USING SPEECH SIGNALS TO IMPROVE VISUAL FACIAL IMAGE RECONSTRUCTION: AN RNN APPROACH TO EXPLORE THE MUTUAL INFORMATION

S-H Luo and R. W. King

Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney

ABSTRACT - We present a novel approach for improving visual facial image reconstruction by utilizing information in the accompanying acoustic speech signal. From analysis of the speech signal and knowledge of the mutual information between speech and visual features of facial images, the method can be used to synthesize moving facial images. A recurrent neural network is used to map between the acoustic and visual spaces: the input to the RNN is 21 acoustic speech features and the output is the position values of 15 facial feature points and the first 20 coefficients of a principal components representation of the mouth area.

INTRODUCTION

This research explores the mutual information which exists between an acoustic speech signal and the corresponding visual facial image. The work is motivated partly by potential applications: to improve the quality of the mouth image in low-bit rate videophone and videoconference systems; and to provide facial image animation in multimedia communications. The articulatory connections between the audible speech signal (what the speaker says) and the visible speech signal (what the speaker's face looks like) provide the possibility of deriving one kind of information from the other (Massaro, 1987).

In natural face to face speech perception, acoustic speech is reinforced to some degree by observation of the speaker's mouth. Visible speech perception is especially useful when the acoustic signal is degraded by noise, while for hearing impaired people, lip-reading provides direct access to speech. Several workers have shown that recognising the visual speech signal can significantly improve speech recognition performance, compared to acoustic recognition alone (Petajan, 1987, Duchnowski et al., 1994).

Running parallel with these studies have been several endeavors to utilize the acoustic speech signal to improve visual facial image reconstruction or synthesis. Computer graphics animation methods for generation of talking faces, with the synthesized mouth image being derived from phonetic descriptions of text have been demonstrated (Brooke et al., 1994; Massaro and Cohen, 1994). Brooke's method uses hidden Markov modelling and principal component analysis and incorporates visible features of primary articulators such as lips, teeth, and tongue derived from real images. Lewis (1991) has described an automatic lip synchronizing method that created mouth animation synchronized to the speech input by recognizing phonemes by linear prediction and associating the phonemes with mouth positions to provide key frames. Despite the speech to image synchronization, the synthesized images were reported as lacking reality.

The derivation of acoustic speech features from visual speech features, or vice versa, can be considered as a transformational mapping between the two feature spaces. While no explicit relationships between acoustic and visual speech features have been found; artificial neural networks have become a widely accepted and efficient approach to the computation of the transformation. Promising results have been obtained for utilizing neural networks to estimate the acoustic speech structure from the concurrent visual speech signals (Sejnowski, et al., 1989). There has been relatively little research on the inverse process of inferring visual speech information from acoustic speech signals, although Knotts et al. (1993) have discussed a system that can aid hearing-impaired persons to produce normal speech. In this, an artificial neural network was used to transform from the acoustic parameters into visible facial movements with particular emphasis on the tongue, lips and jaw position.

This paper presents a novel approach to facial image reconstruction by utilizing the acoustic speech signal. Selected speech feature parameters are applied to the input of a feature mapper with two recurrent neural networks (RNN). This generates feature parameters which describe facial actions and expressions, as described in the next section. Continuously varying images of the face corresponding to the speech can be synthesized (Luo, 1994) with the visual feature parameters derived from the transformation combined with a database which includes prototype 3D human head models, mouth images and eigenmouths.

Real data, both speech signals and corresponding facial images, have been used to train the feature mapping RNNs and justify the algorithms. Quite good reconstructed image quality is achieved, as assessed against subjective fidelity criteria of "naturalness" and "faithfulness" to assess the overall head images, and an objective image quality criterion which measures the differences between the synthesized and original mouth area images corresponding to the enunciation of the speech.

MAPPING ACOUSTIC SPEECH TO VISUAL SPEECH

With acoustic features at their input, the trained RNNs can output parameter values describing the positions of characteristic facial and mouth points from which the whole facial image may be reconstructed. In this section, the selections of both the acoustic speech features and the visual speech features are first introduced, then the details of the feature mapping methodology are discussed.

