A distributed model of memory for the McGurk effect

Ioana Sporea, *Member, IEEE,* André Grüning

***Abstract*—The present paper is investigating the modelling of the McGurk effect, an audio-visual speech perceptual illusion, with a distributed model of memory. The network is trained with congruent auditory and visual patterns and tested with incongruent sets of patterns considered to produce the McGurk effect.**

I. Introduction

THE McGurk effect is an auditory visual perceptual illusion in the speech perception domain [1]. The effect is observed when incongruent auditory visual stimuli are presented to subjects. Studies performed on speech perception show that when an auditory stimulus, such as /ba/, is combined with a different visual stimulus of mouth movements, such as /ga/, people often perceive an intermediate sound, /da/. Our previous work on modelling the McGurk illusion [2] with a feed-forward neural network suggests that the appearance of the effect highly depends on the range of phonemes. These results are consistent with empirical studies conducted on Japanese speakers that resulted in a weaker McGurk effect [3].

Our simulations also suggest that the frequencies of the phonemes have an important role in the occurrence of the McGurk illusion. When the network model was trained with a random sequence of combined auditory (phonemes) and visual (visemes) patterns with equal probability of appearance, the network recognized the sound /d/ 39% of the cases when presented with the incongruent pair /b/-/g/. When the network was trained with a random sequence of patterns having English phonemes' frequencies as found in conversational English [4], the percentage of the produces pattern /d/ increased to 68% reflecting findings in psycho-linguistic experiments [1].

The current paper is extending the modelling of the McGurk effect by using a distributed model of information processing and memory [5]. The distributed memory model is an autoassociative network and consists of a collection of simple and highly interconnected processing units.

Figure 1 shows the internal structure of a network with eight interconnected processing units according to [5]. Each unit receives an external input, within the range of -1 to +1 and an internal input representing the weighted sum of the activations of the other units in the module, as each processing unit is connected to all other units. The net input of the processing unit, n_i, represents the sum of all the internal inputs and the external input:

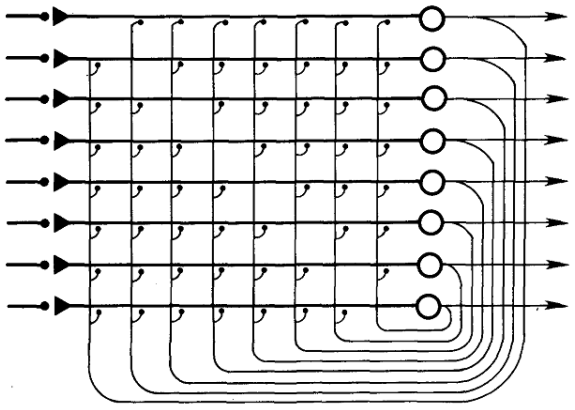

Fig. 1. An example of a network with eight interconnected units. Reproduced from McClelland and Rumelhart, Fig. 1 [5]

$n_i = e_i + \sum_j w_{ij} a_j$, where e_i is the external input, a_j is the activation of the unit j and w_{ij} is the weight between the units i and j. If the net input is positive the activation of this unit is then incremented by an amount proportional to the distance left to the ceiling activation of +1. If the net input is negative the activation is then decremented by an amount proportional to the distance left to the floor activation of -1. The equations used for updating the activations are:

$\Delta a_i = En_i(1-a_i) - Da_i$, if $n_i > 0$,

$\Delta a_i = En_i[a_i - (-1)] - Da_i$, if $n_i \le 0$, where E and D are global parameters that represent the rates of excitation and decay, respectively.

The processing ends when the pattern of activation is that the network produced settles down and stops changing. In the simulations described in the present paper, the network runs for a maximum of 50 processing cycles. The units have activations values which range from -1 to +1, with the value 0 representing a neutral resting value [5].

