Modelling the McGurk effect

Ioana Sporea[1] and Andre Gruning[1]

1- University of Surrey - Department of Computing
Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, GU2 7XH - United Kingdom

Abstract. The current study investigates the McGurk effect by modelling it with neural networks. The simulations are designed to test the two main theories about the moment at which the auditory-visual integration happens. To further analyze the causes behind the McGurk illusion, the neural network that best models the effect is used to simulate the influence of language and the frequency of phonemes on auditory-visual speech perception, using two phonetic distribution from English and Japanese, with different empirical results in the McGurk effect.

1 Introduction

The McGurk effect is a perceptual illusion in the auditory visual speech perception domain. The effect occurs when an auditory stimulus, such as /ba/, is combined with a different visual stimulus of mouth movements, such as /ga/. In this situation of an incongruent auditory-visual input, people often perceive a different sound, in this case /da/ [1], which is from a phonetic point of view, an intermediate sound between the two stimuli.

Studies performed on auditory-visual speech perception show the importance of facial articulators. In noisy environments, seeing the speaker's face has a significant improvement in speech perception ([2], [3], [4], and [5]). The importance of visual articulation in speech perception is also emphasized by neuroimaging studies that show that during lip-reading in the absence of auditory speech input both primary auditory cortex ([6] and [7]) and secondary auditory cortex [8] are activated.

Several studies have been conducted in order to establish the moment of the auditory-visual integration during the processing of speech. While some researchers believe that the signals are processed parallel and independently and the integration occurs at a later stage [9], others suggest that the integration is produced at an early point in the speech processing ([10] and [11]).

Other studies suggest that the phonological repertoire influence the appearance of the McGurk effect. One such evidence is shown in [12], where Japanese subjects have been tested for the McGurk effect. The results indicate that in noise free environments the "Japanese McGurk effect" is weaker than the English one. The perception of the incongruent auditory-visual signals, produced by a Japanese speaker, was dominated by the auditory stimuli for about 80% of the time. For the most common set of incongruent syllables, auditory /ba/ combined with visual /ga/, has been heard 100% as /ba/ contrasting with the original result found by McGurk and MacDonald for English where for the same pair of stimuli /da/ was perceived from 64% to 98% of the time ([13] and [1]). Such differences in the perception of incongruent stimuli may be caused by the phonetic dissimilarities in the two languages or by cultural factors, such as the production or perception of speech.

2 Method

2.1 Pattern representation

The input of the neural network is represented by binary patterns corresponding to phonemes, the smallest units of sound, and to visemes, which are the basic units of speech in the visual domain [14]. The output of the network is the recognized sound, and therefore the output vector is identical with the auditory input pattern.

The phoneme is represented using the features utilized by the International Phonetic Alphabet [15]. The auditory patterns incorporate vectors that indicate the voice, the manner and the place of articulation. The viseme is represented by randomly generated vectors. The visual unit of sound contains less information than the auditory stimulus since several phonemes are mapped to one viseme.

2.2 Network structures

In order to test the main theories regarding the point at which the integration of the stimuli occurs, two feed-forward networks structures [16] have been used and compared. Both networks have two bands of inputs, consisting of the auditory and visual stimuli, and one set of outputs, which is the recognized sound.

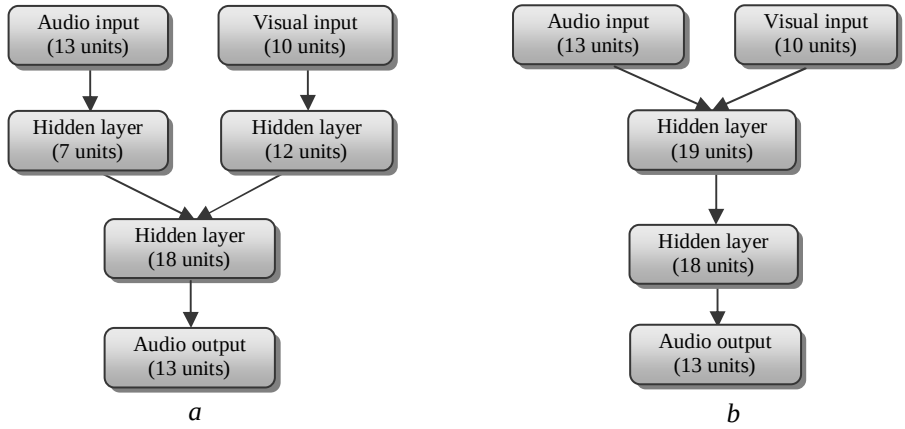

Fig. 1: (a) Late integration model. (b) Early integration model

The network in Figure 1 (a) corresponds to the late integration hypothesis. The structure has two individual and parallel hidden layers, and an integration layer. The network in Figure 1 (b) corresponds to the early integration hypothesis. This structure has an integration layer instead of the parallel hidden layers, without any individual pre-processing of the two stimuli. The two neural networks have the same number of neurons and have been trained and tested in the same conditions as described below.

