Speech Data Analysis and Recognition Using Fuzzy Neural Networks and Self-Organising Maps

N. Kasabov, R. Kozma, R. Kilgour, M. Laws, M. Watts, A. Gray, J. Taylor

Department of Information Science
University of Otago, P.O Box 56, Dunedin, New Zealand
Phone: +64 3 479 8319, fax: +64 3 479 8311
nkasabov@otago.ac.nz

Summary. This paper presents results from a research project on speech data analysis and speaker independent adaptive speech recognition using novel neuro-fuzzy techniques and a new system architecture. A speech recognition system of English is presented in the chapter that utilises fuzzy neural networks and self-organising spatial-temporal maps. Experimental results on speech recognition are given.

Key words: hybrid intelligent information systems, speech recognition, cognitive engineering, fuzzy neural networks.

Introduction

Speech recognition is an extremely difficult engineering task to be performed by a computer system because the variability in the way people speak, which is reflected in tremendously complex speech signals to be processed in the automatic speech recognition systems (ASRS). There are several key areas of future research which have been pointed out in [3] as significant for the future development of the spoken language systems. These include robust speech recognition; automatic training and adaptation; spontaneous speech; dialogue models; natural language response generation; speech synthesis and speech generation; multi-lingual systems; interactive multi-modal systems.

We take the view that these goals can be achieved if an integrated approach is used, where everything which is known about the speech and the language recognition task should be brought together and used in one system [10, 13, 14], for example: speech corpus data, phonemic knowledge, linguistic knowledge, skills from pedagogy and teaching languages. Advanced knowledge engineering techniques are used to facilitate this integrated approach.

The task of speech recognition becomes even more complicated when the ASR system is used as part of an intelligent human computer interface that allows for retrieving information from a database or for connecting to other communication ports by using both speech and text.

This paper presents a novel approach to adaptive speech recognition that utilises fuzzy neural networks and spatial-temporal self-organising maps.

This approach is illustrated on a hybrid speech recognition system called HySpeech. First, the issue of speech data analysis is discussed, then the neuro-fuzzy methods for modelling and recognition of speech data are presented. Finally some applications and experimental results are given.

1. Speech Data Variability and Analysis

1.1 Variability in spoken language

Although the acoustic signal contains a high degree of redundancy (multiple cueing of phoneme segments, temporal spreading of cues), numerous factors conspire to render the signal potentially ambiguous. Even in slow careful speech, the cues for any given phoneme are liable to vary according to: (a) phonemic environment, (b) position vis-à-vis syllable, word, and utterance boundaries, (c) prosodic factors (for example, stress), and (d) speaker-dependent aspects (for example, gender and accent). To resolve ambiguity it is necessary to call upon higher-order linguistic knowledge (phonological, lexical, syntactic, semantic), and the expectations that higher-order knowledge generates.

This project deals with New Zealand English which is undergoing rapid change (especially in its vowel system), and is becoming increasingly distinct from other varieties of English. Changes currently underway affect (a) the vowel inventory, (b) the location of the vowels in formant space, and (c) the distribution of the vowels. The data used for training the recognition models is organised in the Otago Speech Corpus as described below.

1.2 The Otago Speech Database of New Zealand English

The Otago Speech Corpus [27] contains 37 speakers uttering selected isolated words chosen to represent allophonic realisations of the 43 NZ phonemes. The data was recorded at a sample rate of 22050Hz and a resolution of 16 bits. Manual segmentation was used to extract selected phonemes from either the initial, medial or final position. When a speech file is transformed into the Mel Scale Coefficients, it is saved as three time steps of 26 element MSC vectors merged to make up the 78 MSC vectors required for input to the recognition units, below.

The HySpeech system described in the next section is used for several applications one of them being English-to-Māori word and phrases translation and pronunciation. For this reason a Māori speech database is created based on the same NZ English corpus construction, where new and existing Māori speakers/words will have similar file formats and codes. There are currently 17 speakers of Māori in the corpus. The NZ English speech corpus is available free on the WWW from: "http://kel.otago.ac.nz/hyspeech/corpus/"

1.3 Statistical methods for speech data analysis. Linear discriminant analysis

Multiple discriminant analysis (MDA) is a statistical technique that separates observations into two or more groups based on orthogonal linear functions of the independent variables. The technique assumes a reasonable degree of multivariate normality, with logistic regression an alternative where this is not the case. When working with speech data, two methods for recognising phonemes using MDA are available; firstly to develop a single model with multiple discriminant functions, and secondly to follow the modular approach of using separate discriminant functions for each phoneme. For the speech data being analysed here the two approaches were both used with the results presented in Table 1.1. This table shows the performance of the two sets of models using 14869 observations for model development, and 4955 observations for validation. The phoneme selected for the multiple models case was based on a simple highest probability vote scheme. As can be seen the use of a single model provides better results in this case, as would be expected.