Input Acoustic Feature Selection

The following 21 acoustic speech features are selected to characterise the acoustic speech signal: the first 12 mel frequency cepstral coefficients (MFCCs), log energy of the speech, the first 4 formant frequencies and bandwidths (f_1, f_2, f_3, f_4, b_1, b_2, b_3, b_4). These features have been chosen for their ability to discriminate both phonemes and visemes (Luo and King, 1994). Figure 1 demonstrates the methodology for calculation of these features.

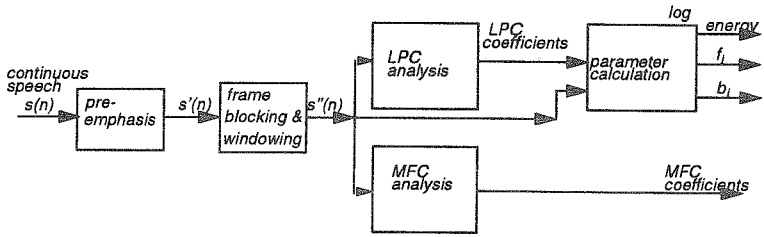

FIGURE 1. The block diagram of the acoustic speech feature calculation

Note: LPC: linear predictive coding; MFC: mel frequency cepstral;
f_i & b_i: formant frequency & bandwidth;

Output Visual Feature Selection

Two groups of visual speech features are involved in the mapping. The first one is the 30 values (fp_1, fp_2, ..., fp_{30}) representing facial feature points characterizing the facial expression with speech, including 14 mouth contour feature points and the chin point; the other is the 20 coefficients (ω_1, ω_2, ..., ω_{20}) of the inner mouth area image principal components derived with the optimal Karhunen-Loeve transformation (KLT) (Gonzalez, 1987).

For training these networks from real images, the chin point and the 14 mouth contour feature points are extracted from a front-viewed facial image in a 2-step procedure (Luo, 1994). First, the head contour and mouth contour are located by combining an active contour model with multi-layer neural networks. Then these feature points are detected by the calculation of object orientation and mass centre under the assumption that the face is symmetric. Using these facial feature points, a complete facial image can

be synthesized through a 3-D head model-based transformation. This does not, however, provide satisfactory modelling of the inner mouth area such as lips, teeth, tongue and palate. In order to get over this problem, the inner mouth images are represented by coefficients derived with the optimal KLT.

The KLT provides an optimal coordinate system to describe the inner mouth image. This is a set of basis vectors, also referred to as principal components or, in this application, eigenmouths (Sirovich & Kirby, 1987). These basis vector μ_i's are obtained from a training ensemble of inner mouth images. The first basis vector encodes the maximum variation between the training images and successive basis vectors encode progressively smaller variations. It is an optimal transformation in the sense that the mean-square error introduced by truncating the expansion is a minimum. In the optimal system, any inner mouth image X_i' (within or outside the training ensemble) can be represented, with minimal truncating error, with M eigenmouths as

$$\hat{X}'_i = \sum_{k=1}^{M} \omega_k u_k + m_X \qquad \text{(EQ 1)}$$

where $\hat{X}'_i$ is the approximation version of X_i' constructed with M eigenmouths; m_X is the mean vector of training set; the coefficient ω_k describes the contribution of the kth eigenmouth in representing X_i', and is given by

$$\omega_k = u_k^T (X'_i - m_X) \qquad k = 1, \ldots, M \qquad \text{(EQ 2)}$$

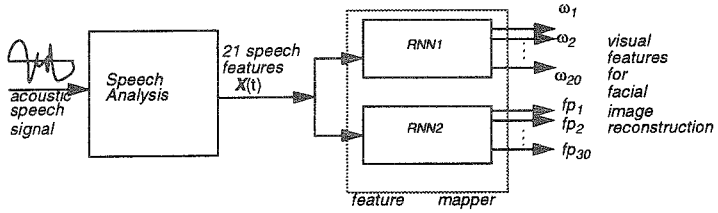

FIGURE 2. Mapping acoustic speech features into visual speech features

Note: ω_i's: the principal component coefficients of the inner mouth area image
fp_i's: positional coordinates of the facial feature points

Feature Mapping Methodology

The methodology for mapping the acoustic speech into the visual speech is depicted in Figure 2. The feature mapper is composed of two recurrent neural networks, RNN1 and RNN2, which infer the 20 principal component coefficients of the inner mouth area image and the 30 positional coordinates of the facial feature points respectively. The inputs are all the same to both the RNNs.