2.3 **Training and testing**

Both networks have been trained with congruent audiovisual information arranged in a randomly generated sequence of a hundred patterns, replicating the way human subjects hear and see people producing sounds. The training sequence contains all the consonants from one language (e.g. English with 24 phonemes or Japanese with 16 phonemes).

Apart from the original patterns, the simulations take into consideration the influence of noise while training in order to simulate the presence of noise to audio-visual speech perception.

The training sequences contain only original patterns or a random combination of original patterns and blind channel patterns (the audio or visual input has null values) and/or noisy channel patterns (the audio or visual input has each of its values inverted with a probability of 10%). The combined training sequences consist of blind channel and/or noise channel patterns with a probability of 10% depending of the type of training sequence.

The neural networks have been tested using the following steps:

- Initialize the network with random weights, within the range of -0.1 and 0.1 uniformly distributed.
- Generate a new random training sequence of a hundred congruent patterns.
- Train the network with the back-propagation algorithm [16] with a learning rate of 0.1. The learning stops when the total mean squared error of the training set is sufficiently small (0.1).
- After each session of training the network is tested with the sets of incongruent patterns considered to produce the McGurk effect, the winning phoneme being determined as having the smallest Euclidean distance from the original vector to the output vector.

The results are stored and these steps are repeated 100 times, the networks being trained with a new generated sequence of patterns and new random weights.

3 Results

The tables below show the percentage of recognized phonemes (see 1–3 below) corresponding to the McGurk effect averaged across 100 trials, when trained with all the consonants from the English or Japanese phonetic alphabet and tested with three incongruent auditory-visual pairs of stimuli.

3.1 **Early and late integration models**

Table 1 shows the summarized results corresponding to the late and early integration models. For all three set of incongruent auditory-visual patterns, there can be seen a significant difference between the results of two neural network models.

Late integration	Training set	1.	2.	3.	Early integration	Training set	1.	2.	3.
	a.	39	46	65		a.	0	0	0
	b.	27	23	38		b.	5	3	3
	c.	50	41	53		c.	1	1	0
	d.	32	45	56		d.	2	5	2

Table 1: The output when trained with a random sequence of patterns having equal probability of appearance. The training is stopped when the network has reached the performance criterion on the congruent data.

A description of the notation found in the table follows:

The percentage of recognized phonemes corresponding to the McGurk effect for the incongruent sets of phonemes:

1. Audio /b/, visual /g/ - empirical data shows that is often perceived as /d/
2. Audio /p/, visual /k/ - empirical data shows that is often perceived as /t/
3. Audio /m/, visual /n/ - empirical data shows that is often perceived as /n/

The four types of random training sets used in the results tables are:

a. original patterns
b. original patterns and 10% blind channel patterns
c. original patterns and 10% noisy channel patterns
d. original patterns, 5% blind channel patterns, and 5% noisy channel patterns

3.2 **English and Japanese phonetic alphabets**

To further investigate the influence of language on the appearance of the McGurk effect, the late integration neural network is trained with English phonemes using the frequency of phonemes as found in conversational English [17] and with Japanese phonemes with frequencies found in the Japanese newspaper Asahi [18] and tested with the same incongruent stimuli. Table 2 shows the summarized results of the simulations when the network is trained and tested in the same conditions as above.

English phones	Training set	1.	2.	3.	Japanese phones	Training set	1.	2.	3.
	a.	68	83	83		a.	27	11	65
	b.	68	80	80		b.	32	17	58
	c.	42	53	89		c.	9	3	84
	d.	48	75	72		d.	27	15	66

Table 2: The output when trained with a random sequence of patterns having English and Japanese phonemes' frequencies

When the training set contains English consonants with English phonemes' frequencies, the results show a stronger McGurk effect for all three sets of incongruent auditory-visual phonemes compared to the results of the same network trained with English phonemes with equal probabilities of appearance.

Unlike the English phonetic alphabet, the Japanese phonetic alphabet does not contain certain phonemes, such as /r/ or /l/, and contains others that do not exist in English, such as /N/ [15]. As a consequence, for the plosive sets of consonants (auditory-visual /b/-/g/ and /p/-/k/) the results are considerably lower for fusion response than the results of the equivalent model when trained with English phonemes. In the case of nasal incongruent pair (auditory-visual /m/-/n/), the result are similar when comparing to the corresponding English trained network.

The late integration network has also been trained with patterns having equal probability of appearance and the results are similar to those of the network trained with the set of patterns with the frequency of Japanese phonemes.

