AUDIO AND AUDIO-VISUAL PERCEPTION OF CONSONANTS DISTURBED BY WHITE NOISE AND 'COCKTAIL PARTY'

László Czap

University of Miskolc, Department of Automation

H 3515 Miskolc, Egyetemváros
E-mail: czap@malacka.iit.uni-miskolc.hu

ABSTRACT

Some research questions regarding the speech perception can only be answered with natural speech stimuli especially in noisy environment. In this paper we are going to answer a couple of questions concerning the visual support of audio signal at speech recognition. How much support the video signal can give to the audio one? The impact of nature of the noise. How can the visual information help to identify the place of articulation? Do the voices of different classes of excitation get the same visual support? In order to answer these questions we have performed intelligibility study on consonants between the same vowel supported or not by the speaker's image with different signal to noise ratios. The noise is either white noise or a mix of other speakers' voice.

1. INTRODUCTION

It is well known that visual information obtained by speechreading and interpretation of body gestures improves perception of speech, especially in noisy environment. The visual information is even more important to persons with a hearing loss. Probably there is some relationship between the performance of a human recogniser disabled by noise and that of a machine with limited capability.

To understand bimodal recognition the first stage is to perform intelligibility tests for quantification of the information transmitted by the visual channel. This work has not been carried out yet for Hungarian visemes. The first stage of our bimodal recognition research aims at getting information on the visual support of different consonants.

2. METHOD

In the test series the subjects were university students without prior phonetical study. They were asked to listen to VCV words with a consonant between the same vowels. (e.g. ete, ama) twice, then they wrote down the consonant. They had limited time for the answer (appr.2 seconds). They were listening to the noisy voice of a series of 23 words each containing one consonant and then to the same audio signal supported by the speaker's image. They watched the image on the same TV monitor and listened to the voice from a loudspeaker. The momentary signal to noise ratio was fixed in every 5 milliseconds to -6, 0, 6 or 12 decibels. To avoid disturbing the examined consonants more than the surrounding vowels by an average level of noise, the noise level was fixed for keeping the desired signal to noise ratio. There were two types of noise:

- white noise (W)
- mixture of 4 speakers' voice modeling the cocktail party effect (P)

The results were obtained after evaluating 10,166 answers.

3. RESULTS

The recognition has been described as a function of the signal to noise ratio. The obtained recognition rates for a certain SNR and audio stimuli only are close to those of audio-visual stimuli with 6 dB lower SNR.

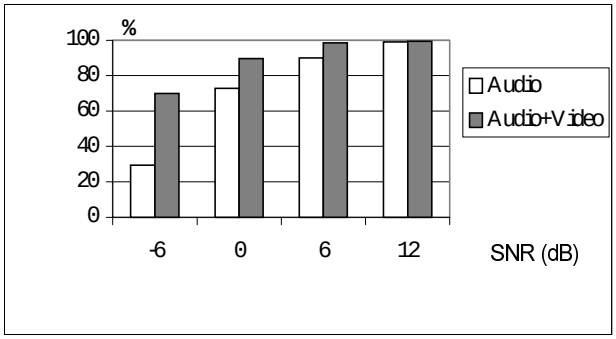

Figure 1. Recognition rates vs. signal to noise ratio for audio and audio-visual stimuli.

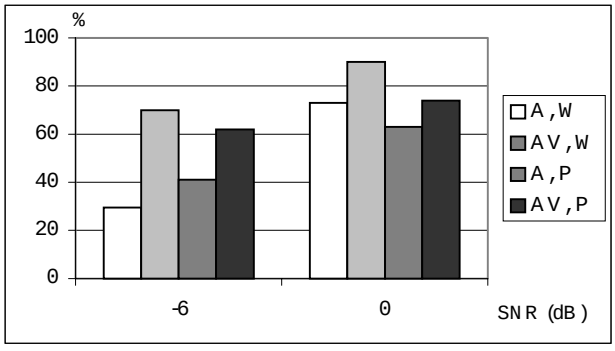

Figure 2. Recognition rates of consonants disturbed by white noise (**W**) and 'cocktail party' (**P**) with audio (**A**) and audio-visual (**AV**) stimuli.