The process of mapping acoustic speech features into visual speech features is one of parameter estimation. Thus the output space is described with parameters of continuous value, generally normalized to the range of -1 and +1. The RNN can be considered as a sequence of backward error propagation networks (Rumelhart, Hinton & Williams, 1986). For both RNNs we have used the topology (Werbos, 1990) illustrated in Figure 3. There are both forward connections (dark shaded arrows) between current input neurons and current hidden neurons and output neurons, and backward connections (shallow shaded arrows) operating over the previous time period. The RNNs have been trained using back propagation through time, with the delta-bar-delta procedure for updating network weights (Jacobs, 1988).

EXPERIMENTS AND RESULTS

Data Collection and Pre-processing

To test the feasibility of the mapping from acoustic speech features to visual speech features, acoustic speech signals (digitized speech waveforms) and visual speech signals (facial images), corresponding to the enunciation of the English phonemic alphabet and 50 selected continuous sentences, were

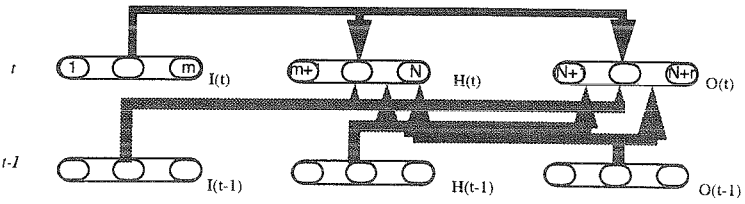

FIGURE 3. The RNN used in feature mapping

Note: I(t), H(t), O(t): input, hidden, and output neurons at time t; t-1: 1 time period ahead time t;
forward connection backward connection

recorded. The speakers facial images were recorded frame by frame with 40ms frame length using the PAL video standard. The images were digitised to 316*450 pixels monochrome with a 256 level grey scale. For each image, two kinds of feature extraction procedures were employed. Firstly, motion analysis was applied to track the contour of the head and the contour of the mouth, then the 30 values representing facial feature points were calculated and normalized to [-1, +1]. Secondly, the mouth area was automatically located and cut to form a mouth area sub-image of size 100*50 pixels, to which the KLT analysis was applied to derive the coefficients of the eigenmouths.

The acoustic speech signals were analysed to produce the 21 input speech features for the mapping. Since each image frame has a length of 40 ms, the speech analysis process provides a set of speech feature every 40 ms. In practice, the speech features are calculated every 20ms, but only the set of every other 20 ms is selected as the speech features of the current 40ms segment.

The number of free parameters in the feature estimation neural network is a trade-off between computational efficiency, desired accuracy, and the potential problem of over-learning the data. On the one hand, if the network size is not large enough, the network is unable to pass the relevant information in the acoustic speech signals to the output units and the network performs poorly. On the other hand, if the network size is too much large, the network will try to memorize the idiosyncratic details of the training set, and may fail to generate rules that apply to the testing set. Various network topologies were selected and tested for RNN1 and RNN2 in order to choose an adequate network size while minimizing the effects of over-learning. The training process was set to stop after 200,000 iterations considering the practical computing time and the accuracy achieved. Each of the networks was trained with 5585 frames of training data set and tested on 620 frames of testing data. Two kinds of network topology for RNN1 were investigated, with sizes 22-10-20, and 22-40-20. The sizes of 22-10-30 and 22-20-30 were used in the investigation for RNN2.

Results and Discussion

Table 1 summarizes the mapping performance of the two network configurations and the quality of the reconstructed images based on these principal components for both the training set and the testing set.

