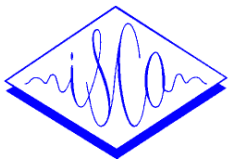


Auditory-Visual Speech Processing
(AVSP'99)
Santa Cruz, CA, USA
August 7-10, 1999

AUTOMATIC COMPUTER LIP-READING USING FUZZY SET THEORY

James F.Baldwin, Trevor P.Martin, Mehreen Saeed

Artificial Intelligence Research Group, Department of Engineering Mathematics
University of Bristol, UK
Email: (Jim.Baldwin,Trevor.Martin,Mehreen.Saeed)@bristol.ac.uk

ABSTRACT

This paper presents the application of fuzzy set theory to automatic computer lip-reading from video images. Simple rules based on fuzzy sets were generated using the mass assignment theory and were used for automatic feature extraction from video sequences. Probabilistic grid models were used to derive a knowledge base representing the visual data for phonemes or sounds. Phonemes from a medium sized vocabulary of words were used for training and testing and a reasonable accuracy for classification was achieved. The methods were also applied to the Tulips1 database and the results illustrate that the learning techniques are efficient and general enough to be applied to different speakers.

1. INTRODUCTION & MOTIVATION

Automatic computer lip-reading is becoming an increasingly popular area of research amongst the speech recognition community. The growing interest arises because the information regarding speech contained in visual signals is both supplementary and complementary to the information contained in audio signals, especially in the presence of noise [13]. Hence using visual signals in addition to audio signals has been shown to enhance the accuracy of speech recognition programs [11,12] to a considerable extent. In addition to this speechreading can find its use in many other areas such as building aids for the deaf, video conferencing etc.

The aim of the undertaken project is to develop a system for automatic computer lip-reading using image sequences of the lip movements of a speaker without the use of sound. The methods employed are based on learning using fuzzy set theory [4]. Two different databases have been selected on which training and testing have been performed. The first database consists of a medium sized vocabulary of words uttered by one person. The words have been segmented manually on phoneme boundaries and phoneme classification has been performed on this database. To illustrate that the methods are speaker independent, a second database, the Tulips1 database [10], which consists of four digits spoken by 12 different speakers, has been taken into consideration.

The purpose of this paper is to show the effectiveness of the methods employing fuzzy set theory and to demonstrate their linguistic qualities. In the past neural networks or Hidden Markov Models have been used for visual speech recognition, but these methods are black box techniques which are difficult to understand. The use of fuzzy set theory however makes the speech reading techniques simple and easy to comprehend. They also provide a means of representing data with simple linguistic terms, hence making the methods glass box methods. Moreover the speech data are represented by means of fuzzy sets which are generated by taking into account the probability distribution of data. This methodology of knowledge representation is not only intuitive but also a compact way of explaining the nature and the trend of data.

2. DATA & FEATURES

The following features have been used for this application and are illustrated in Figure 1.

1. Width of lips
2. Height of lips
3. Width of mouth
4. Height of mouth
5. Height of lower lip
6. Height of upper lip

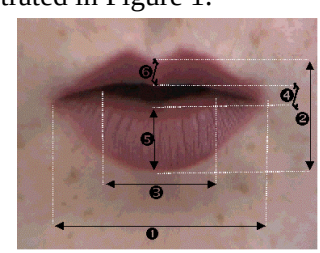


Figure 1: Features used

Additional features, which are not always visible, have also been taken into account:

- Height of tongue below upper lip
- Height of tongue above lower lip
- Height of tongue between teeth
- Height of upper teeth and lower teeth
-