3.1. Nature of Confusions

Three classes of confusions were defined:

- Confusing the consonants of the same articulation place. (**PA**)
- The class of excitation (stops, fricatives and whisper, semivowels and nasals, affricates) is correct.(**EC**)
- Others, when neither the place of articulation nor the sound class is correct.(**OT**)

As there occurred few confusions at 6 and 12 dB SNR, the confusion analysis concentrates on the –6 and 0 dB tests.

In case of audio stimuli most confusions are of third (OT) class, while with audio-visual stimuli most errors are on the same place of articulation (PA).

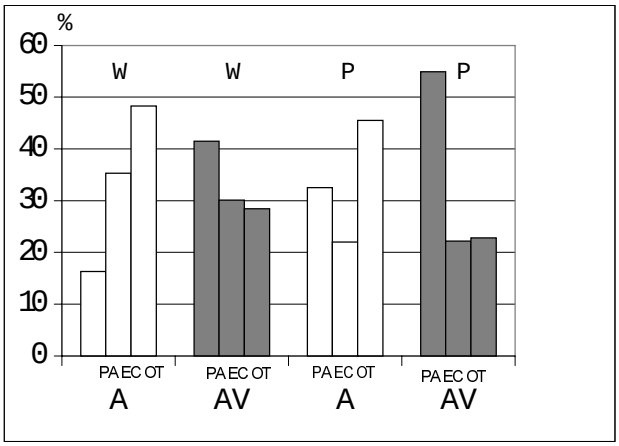

Figure 3. Confusion classes: correct place of articulation (**PA**), correct class of excitation (**EC**), neither of them is correct (**OT**).

3.2. Impact of Noise Type

The overall recognition rate of consonants is similar for white noise and cocktail party, but the impact of the two types of noise is different for voiced and unvoiced sounds. In white noise tests the recognition rate is identical for voiced and unvoiced sounds, 66.4% and 66.9%, respectively. Higher recognition rate was expected for voiced sounds. At cocktail party effect the recognition rate was much higher for unvoiced sounds (78.2%) than for voiced ones (45.6%). This tendency is not surprising if we take into consideration the dominance of voiced sounds in the mix caused by the high duration and magnitude of vowels.

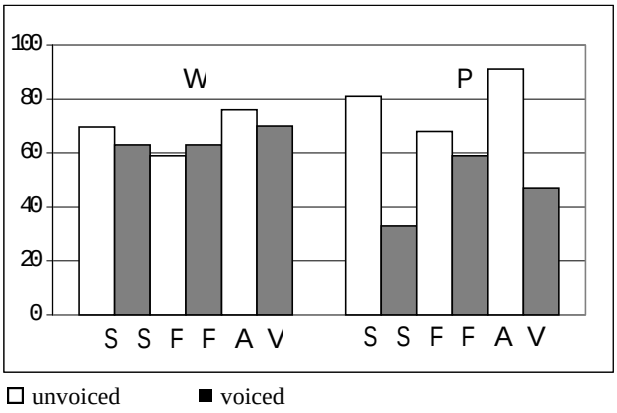

Figure 4. Recognition rates of voiced and unvoiced stops (**S**), fricatives (**F**), affricates (**A**), and semi vowels and nasals (**V**) disturbed by white (**W**) noise and cocktail party (**P**).

3.3. Excitation Classes

There is no significant difference between the recognition rates of sounds of different excitation classes with either audio or audio-visual stimuli. The higher rates of affricates are due to the previous effect as only unvoiced affricates were examined.