Net Input	Net Output	Net Size	Output Error (OE) ($*10^{-3}$)		Mean Square Error (MSE) ($*10^{6}$)	
			train	test	train	test
1st 12 MFCCs, f_1, f_2, f_3, f_4, b_1, b_2, b_3, b_4, log of energy	ω'_1, ω'_2, ..., ω'_{20}	22-10-20	2.3	4.9	1.744	3.478
1st 12 MFCCs, f_1, f_2, f_3, f_4, b_1, b_2, b_3, b_4, log of energy	ω'_1, ω'_2, ..., ω'_{20}	22-40-20	1.7	3.3	1.332	2.891

TABLE 1. Eigenmouth mapping performance of RNN1 with different topologies

The output error OE reflects the network's ability of inferring the 20 principal components (PC), and is defined as the average square error of the actual network output ω_i's to the desired network output

summed over the whole training set or testing set. The better mapping result with OE equating 0.0033 is achieved on the testing set by the RNN1 with 40 hidden neurons. Figure 4 gives the original first 5 PC coefficients and the inferred ones corresponding to the enunciating of phonemes /ʃ/ and /ʒ/ expanding over 30 frames. From the figure, it can be seen that the shapes of the PC coefficients mapped by RNN1 coincide with those of the original ones quite well.

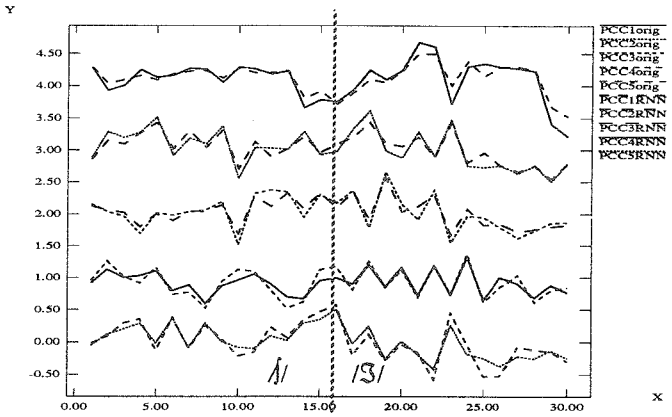

FIGURE 4. Variation of the first 5 principle component coefficients, both the original ones and those inferred by RNN1, for phonemes /ʃ/ and /ʒ/ expanding over about 30 frames.

Note: X direction: time sequence; Y direction: normalized ([-1,+1]) coefficients spaced at one unit;
PCCiorig: original coefficient of the ith principle component;
PCCiRNN: inferred coefficient of the ith principle component;

The influence of the OE value on the image quality can be measured with the objective criterion MSE which is defined as the mean square error, averaged over the training set or testing set, between the original face image and the image reconstructed with M principal components derived by RNN1. Figure 5 shows the feature mapping results by RNN1 with structure 22-40-20 on 4 consecutive mouth area images in testing set. The first row gives 4 real mouth area images corresponding to the speech of which the acoustic features are input to RNN1 to infer 4 visual feature vectors, with each vector containing 20 coefficients of eigenmouths. The second row shows the corresponding mouth area images reconstructed with the eigenmouths according to the inferred coefficients. It can be seen that the mouth area images reconstructed according to the mutual information between acoustic speech and visual speech are generally "faithful" to the original expressions.

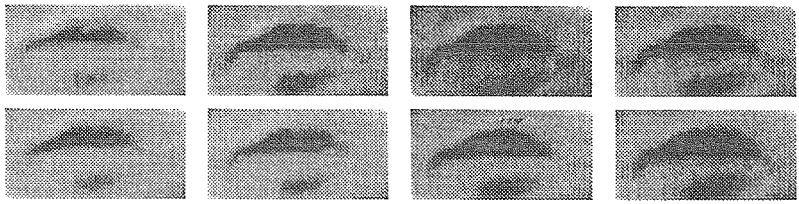

FIGURE 5. Results of mouth area images mapping from speech by RNN1

(The upper row is 4 consecutive mouth area images from testing set; the lower row is the corresponding images reconstructed with the eigenmouth coefficients inferred from speech by RNN1.)