After computing the net input, n_i, of the unit i, the activations of the units are updated. The weights between the processing units are adjusted using the delta rule in order to determine the amount and direction of the change of the connection weights. The weights are modified according to the following formula [5]:

Manuscript received April 26, 2010.

I. Sporea is with the Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK (corresponding author phone: +44 (0)1483 68 6056; e-mail: i.nica@surrey.ac.uk).

A. Grüning is with the Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK (e-mail: a.gruning@surrey.ac.uk).

$\Delta w_{ij} = S(e_i - \sum_j a_j w_{ij}) a_j$, where S is a constant that controls the amount of the modifications in the weights.

The algorithm based on the delta rule is reducing to zero the difference between the external and the internal input. As a consequence, when a partial or distorted pattern is presented, part of the units will be active and will tend to reproduce the rest by the connections between units [5].

II. Modelling the McGurk Effect with the Distributed Model of Memory

The different perceived sounds when presented with incongruent auditory visual stimuli suggest a strong association between auditory and visual inputs. The distributed model of memory and information processing model [5] described above is assumed to associate the auditory and visual stimuli.

The network's input consists of patterns representing the auditory stimulus (the phoneme, which is the smallest unit of sound) and the visual stimulus (the viseme [6], the basic unit of speech in the visual domain). The representation of the patterns is similar to the one used in our previous work [2]. The new patterns are vectors with elements of -1 and 1, instead of the previous representation with elements of 0 and 1 [2].

The phonemes are represented by bipolar 13-element vectors which encode speech features utilized by the International Phonetic Alphabet [7]. The auditory patterns incorporate vectors that indicate the voice, the manner and the place of articulation. Each feature is represented by vectors; while the vectors used to encode each manner of articulation were generated randomly, the vectors used to represent the place of articulation have been constructed using the Gray code [8] in order to reflect their order in the vocal tract from bilabial to glottal. The Gray code is a binary system where two successive values differ in only one bit. Using the Gray code, this encoding reflects the places of articulation as they are located in the vocal tract.

The visemes correspond to groups of phonemes, as the visual input contains less information than the auditory input. Therefore, several phonemes are mapped to one viseme, for example the phonemes /f/ and /v/ are in the same visual group. The visual patterns are represented by randomly generated bipolar vectors with 10 elements. The vectors are independent since viseme clusters do not have common features.

III. Results

The distributed memory system has been trained with congruent audiovisual patterns arranged in a randomly generated sequence of a hundred patterns. The training sequence consists of all 24 consonants from English as found in the International Phonetic Alphabet [7].

The distributed memory system has been tested using the following steps:

- Initialize the internal weights with null values;
- Generate a new random sequence of congruent patterns to be used in the training process;
- Train the network for 10 learning cycles with the delta rule;
- After each session of training the network is tested with the congruent patterns, with auditory patterns (the visual part has null values), distorted patterns, and with three incongruent auditory visual pairs of stimuli considered to produce the McGurk illusion. A stable pattern of activations is considered to be achieved when there is no difference in the updated activations or when it reaches a maximum of 50 processing cycles. The produced pattern is determined as having the smallest Euclidean distance from the original pattern;
- The steps are repeated a hundred times, the results being stored and averaged.

In order to test again the theory that the phonemes' frequencies [2] influence the appearance of the McGurk effect two types of training sequences have been used. The network has been trained with a random sequence of patterns with equal probability of appearance and the other one having English phonemes' frequencies as found in conversational English [4].

After the network was trained with a training sequence of congruent patterns with equal probability of appearance and tested with the original congruent patterns and auditory patterns (the visual part has null values). When tested with congruent patterns the average error across a hundred trials is 1.6% with a range between 0% and 9%. When the network is tested with auditory patterns the average error is 9.9% with a range between 0% and 57%.

When the network was trained with a training sequence with English phonemes' frequencies the average error for congruent patterns is 0.6% with a range between 0% and 6%. When the network was presented with auditory patterns the average error across a hundred patterns is 1.9% with a range between 0% and 32%.