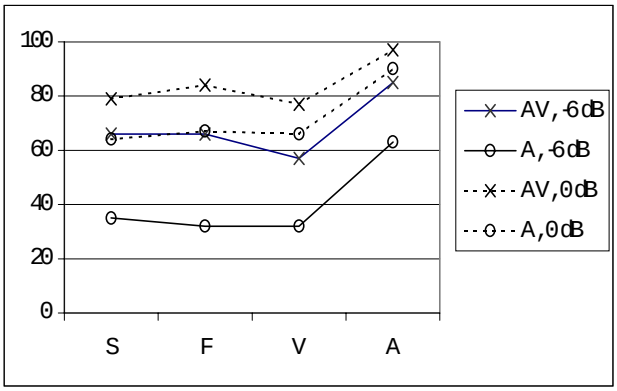

Figure 5. Recognition rates of different excitation classes: stops (**S**), fricatives and whisper (**F**), semi vowels and nasals (**V**) and affricates (**A**) with audio (**A**) or audio-visual (**AV**) stimuli and different signal to noise ratios.

The accuracy of excitation class is slightly improves with visual support. 45.4% of the total number of confusions are in the correct excitation class with audio stimuli. It is 58.6% with audio-visual stimuli.

The confusions are asymmetrical, e.g. 13.5% of the total semi vowel and nasal events were considered to be stops, while only 3.9% of stop consonant experiments were taken for semi vowels and nasals. 16.6% of fricatives was confused with stops but only 5.5% of stops were considered to be fricatives.

3.4. Place of Articulation

The following places of articulation were considered: bilabials (**bi**: p,b,m), labiodentals (**ld**: f,v), alveolars (**al**: t, d, c, sz, z, n), prepalatals (**pr**: ty, gy, cs, s, zs, l), palatals (**pa**: r, j, ny), velars (**ve**: k,g,h) *

The acoustic information alone is most important for back consonants, while the visual information is most important for front consonants (bilabials and labiodentals). The recognition rate improves 1.7 times with image support for bilabials and labiodentals and 1.1 times for other consonants.

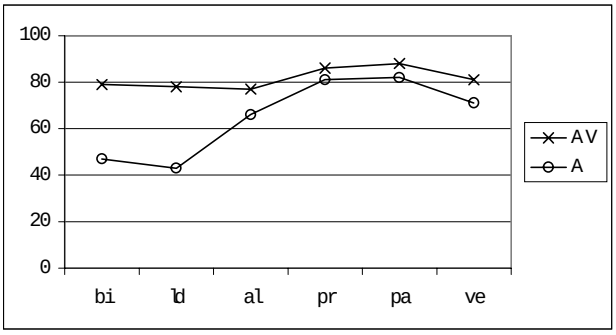

Figure 6. Recognition rates of consonants with different places of articulation.

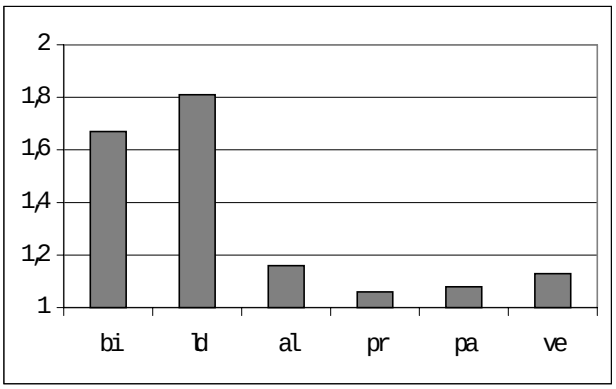

Figure 7. Improvement of recognition rates with video signal.

Confusion analysis shows a great improvement in identifying the place of articulation.

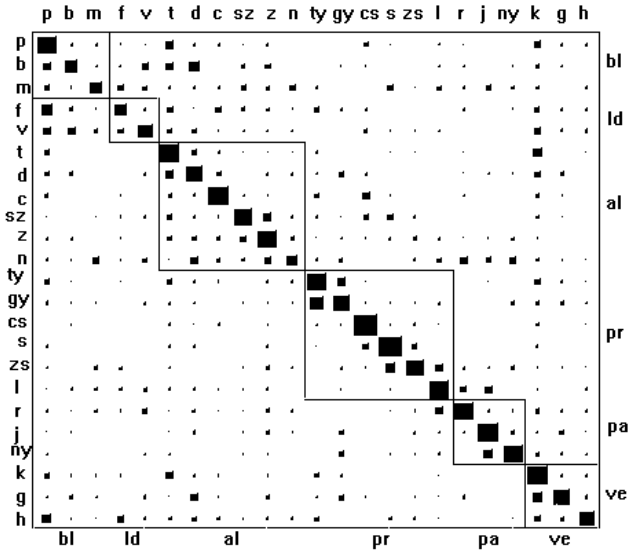

Figure 8. Hinton diagram of the confusion matrix with audio stimuli, showing the place of articulation. *