Table 1.1. Results for the MDA models

	Multiple MDAs	Single MDA
Average correct positives on validation set	74.8%	–
Average correct negatives on validation set	87.9%	–
Overall correct on validation set	34.9%	38.3%

The advantages of using a technique such as MDA include that the models are well founded, with the outputs providing the probabilities of group membership. This allows for the use of Bayesian statistics where probabilities such as P(A—B) can be used. For example, the probability that the phoneme is in fact number 26 (A) given that the discriminant function classifies it as 26 (B). Alternatively, the probability that the phoneme will be classified as 26 (A) given that it is in fact phoneme 27 (B) can be calculated. Such an approach also allows the analyst to combine other forms of linguistic knowledge, classifications from the same utterance, and other features of the speech signal in a meaningful manner using Bayesian rules of evidence. The FuNN models could also be used in this way to arrive at the final, most likely, classification.

The initial results for the models based on the fuzzy neural networks (FuNN), as presented further in the paper, are very similar to those produced by the MDA. In general, the confusion matrices for these two approaches are almost identical, with the FuNNs providing slightly better average correct positives and negatives but with slightly lower overall correct classifications

after voting. This suggests that the level of accuracy achieved by FuNN models on the individual phonemes is in fact as high as can be expected, with the remaining errors due to the inherent noisiness contained in the data. One of the advantages of the FuNN-based phoneme modelling architecture is that a separate FuNN-model is used for each of the 45 phonemes in English. This allows for individual tuning of the models which ultimately leads to a significant improvement in the classification rate. The FuNN model allows also for adaptation to new speakers, for rule extraction capturing the main phonetic characteristics of the speech data in the form of IF-THEN rules and for a possible extension of the speech recognition system to multi-lingual one [14].

2. The Principles and the General Architecture of a Hybrid Speech Recognition System HySpeech

The block diagram of the HySpeech framework is given in Figure 2.1. It consists of the following modules: speech pre-processing; modular phoneme recognition; word mapping; answer formation.

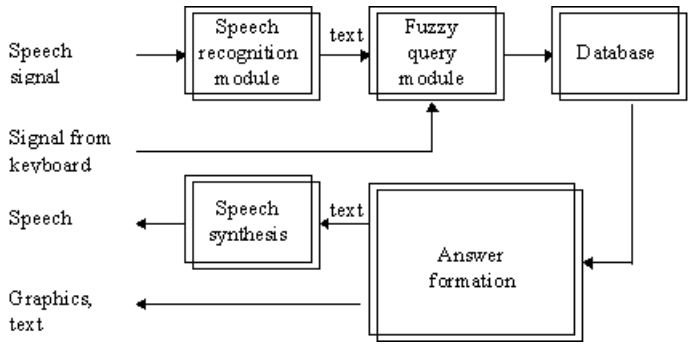

Fig. 2.1. A block-diagram of HySpeech

The speech pre-processing module does the transformation of the raw speech signal into Mel-scale coefficients (MSC), where each time frame of 5.8ms of speech data is transformed into a 26-element vector, with an overlapping of 50%. Three time-lags is thus used to represent approximately 11.6ms of speech. The phoneme recognition module utilises one fuzzy neural network (FuNN) [12, 16] for each of the abstract elementary sounds (phonemes) in the language. FuNNs are useful tools for learning and adaptation from speech data. A novel algorithm for FuNN training with forgetting allows the FuNN to structurally capture the main characteristics of the speech signal. Brief description of FuNNs, their learning and adaptation algorithms and applications for phoneme recognition are given in Section 3. Using genetic algorithms for adaptation in the FuNN system is discussed further in the paper.

The word mapping module uses a novel architecture called fuzzy spatial-temporal maps (FuST). These are an extension of the well established SOMs [22]. They are used here to store the words of the speech dictionary in a 'sounds-like' map for fast retrieval and quick updating to include new words in the dictionary. The major principles of FuST and their application for word mapping are presented further in the paper.

The HySpeech framework can be applied to any spoken language recognition. Before applying it to a particular language, in our case English, the available speech data and linguistic, phonetic knowledge must be analysed in terms of language features, basic characteristics of the elementary sounds, similarity and differences between the phonemes as discussed in the previous section.

The HySpeech architecture is used in several application-oriented projects one of them being English to Māori talking dictionary, an English word/phrase is recognised and the Māori equivalent is returned to the speaker as both text and speech.