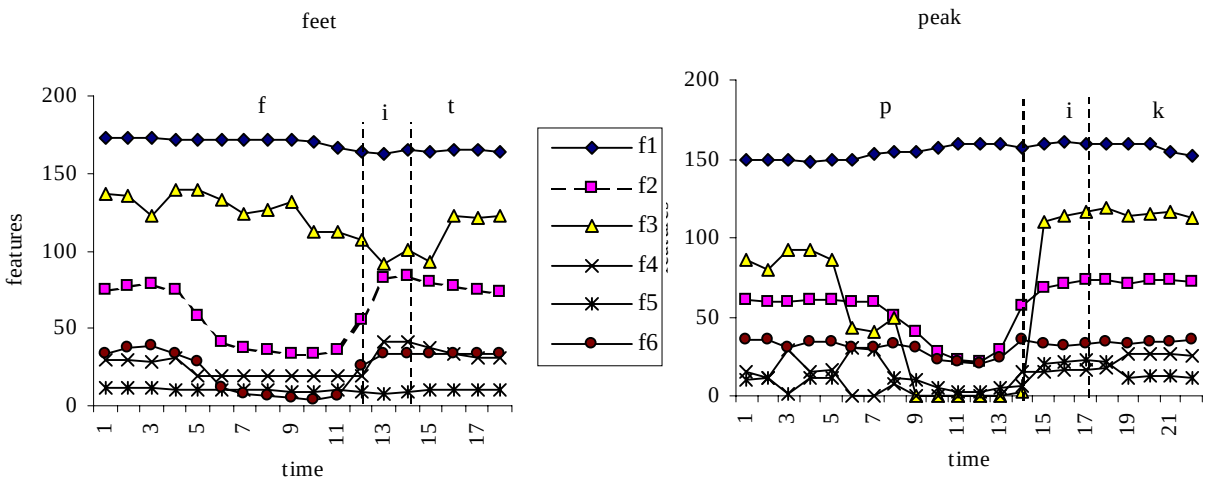


Figure 2: Plot of feature values against time. The phoneme boundaries have been marked by hand

2.1. Feature Extraction from Data

The features described in Section 2 are extracted automatically from the video sequences using the method described in [5]. Baldwin *et al.* have developed a lip detection system for their work on facial feature extraction. A brief overview of the techniques employed to obtain the various lip measurements follows.

A database of the normalized red, green, and blue pixel values of the lips, skin and teeth images is obtained from a small training set. This database is used to obtain a probability distribution of the normalized RGB values over the 3 classes, namely lips, skin, and teeth. The probability distributions are then converted into their corresponding fuzzy sets using the theory of mass assignments [2]. Thus 3 fuzzy sets representing the normalized RGB pixel values are obtained for each of the 3 afore mentioned classes. This leads to the formation of simple Fril rules [4] describing each class (Fril is a logic programming language incorporating fuzzy set theory). A simple Fril rule for one of the skin, lip and teeth classes has the following form:

Pixel is Class$_i$ if
R_n value of pixel is FuzzySet$_{Red}$
and
G_n value of pixel is FuzzySet$_{Green}$
and
B_n value of pixel is FuzzySet$_{Blue}$

where R_n, G_n, B_n represent the normalized red, green and blue pixel transform values respectively. For example R_n (B_n and G_n can be similarly obtained) is given by:

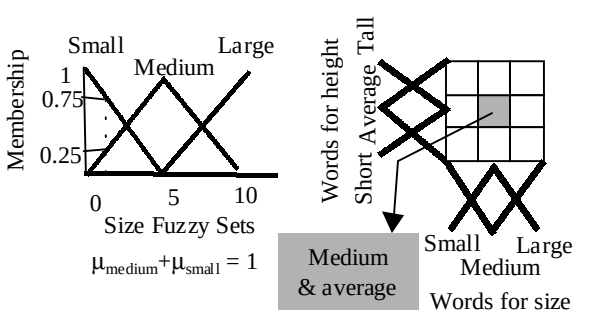
$$R_n = Red/(Red+Green+Blue)$$

The simple Fril rules generated from the training set are used for feature extraction in the database of images. Every pixel in an image is classified individually as lip, skin or teeth. The classified lip and teeth pixels are then checked to see if they lie within the permissible range of intensity values for lips and teeth. Hence an image with clusters of lips, teeth and skin pixels is obtained. Additional searching algorithms are added on top of these classification routines to find the corners of the lips and mouth, and the measurements of the 11 features described in Section 2 are obtained. The plots of the first 6 feature values, presented in Section 2, are displayed in Figure 2 for the words 'feet' and 'peak'. The phoneme boundaries have been marked by hand and are also shown in the graphs.

3.MODELING VISUAL SPEECH

Before a discussion of the technique used in modeling visual speech data is carried out, a brief overview of modeling compound words in Cartesian space and generating extended Fril rules [4] is given.

3.1. Grid Built on Fuzzy Granules in Cartesian Space and Extended Fril Rules



a) *Mutually exclusive fuzzy sets*

b) *Words built on fuzzy granules*

Figure 3: Illustration of mutually exclusive fuzzy sets and words built on fuzzy granules.