The network has also been tested with distorted congruent patterns (each element of the pattern has an independent probability of 20% of being flipped), resulting in an average error of 1.5% across a hundred trials in both types of training sequences of patterns.

These results confirm the fact that the network has learned the patterns and association between auditory and visual inputs. Moreover, the model is capable of pattern completion and noise elimination.

The network was tested with the following sets of incongruent patterns:

- Auditory /b/, visual /g/ - often perceived as /d/ in empirical studies (98% of the cases [1])

- Auditory /p/, visual /k/ - often perceived as /t/ in empirical studies (81% of the cases [1])
- Auditory /m/, visual /n/ - often perceives as /n/ in empirical studies (80% of the cases [9])

Studies performed on audiovisual speech perception shows that in noise-added auditory visual condition Japanese speakers experience a much stronger McGurk effect than in noise-free condition, suggesting that people rely on the visual input in the presence of auditory uncertainty [3]. Therefore, noise has been added to the incongruent patterns, both on the auditory and visual part.

In the table the auditory response means that the network response is closest (in Euclidean distance) to the congruent pattern corresponding to the auditory part of the incongruent pattern - for example for the incongruent audiovisual pair /b/-/g/, the auditory response is /b/, the visual response means that the network response is closest to the congruent pattern corresponding to the visual part of the incongruent pattern - /g/ in this case, the fused response is the intermediate sound different from the auditory and visual patterns considered as the McGurk perceptual illusion - /d/ for this incongruent pattern. Table I shows the obtained results averaged across a hundred trials.

TABLE I
THE OUTPUT OF THE NETWORK AFTER TRAINED WITH CONGRUENT PATTERNS AND TESTED WITH NOISY INCONGRUENT PATTERNS

	Audiovisual input	Auditory response	Visual response	Fused response
1	/b/-/g/	25%	25%	50%
	/p/-/k/	94%	1%	5%
	/m/-/n/	14%	86%	
2	/b/-/g/	14%	41%	45%
	/p/-/k/	14%	41%	45%
	/m/-/n/		100%	

A description of the notation found in the table follows:

1. The network was trained with a randomly generated sequence of congruent patterns having equal probability of appearance.
2. The network was trained with a randomly generated sequence of congruent patterns having English phonemes' frequencies.

Furthermore, the model has been tested with incongruent patterns in relation with prime congruent patterns. Due to the fact that the activations are calculated based on the values of the previous activations, the distributed model of memory is sensitive to priming effects [5]. Therefore, the distributed model of memory is represented as a composition of the patterns of activations determined by the processing of each input presented to the neural network. Each time an input is presented to the network, the activations are changed corresponding to the present stimuli and also to the current state of the memory. The new state of the memory is thus determined according to the processing of the given patterns. As McClelland and Rumelhart explained [5], the experience of perceiving an item affects the subject's later performance. If, for example, a word is perceived twice within a reasonable interval of time, the prior presentation makes it possible for the subject to recognize the word faster, or from a briefer presentation. McClelland and Rumelhart showed that the model provide an account for the existence of priming effect as a function of congruity between the prime event and the test event [5].

After the network was trained with congruent patterns, it has been tested with the congruent pattern and then the partial patterns. This time, the error of the network across a hundred trials dropped to 2.7% with a range between 0% and 17% after training with the sequence of patterns having equal probability of appearance. After the network was trained with a random sequence of patterns with English phonemes' frequencies the error across a hundred trials is 0.1% with a range between 0% and 3%.

In order to investigate the priming effect on the appearance of the McGurk illusion the network was presented with the fused pattern before presenting the incongruent audio-visual pair of phonemes. Table II shows the results obtained across a hundred trials.