3. Fuzzy Neural Networks - A General Architecture and Applications for Phoneme Data Analysis and Classification

3.1 The FuNN Architecture and its Functionality

Fuzzy neural networks are neural networks that realise a set of fuzzy rules and a fuzzy inference machine in a connectionist way [29, 6, 8, 13].

FuNN is a fuzzy neural network introduced first in [13] and then developed in its FuNN/2 version [16]. It is a connectionist feed-forward architecture with five layers of neurons and four layers of connections. The first layer of neurons receives the input information. The second layer calculates the fuzzy membership degrees to which the input values belong to predefined fuzzy membership functions; For example, small, medium, large. The third layer of neurons represents associations between the input and the output variables, fuzzy rules. The forth layer calculates the degrees to which output membership functions are matched by the input data and the fifth layer does defuzzification and calculates values for the output variables. A FuNN has both the features of a neural network and a fuzzy inference machine. A simple FuNN structure is shown in Figure 3.1. The number of neurons in each of the layers can change during operation through growing or shrinking. The number of connections is also modifiable through learning with forgetting, zeroing, pruning and other operations [12, 13, 17, 18].

The membership functions, used in FuNN to represent fuzzy values, are of triangular type, the centres of the triangles being attached as weights to the corresponding connections. The membership functions can be modified through learning.

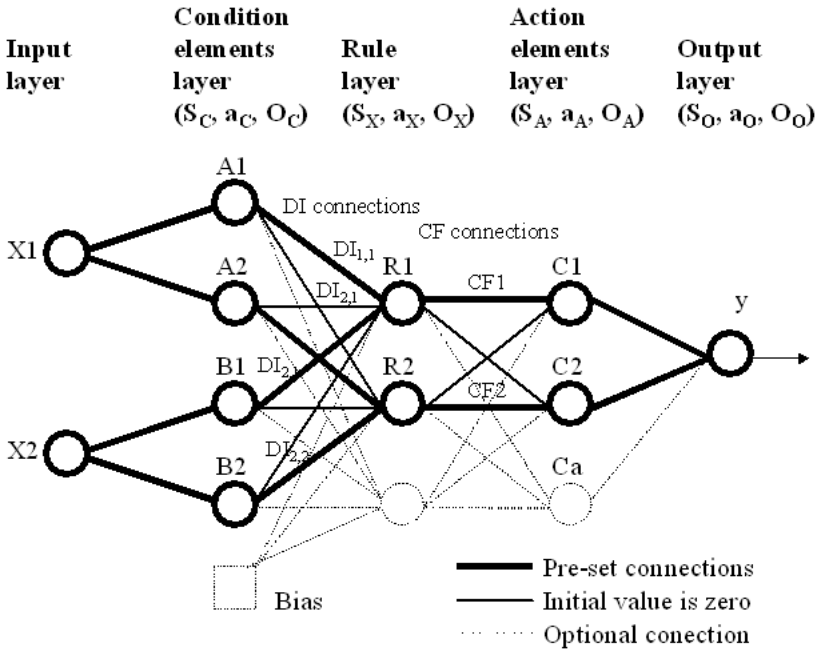

Fig. 3.1. A FuNN structure for two initial fuzzy rules: R1: IF x1 is A1 (DI1,1) and x2 is B1 (DI2,1) THEN y is C1 (CF1); R2: IF x1 is A2 (DI1,2) and x2 is B2 (DI2,2) THEN y is C2 (CF2), where DIs are degrees of importance attached to the condition elements and CFs are confidence factors attached to the consequent parts of the rules (adopted from [14]). The triplets (s,a,o) represent specific for the layer summation, activation, and output functions

Several training algorithms have been developed for FuNN:

(a) A modified back-propagation (BP) algorithm that does not change the input and the output connections representing the membership functions.
(b) A modified BP algorithm that utilises structural learning with forgetting. That is, a small forgetting ingredient, for example 10-5, is used when the connection weights are updated (see [7, 24, 17]).
(c) A modified BP algorithm that updates both the inner connection layers and the membership layers. This is possible when the derivatives are calculated separately for the two parts of the triangular functions which are also the activation functions of neurons in the condition element layer. The triangular membership functions can be considered as non-monotonic activation functions.
(d) A genetic algorithm for training [18].
(e) A combination of any of the methods above used in different time intervals as part of a single training procedure.

Several algorithms for rule extraction from FuNN have been developed and applied [12]. One of them represents each rule node of a trained FuNN as an IF-THEN fuzzy rule as shown in Figure 3.1.