In this section a method based on words modeled by fuzzy sets is described to represent a knowledge base. A word can be represented by a fuzzy set of points representing a clump of elements drawn together by similarity [6]. Figure 3a shows 3 fuzzy sets representing the words 'small,' 'medium,' and 'large.' Moreover they are mutually exclusive triangular fuzzy sets having the characteristic that the membership of a point in all the fuzzy sets adds to one. Any point on the axis has a nonzero membership in, at the most, two fuzzy sets. Each triangular fuzzy set is thus a word (granule or a label). This concept is extended in multidimensions. A two-dimensional grid in Cartesian space with mutually exclusive fuzzy sets on its two axes has each cell modeling a compound word or the cross product of two linguistic labels. So, for example, in Figure 3b the shaded cell represents the situation when the height feature is *average* and the size feature is *medium*.

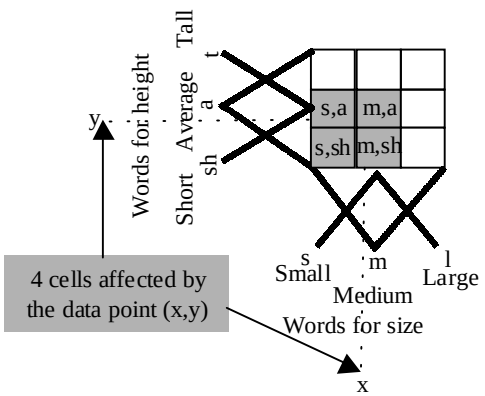


Figure 4: Counting procedure for the data point (x,y)

A simple counting procedure is adopted for constructing the grid from example points in the data set. A data point *(x,y)* shown in Figure 4 can be written as:

$$(x = \frac{f_s}{\mu_s} + \frac{f_m}{\mu_m}, y = \frac{f_{sh}}{\mu_{sh}} + \frac{f_a}{\mu_a})$$

where μ_i is the membership of the point in fuzzy set f_i. According to the voting model [3] and the theory of mass assignments [2], the least prejudiced distribution $\text{lpd}_f(x)$ of a point in one of the mutually exclusive fuzzy sets f is equal to its membership in that fuzzy set. Since the least prejudiced distribution represents the conditional probability of the point given that fuzzy set [6], hence the 4 pair of values

$$\{\text{lpd}_s(x)\,\text{lpd}_{sh}(y); \text{lpd}_m(x)\,\text{lpd}_{sh}(y)$$
$$\text{lpd}_s(x)\,\text{lpd}_a(y); \text{lpd}_m(x)\,\text{lpd}_a(y)\}$$

are equivalent to

$$\{\mu_s\mu_{sh}; \mu_m\mu_{sh}; \mu_s\mu_a; \mu_m\mu_a\}$$

and can be associated with their corresponding cells {s,sh; m,sh; s,a; m,a} respectively in Figure 4. Thus, for a given data tuple, the membership of feature 1 and feature 2 in their corresponding fuzzy sets is determined and the relevant cell is incremented by the product of memberships in the two fuzzy sets. Clearly there is an advantage of using fuzzy granules to crisp sets. If crisp sets are used then the membership of only one cell is affected whereas in this case several cells are affected depending upon their membership in various words. This phenomenon is illustrated in Figure 4 where the point (x,y) affects the four cells {s,sh; m,sh; s,a; m,a}.

After filling each cell of the grid in Cartesian space, each cell in the grid is divided by the total number of entries and a probability distribution θ_{ik} over the cells is obtained for a given $class_k$. Rules based on Bayesian theory are used to derive conditional probabilities. So the conditional probability of a $class_k$ given $cell_i$ i.e. $\Pr(class_k/cell_i)$ is given by:

$$P_{ki} = \frac{\Pr(cell_i / class_k)\Pr(Class_k)}{\Pr(cell_i)}$$
$$= \frac{\theta_{ik}\Pr(class_k)}{\sum_j \Pr(cell_i / class_j)\Pr(class_j)}$$
$$= \frac{\theta_{ik}\Pr(class_k)}{\sum_j \theta_{ij}\Pr(class_j)}$$

Assuming that all classes are equally likely the above form reduces to:

$$\Pr(class_k / cell_i) = \frac{\theta_{ik}}{\sum_j \theta_{ij}}$$

When an example point from the test set is encountered then a test grid is formed and the probability of each cell e_i is determined for the features in the test data point. The $support_{jk}$ for $class_j$ for a single grid which represents a single feature k is given by:

$$sup\ port_{jk} = \sum_i e_i\, Pr(class_j / cell_i)$$

The over all support for $class_j$ using m grids representing m different features is averaged as:

$$sup\ port_j = \frac{\sum_{i=1}^{m} sup\ port_{ji}}{m}$$

When classification is performed then the class with the highest support is the predicted class.