4 Conclusions

The models that simulate late and early integration have different results when tested with incongruent auditory-visual patterns. Although both neural networks learn very well to recognize the congruent patterns, the McGurk effect in the early integration model is below 5% in all training cases. These results support the theory of the independent parallel processing and late audiovisual integration [9].

The outcome of the simulations with the English alphabet shows that the presence of blind channel in the training sequence (b and d) results in a decrease of the McGurk effect in almost all cases. When the training sequence contained noisy channel patterns (c), there can be seen a slightly stronger McGurk effect in some of the cases when trained with both alphabets.

When the late integration model is learning the auditory-visual patterns having the frequency of phonemes found in conversational English the network response to the incongruent stimuli is much closer to the experimental data. In [13] the percentage of fused responses to the incongruent stimuli /ba/-/ga/ is 64%, while the neural network produced the pattern /d/ 68% of times. The other sets of incongruent stimuli produce similar results to empirical data: for the /pa/-/ka/ pair McGurk and MacDonald [1] recorded 81% fused responses and the pair /ma/-/na/ was perceived as /na/ 80% of times, while the late integration model responded with the fused pattern in 83% of times for both sets of incongruent phonemes.

The results of the simulations using the Japanese phonetic alphabet are partly consistent with empirical data showing that in noise free environments the McGurk effect is weaker for Japanese listeners (see [12]). The results of the experiments conducted with Japanese listeners illustrate that speech perception is almost entirely limited to the auditory stimuli when presented with incongruent signals. The results of the simulation with the late integration model, although it presents a weaker McGurk effect for two sets of incongruent patterns, it also presents a strong McGurk effect for the pair /m/-/n/. These results suggest that the range of phonemes is not solely responsible for the weak Japanese McGurk effect found in empirical experiments. The weak “Japanese McGurk effect” may be a result of the difference in the range of consonants combined with cross-cultural dissimilarities in the perception of faces and facial expression. Other explanations for these results can be found in the identical mapping of the viseme to phoneme used for both English and Japanese training sequences, as empirical findings show that the number of viseme clusters depends on individual speakers [19].

References

[1] McGurk, H. and Macdonald, J., Hearing lips and seeing voices. Nature, 264:746-748, 1976.

[2] Helfer, K. S., Auditory and auditory-visual perception of clear and conversational speech. Journal of Speech, Language and Hearing Research, ASHA, 40:432-443, 1997.

[3] MacLeod, A., Summerfield, Q., A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation and recommendation for use. British Journal of Audiology, 24:29-43, Informa Healthcare, 1990.

[4] Dobb, B., The role of vision in the perception of speech. Perception, 6:31-40, Pion, 1977.

[5] Sumby, W., Pollack, I., Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26:212-215, ASA, 1954.

[6] Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R.,Tarkianen, A., et al., Primary auditory cortex activation by visual speech: An fMRI study at 3 T. NeuroReport, 16:125-128, 2005.

[7] Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al., Activation of auditory cortex during silent lipreading. Science. 276:593-597, AAAS, 1997.

[8] Bernstein, L. E., Auer, E. T., Jr., Moore, J. K., Ponton, C. W., Don, M., Singh,M., Visual speech perception without primary auditory cortex activation. Neuroreport, 13:311–315, LWW, 2002.

[9] Massaro, D. W., Stork, D. G., Speech recognition and sensory integration. American Scientist, 86:236-244, Sigma Xi, 1998.

[10] Bernstein, L. E., Independent or dependent feature evaluation: A question of stimulus characterization. Behavioral and Brain Sciences, 12:756-757, Cambridge University Press, 1989.

[11] Green, K.P., and Miller, J.L., On the role of visual rate information in phonetic perception. Perception & Psychophysics, 38:269-276, Psychonomic Society, 1985.

[12] Sekiyama, K., Tohkura, Y., McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. The Journal of the Acoustical Society of America, 90:1797 – 1805, ASA, 1991.

[13] Macdonald, J., McGurk, H., Visual Influences on speech perception process. Perception and Psychophysics, 24:253-257, Psychonomic Society, 1978.

[14] Fisher, C. G., Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 11:796-804, ASHA, 1968.

[15] Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet, Cambridge University Press, Cambridge, 1999.

[16] Beale, R., author, Jackson, T., author, Neural Computing: An Introduction, IOP Publishing Ltd, Bristol, Great Britain, 1990.

[17] Mines, M. A., Hanson, B. F., Shoup, J. E., Frequency of occurrence of phonemes in conversational English. Language and speech, 21(3):221-41, Kingston Press, 1978.

[18] Tamaoka, K., Makioka, S., Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. Behavior Research Methods, 36(3): 531-547, Psychonomic Society, 2004.

[19] Kricos, P., Differences in visual intelligibility across talkers. In: Stork, D. G., Hennecke, M. E., editors, Speechreading by Humans and Machines, NATO ASI Series, pages 43-53, Springer, Berlin, 1996.