Experiments on mapping acoustic speech onto facial features characterizing the whole face actions and expressions with speech have also been conducted. Based on the criteria of the normalized average deviation between the desired feature set and the feature set derived with RNN2, and a shape similarity function which measures the shape similarity between the desired mouth contour and the mouth contour represented by the features derived with RNN2, satisfactory facial feature reconstructions have been achieved.

CONCLUSIONS

We have described a novel approach for mapping acoustic speech feature space onto visual speech feature space by recurrent neural networks. The system has been trained efficiently and tested with spoken and facial data collected from one speaker. Criteria have been developed to assess the quality of the reconstructed facial image sequence. Initial results are promising. Clearly the RNN mapping technique facilitates good modelling of the temporal connections between the acoustic and visual signals.

ACKNOWLEDGEMENT

S-H Luo holds an OPRS award. The authors are grateful to ATERB for their financial support of this work.

REFERENCES

Brooke, N. M., Scott, S. D., (1994) *Computer graphics animations of talking faces based on stochastic models*, 1994 Inter. Sympo. on Speech, Image Processing & Neural Networks, 13-16, April, Hong Kong, pp. 73-76.

Chang, S. C.,Takebe, T., Harashima, S.H. (1992) *Analysis of facial expressions using a three-dimensional facial model*, Systems and Computers in Japan, Vol. 23, No. 12, pp. 13-27.

Duchnowski, P, Meier, U. and Weibel, A (1994), *See me, Hear me: Integrating Automatic Speech Recognition and Lip-Reading*, Proc. ICSLP-94, pp. 547-50.

Gonzalez, R. C. (1987) *Chapter 3 "Image Transforms", Digital Image Processing*, Addison-Wesley Publishing Company, USA, pp. 61-137.

Jacobs, R. A. (1988) *Increased rates of convergence through learning rate adaptation*, Neural Networks, Vol. 1, pp. 295-307.

Knotts, S. L., Freeman, J. A., Harman, T. L. (1993) *Perceptual linear prediction and neural networks for speech analysis*, 3rd Workshop on Neural Networks: Academic/Industrial/NASA/Defence, pp. 322-327.

Lewis, J., (1991) *Automated lip-sync: background and techniques*, The Journal of Visualization and Computer Animation, Vol. 2, pp. 119-122.

Luo, S. H. (1994) Speech-Enhanced Model-Based Video Facial Image Coding, Ph. D. thesis to be submitted, Department of Electrical Engineering, The University of Sydney.

Luo, S.H and King, R.W. (1994), *A Novel Approach for Classifying Continuous Speech into Visible Mouth-Shape Related Classes*, Proc. ICASSP -94, Vol. 1, pp. 465-468.

Massaro, D. W. (1987) Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry, Lawrence Erlbaum Associates, Publishers, London.

Massaro, D.W. and Cohen, M.M.(1994) *Auditory/Visual Speech in Multimodal Interfaces*, Proc. ICSLP-94, pp. 531- 534.

Petajan, E. D. (1987) *An improved automatic lipreading system to enhance speech recognition*, Bell Labs. Tech. Report, No. 11251-871012-111TM.

Rumelhart, D. E., Hinton, G.E. & Williams, R.J. (1986) *Learning internal representations by error propagation Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Chapter 8. Bradford Books/MIT press, Cambridge, Massachusetts.

Sejnowski, T. J., Yuhas, B. P., Jenkins, R. E. (1989) *Combining visual and acoustic speech signals with a neural network improves intelligibility*, Advances in Neural Information Processing Systems, Edited by D. S. Touretzky, K. Kaufmann Publisher, San Mateo, CA, pp. 232-239.

Sirovich, L., Kirby, M. (1987) *Low-dimensional procedure for the characterization of human faces*, Journal of the Optical Society of America, Vol. 4, No. 3, Mar., pp. 519-524.

Werbos, P. J. (1990) *Backpropagation through time: what it does and how to do it*, Proc. of IEEE, Vol.78, No.10, Oct., pp. 1550-1560.