3.2. Generating Rules for Speech Data

After successful feature extraction, data for a word or a phoneme is obtained as a plot of feature values against time. Those data are composed of sequences of different lengths. To model the temporal characteristics of each feature, one axis on the grid, described in Section 3.1, represents the feature space and the second axis represents time. The values on the time axis range from the first time frame to the last frame comprising the sound. The number of mutually exclusive fuzzy sets placed on the time axis remains the same for every sound or word and hence they are made broader or narrower depending upon the number of sequences comprising the word. At each time step the membership of a feature in its corresponding feature fuzzy sets and time fuzzy sets is determined and a grid on the cross product space of words representing features and time is obtained. It has very simple semantic meanings and an example is illustrated in Figure 5. The figure depicts the following situation:

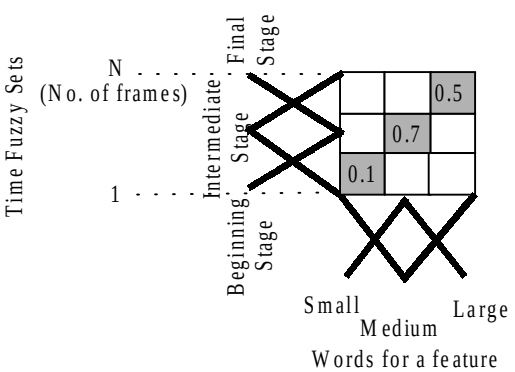


Figure 5: Modelling speech data

- Feature value is S*mall* in the *Beginning Stage* with a probability of 0.1
- Feature value is *Medium* in the *Intermediate Stage* with a probability of 0.7
- Feature value is *Large* in the *Final Stage* with a probability of 0.5

4. DATABASE & RESULTS

The rules described in Section 3 have been generated for and tested on two different types of databases. They are described in detail below.

4.1.Phoneme Based Classification and Results

The database used for phoneme classification has been formed from the video of a single person saying 310 words in a row. The speaker does not have a strong accent in English. The video has been taken at 25 frames per second. The length of the video sequences for each word ranges from 11 to 37 frames. There are around 6000 colored images, 250x160 pixels in size, occupying about 720M bytes of disk space.

All the words of the vocabulary have been segmented manually on phoneme boundaries. A total of 44 phonemes consisting of 25 consonants and 19 vowels and diphthongs have been used. All phonemes with a similar lip movement have been grouped into visemes. A viseme is comprised of phonemes having the same lip movement and are visibly indistinguishable, e.g. p/b/m or f/v, etc. The placing of phonemes into various visemes has been carried out using [8] with some modifications according to the speaker of the video sequences. They are shown in Appendix A.

	Plain		**Position Information**	
	Train	**Test**	**Train**	**Test**
Consonants	57.68%	51.09%	73.78%	66.15%
Vowels	70%	54.58%	70%	55%
Overall	61.92%	52.29%	72.48%	62.32%

Table 1: Percentage accuracy for consonants and vowels

Out of the 310 spoken words, 116 words have been placed in the training set and the rest have been used in the test set. Each phoneme in the manually segmented words of the test set has to be classified into its corresponding viseme group. Table 1 presents the results obtained on consonants and vowels. The table shows two types of results. The first two columns under the *Plain* results illustrate the accuracy of phonemes when no position information was used for their classification. For the last two columns under *Position Information*, the consonants and vowels were grouped together based upon the information of their position in a word. The phonemes such as s/z or t/d/n/l have a more prominent movement when they occur in the beginning of a word as compared to when they