TABLE II
THE OUTPUT OF THE NETWORK AFTER TRAINED WITH CONGRUENT PATTERNS AND TESTED WITH NOISY INCONGRUENT PATTERNS AFTER PRESENTING A PRIME

	Audiovisual input	Auditory response	Visual response	Fused response
1	/b/-/g/	2%		98%
	/p/-/k/	14%		86%
	/m/-/n/	21%	79%	
2	/b/-/g/			100%
	/p/-/k/			100%
	/m/-/n/		100%	

Presenting a prime to the network before presenting incongruent patterns resulted in a much stronger McGurk effect for the network trained with the sequence of patterns with equal probability of appearance. The network trained with the sequence of patterns having English phonemes' frequencies responded only with the fused patterns.

IV. CONCLUSION

In our previous model of the McGurk effect we have used a feed forward network trained with the back propagation algorithm. The network architecture that best modelled the McGurk illusion has three hidden layers making the training of the network very slow. In comparison, the distributed memory system is able to learn the same set of patterns considerably faster. The autoassociator has significantly less weights to adjust and the delta rule requires fewer computations to be performed than the back propagation algorithm.

When the distributed model was tested with partial patterns, the network was able to retrieve the original

pattern 94.1% of the cases on average, respectively 98.5% of the cases when tested with distorted patterns. Although the patterns are not orthogonal vectors, this type of autoassociator is performing very well on learning the patterns as well as on pattern completion and noise elimination.

The fused responses to noisy incongruent patterns can be observed as dominant for most of the incongruent pairs of phonemes, these results being consistent with empirical studies on audiovisual speech perception in noise conditions [3]. One explanation for this behaviour can be found in the network's ability to recognize incomplete and distorted patterns. The patterns presented to the distributed model of memory were both incongruent and distorted. As the network tried to complete both parts of the input, the audio and visual stimuli, the response was the fused pattern corresponding to the McGurk effect. Since the encoding of the auditory patterns reflects the way phonemes are produced by the human vocal tract, we can conclude that the network behaved in a similar manner as human subjects.

When the network was also presented with a prime pattern, the model responded with a much stronger McGurk effect. However, there is not sufficient evidence from psycholinguistic experiments to suggest that the McGurk illusion is sensitive to priming effects. On the other hand, priming events were likely to occur in McGurk effect experiments as the order of stimuli was chosen randomly [9]. The priming event acted as a facilitator for the neural network to recognize the fused pattern.

While the distributed memory model has been tested on partial and distorted congruent patterns and on incongruent audiovisual patterns, there have been two types of sequences of congruent patterns used in learning. In all the conducted tests, the network trained with the sequences of patterns having English phonemes' frequencies performs better, having a higher accuracy than in the case of training with patterns with equal probability of appearance. Moreover, when the network is presented with the sets of incongruent patterns known to produce the McGurk effect, the network responded with the fused pattern more consistently and stronger for most of the incongruent pairs of patterns. The distributed model of memory is thus susceptible to the frequency of the presented patterns as one might expect a human subject to be. These results are consistent with our previous work, as using English phonemes' frequencies resulted in a stronger McGurk effect in our simulations [2].

REFERENCES

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

[2] I. Sporea, A. Gruning, "Modelling the McGurk effect," Proceedings of the 18th European Symposium on Artificial Neural Networks (ESANN 2010), to be published, Bruges, Belgium, 2010.

[3] K. Sekiyama,Y. Tohkura," 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*, vol. 90, pp. 1797 – 1805, 1991.

[4] M. A. Mines, B. F. Hanson, J. E. Shoup, "Frequency of occurrence of phonemes in conversational English," *Language and speech*, vol. 21, issue 3, pp. 221-41, 1978.

[5] J. L. McClelland, D. E. Rumelhart, "Distributed memory and the representation of general and specific information," *Journal of Experimental Psychology: General*, vol. 114, pp. 159-197, 1985.

[6] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 1i, pp. 796-804, 1968.

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

[8] F. Gray, "Pulse code communication," U.S. Patent 2,632,058, March 17, 1953.

[9] J. Macdonald, H. McGurk, "Visual Influences on speech perception process," *Perception and Psychophysics*, vol. 24, pp. 253-257, 1978.