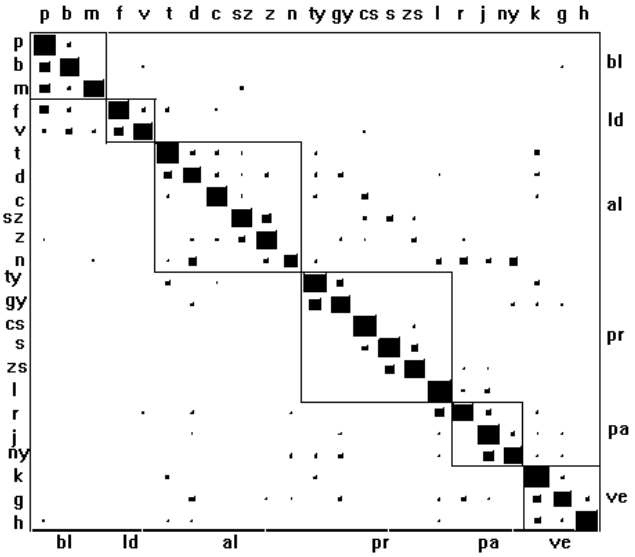

Figure 9. Hinton diagram of the confusion matrix with audio-visual stimuli, showing the place of articulation. *

***** SAMPA simbols of the Hungarian consonants: p (**p**), b (**b**), m (**m**), f (**f**), v (**v**), t (**t**), d (**d**), c (**ts**), sz (**s**), z (**z**), n (**n**), ty (**t'**), gy (**d'**), cs (**tS**), s (**S**), zs (**Z**), l (**l**), r (**r**), j (**j**), ny (**J**), k (**k**), g (**g**), h (**x**)

25.1% of the confusions are correct in the place of articulation with audio stimuli and 49% of the confusions are in the right place of articulation in case of audio-visual signal. Great support can be obtained from the image signal in the identification of the place of articulation.

4. CONCLUSIONS

The visual information is most important for front consonants. The recognition rates of bilabials and labiodentals are much higher with visual information. The visual signal can hardly support the identification of the excitation class. It does improve, however, the recognition of the place of articulation. The recognition rate of unvoiced sounds is much higher than that of voiced ones with cocktail party effect.

5. REFERENCES

1. J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K. E. Spens, T. Öhman: *The Teleface Project Multi-Modal Speech-Communication for the Hearing Impaired*. Eurospeech'97, Patras, Greece Proceedings

2. H. M. Saldana, D. B. Pisoni, J. M. Fellowes, R. E. Remez: *Audio-Visual Speech Perception Without Speech Cues: A First Report*. Speechreading by Humans and Machines Editors: D. G. Stork, M. E. Hennecke Springer-Verlag 1996.

3. M. M. Cohen, R. L. Walker, D. W. Massaro: *Perception of Synthetic Visual Speech*. Speechreading by Humans and Machines Editors: D. G. Stork, M. E. Hennecke Springer-Verlag 1996.

4. P. Cosi, E. M. Caldognetto: *Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications.* Speechreading by Humans and Machines Editors: D. G. Stork, M. E. Hennecke Springer-Verlag 1996.

5. C. Benoit, T. Guiard-MarignyB. Le Goff, A. Adjoudani: *Which Components of the Face do Humans and Machines Best Speechread*. Speechreading by Humans and Machines Editors: D. G. Stork, M. E. Hennecke Springer-Verlag 1996.