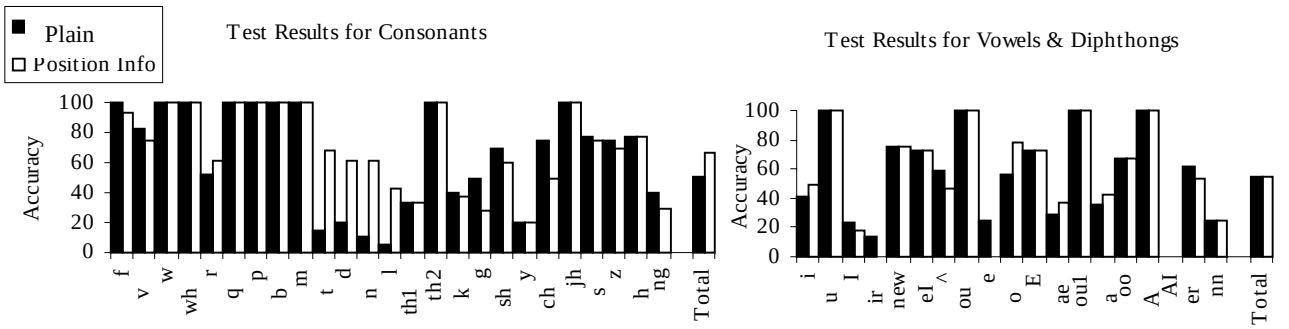


Figure 6: Detailed results for various phonemes. (See Appendix A for an explanation of symbols)

occur in the middle of the word. Clearly the accuracy for consonant classification is improved considerably when information about their position inside a word is used.

Figure 6 illustrates the detailed results obtained for consonants and vowels. The consonants such as p/b/m or f/v have a prominent lip movement and therefore have a high accuracy of classification even without using contextual information. On the other hand, for consonants whose lip movement is very subtle (e.g. r/t/d), using the information about their position within a word increases their chances of falling into the right group considerably.

It is difficult to have a means of comparison for the results obtained on the above database since there is no other known lip-reading system to use the same video sequences. But the results can be compared to expert human lip-readers who are estimated to achieve 30% accuracy on nonsense words [7].

4.2. Tulips1 Database

The Tulips1 Database has been formed from 12 speakers, 9 male and 3 female, saying the words 'one', 'two', 'three,' and 'four' twice. There are 934 gray-scale images of 100x75 dimensions taken at 30 frames per second. The audio signals included in the database have not been taken into account. For these images the corners of the mouth and lips have been marked by hand and the changes between each successive frame for the first six features described in Section 2 have been extracted. Training has been performed by generating rules from 11 speakers and leaving one out for testing. The process is repeated by including each speaker in the test set and the results are averaged over all speakers.

Feature fuzzy sets	7	7	6
Time fuzzy sets	2	3	2
Percentage Accuracy	91.67%	91.67%	92.71%

Table 2: Various results for Tulips1 database using extended Fril rules

Four extended Fril rules have been generated for each word 'one,' 'two,' 'three,' and 'four'. Viseme grouping is not required in this case because rules have been generated on whole words. The results obtained for different parameters are shown in the Table 2. The results in the table illustrate the generalizing capabilities of the extended Fril rules. The rules are speaker independent and hence robust enough to handle different speakers with different ethnic origins. The following table shows the results obtained for individual speakers:

Speaker	1	2	3	4	5	6
Out of 8	8	6	7	8	8	7
Accuracy %	100	75	87	100	100	87
Speaker	**7**	**8**	**9**	**10**	**11**	**12**
Out of 8	8	8	7	6	8	8
Accuracy %	100	100	87	75	100	100

Table 3: Results obtained for the individual speakers

The results obtained are comparable to Luettin and Thacker [9] who get an accuracy of 90.6% on this database by training Hidden Markov Models on the 5 most discriminant features representing shape and intensity and also their delta parameters. Humans with no lip-reading knowledge achieved an average of 89.93% on this database while hearing-impaired people with knowledge of lip-reading obtained 95.49% on this database [10]. The comparison is shown in Table 4.

Lip-reading by…	Accuracy
Extended Fril rules	92.71%
Luettin and Thacker, 1997 [9]	90.6%
Humans with no lip-reading knowledge	89.93%
Humans with lip-reading knowledge	95.49%

Table 4: Comparison for Tulips1 Database

5. SUMMARY & DISCUSSION

In this paper a general discussion of the application of fuzzy set theory to automatic computer lip-reading i.e. speech recognition without the use of

audio, has been carried out. A cross product of linguistic variables representing time and feature space has been used to model the lip movements of sounds defining whole words or phonemes in visual speech. The techniques employed are efficient and not computationally demanding. Moreover their linguistic nature makes them simple and easy to understand. Further work is being carried out on the automatic segmentation of words on phoneme boundaries and performing phoneme-based isolated word recognition using only the visual signals.

