effects of word meaning and sentence context”, Speech Communication, 1988, in press.
8. Kimura, D. “Functional asymmetry of the brain in dichotic listening”, *Cortex* 3: 163-178, 1967.
9. Reite, M., Zimmerman, J.T., Zimmerman, J.E. “Magnetic auditory evoked fields: response amplitude vs. stimulus intensity”, *Electroenceph. Clin. Neurophysiol*. 51: 388-391, 1981.
10. Mäkelä, J., Ahonen, A., Hämäläinen, M., Hari, R., Ilmoniemi, R., Kajola, M., Knuutila, J., Lounasmaa, O.V., McEvoy, L., Salmelin, R., Salonen, O., Sams, M., Simola, J., Tesche, C. and Vasama, J.-P. "Functional differences between auditory cortices of the two hemispheres revealed by whole-head neuromagentic recordings," *Human Brain Mapping* 1: 48-56, 1993.
11. Middlebrooks, J.C. and Pettigrew, J.D. “Functional classes of neurons in primary auditory cortex of the cat distinguished by sensitivity to sound localization”, *J. Neurosci.* 1: 107-120, 1981.
12. Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.-T. and Simola, J. "Seeing speech: visual information from lip movements modifies activity in the human auditory cortex," *Neurosci. Lett.* 127: 141-145, 1991.
13. Sams, M. and Levänen, S. “Where and when are the heard and seen speech integrated: magneto-encephalographic (MEG) studies”. In D. Stork and M. Hennecke (Eds.), *Speechreading by humans and machines*, Springer, Berlin, 1996, 233-238.
14. Sams, M. and Levänen, S., A neuromagnetic study of the integration of audiovisual speech in the brain. In Y. Koga, K.Nagata and K. Hirata (Eds.), *Brain Topography Today*, Elsevier, Amsterdam, 1998, 47-53.
15. Mondor, T.A. and Bryden, M.P. “The influence of attention on the dichotic REA”, *Neuropsychol.* 29: 1179-1190, 1991.