5. REFERENCES

1. Patricia Ashby. *Speech Sounds.* T J Press (Padstow) Ltd, Padstow, Cornwall, UK, 1995.

2. James F. Baldwin. A theory of mass assignments for artificial intelligence. In *Lecture notes in artificial intelligence*, pages 22-34, Springer-Verlag, Sydney, Australia, August 1991.

3. James F. Baldwin. The management of fuzzy and probabilistic uncertainties for knowledge based systems. *Encyclopedia of AI*, editor: S.A. Shapiro, pages 528-537, John Wiley (2nd ed.), 1992.

4. James F. Baldwin, Trevor P. Martin, Bruce W. Pilsworth. *FRIL-Fuzzy and Evidential Reasoning in Artificial Intelligence.* Research Studies Press (Wiley Inc.), 1995.

5. James F. Baldwin, Simon J. Case,Trevor P. Martin. Machine interpretation of facial expressions. *BT Technology Journal*, 16(3):156-164, 1998.

6. James F. Baldwin. Logic programming with uncertainty and computing with words. *Logic Programming and Soft Computing*, editors: Trevor P.Martin and F.A.Fontana, pages:19-48, RSP/Wiley, 1998.

7. Alan J. Goldschen. *Continuous Automatic speech recognition by lipreading*. PhD thesis, George Washington University, Washington, D. C., 1993.

8. Janet Jeffers and Margaret Barley. *Speechreading (Lipreading).* Charles C Thomas Publisher, Springfield, Illinois, U.S.A., 1971.

9. Juergen Luettin and Neil A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163--178, February 1997.

10. Javier R. Movellan. Visual speech recognition with stochastic networks. *Advances in Neural Information Processing Systems*, editors: G. Tesauro, D. Toruetzky and T. Leen (eds.), Vol 7, MIT Pess, Cambridge, 1995.

11. Eric Petajan, Bradford Bischoff, David Bodoff, N.Michael Brooke. An improved automatic lipreading system to enhance speech recognition. *ACM SIGCHI-88*, pages 19--25, 1988.

12. Peter L. Silsbee and Alan C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337--351, September, 1996.

13. Quentin Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. *Hearing By Eye: The Psychology of Lip-reading*, editors: Barbara Dodd and Ruth Campbell, pages 3--51, Lawrence Erlbaum Associates Ltd., London, 1987.

APPENDIX A

The viseme grouping used for the results obtained in section 4.1 are shown in table A. The '*' denotes phonemes that occur in the beginning of a word and '&' denotes phonemes that occur at the end e.g. r* denotes r occurring at the beginning of a word and r& means r occurring in a position other than the beginning of a word.

Plain		**Position Information**	
f, v	I, ^, e	f, v	w, wh, r*, q
ch, jh, sh, y	i, eI, E	ch, jh, sh, y	(t, d, n, l, s, z)&
p, b, m	ir	r&	i eI, E
t, d, n, l	ae, a	p, b, m	I, ^, e
th1, th2	AI	(t, d)*	U, u, ou, o, ou1, oo, A
k, g, h, ng	er, nn	(n, l)*	ir
w, wh, r, q		(s z)*	ae, a
s, z		th1, th2	AI
U, u, ou, o, ou1, oo, A		k g h nk	er, nn

Table A1: The viseme table used for the results shown in section 4.1.

Most of the above symbols correspond to the English alphabet, however a guide to the symbols with which the reader may not be familiar with is given below:

A p*o*t	eI b*ai*t	ir b*ir*d	o o-bey	th2 *th*eta	q *q*uite
AI b*y*	er h*er*	jh *j*ud*ge*	oo B*o*b	U b*oo*t	e cha-otic
ae b*a*t	a d*a*nce	ng si*ng*	ou1 l*aw*	u b*oo*k	wh *wh*ich
E b*e*t	I b*i*t	nn si*nk*	sh *sh*e	y *y*es	
ch *ch*air	i f*ee*t	ou g*o*	th1 *th*is	^ p*u*n	

Table A2: A guide to the symbols used in this paper

ACKNOWLEDGMENTS

The authors would like to thank the Association of Commonwealth Universities, London for their support.