FuNNs have several advantages when compared with the traditional connectionist systems or with the fuzzy systems:

(a) They are both a statistical and a knowledge engineering means.
(b) They are robust to catastrophic forgetting. That is, when further trained only on new data, they keep a reasonable memory of the old data.
(c) They interpolate and extrapolate well in regions where data is sparse.
(d) They can be used as replicators, where same input data is used as output data during training; in this case the rule nodes perform an optimal encoding of the input space.
(e) They are appropriate tools to build multi-modular IIS as explained next.

3.2 Using FuNNs for phoneme recognition

Fuzzy neural networks have been used so far for tasks from speech recognition. Some of the experiments use a large, single, neural network for classifying phonemes on their formant input values [25, 26] and to extract fuzzy rules. Other experiments use hybrid neuro-fuzzy systems in a modular approach [11, 12, 13, 15, 19, 20].

Here FuNNs are used to learn and classify single phonemes. Three 26-element Mel-scale coefficients (MSC) vectors, representing the speech signal at three consecutive time frames of approximately 12ms each, are used as initial inputs. Through training with forgetting, each FuNN unit is tailored to the specific phoneme. After the training procedure and a consecutive pruning of the very small connections, only the important inputs that correspond to significant for the phoneme time-lags, and the important MSC, are kept in

the FuNN structure. This is illustrated on Figure 3.2. A FuNN structure is initialised as 78-234-10-2-1 and then trained with forgetting on both male and female data of the phoneme /e/, as positive data and the rest of the phoneme data as negative data. The training and testing data is taken from 139 words pronounced three times each by 3 male and three female speakers. The FuNN structure has been significantly simplified through training with forgetting and a consequent pruning. As it can be seen from the third figure of Figure 3.3, only several rule nodes remain. The condition element nodes and the left connections from them to the rule nodes, correspond to the main frequency of the phoneme /e/ realisation as shown on the top of Figure 3.2. The bright areas there show high energy of the signal for a particular MSC. It can also be seen that more connections from the first time-lag input vector are left which suggests a higher importance of this time-lag. The trained /e/ FuNN, when tested on new data, showed correct true positive and true negative activation (the bottom figure on Figure 3.2).

Fig. 3.2. A FuNN used to classify the phoneme /e/ from 8 male and 8 female speech data: (a) A FuNN trained to classify phoneme /e/ data without forgetting; (b) The same FuNN trained with forgetting; (c) Test accuracy of recognition

3.3 FuNN-based Intelligent Multi-modular Systems.

HySpeech is based on a more general architecture for FuNN-based IIS. It consists of a FuNN-based module which includes single FuNN-units for each class (elementary event, patter, etc.), a module for rule extraction and explanation, and a module for adaptation. Here, adaptation is the process of on-line training when a single FuNN improves its performance based on observation and analysis how its performance compares to the reality.

Adaptation in a multi-modular FuNN structure is based on individual tuning of single FuNN-units if the analysis of the performance of the whole system shows that those are reasons for unsatisfactory performance or points of improvement. One scheme for adaptation, for example, is where a copy of a FuNN is trained on-line to improve its performance while the main FuNN-unit is in operation. After a certain time interval the copy substitutes the main FuNN unit.

3.4 Multi-modular FuNN-based Systems for All-Phoneme Classification

The overall test accuracy of individual FuNNs, trained with 6 male and 6 female speakers data from the Otago Speech Corpus, is shown in Table 3.1. The testing set consisted of 3 Male and 3 female speakers that the system had not previously encountered.

Table 3.1. Test accuracy of the phoneme recognition in the FuNN units across all the phonemes in New Zealand English.

Phoneme Number	Correct Positive	Correct Negative	Phoneme Number	Correct Positive	Correct Negative
1	82.5%	97.7%	25	69.4%	91.5%
2	66.7%	96.8%	26	87.9%	96.1%
3	97.0%	85.5%	27	67.4%	95.2%
4	69.2%	95.2%	28	80.0%	88.9%
5	94.2%	88.9%	29	76.7%	94.2%
6	92.3%	88.4%	30	54.2%	92.0%
7	98.0%	92.6%	31	61.6%	87.4%
8	94.3%	83.9%	32	85.5%	92.4%
9	98.4%	90.8%	33	76.1%	92.3%
10	86.8%	88.3%	34	68.8%	92.4%
11	92.7%	98.1%	35	70.3%	92.6%
12	86.3%	94.0%	36	69.8%	83.5%
13	92.4%	97.4%	37	77.9%	76.0%
14	79.5%	92.6%	38	70.3%	82.0%
15	94.7%	85.5%	39	44.3%	88.8%
16	92.6%	95.7%	40	61.6%	85.6%
17	88.9%	93.5%	41	68.3%	82.7%
18	90.0%	92.4%	42	58.7%	90.7%
19	100.0%	85.6%	43	48.0%	91.7%
20	93.5%	85.4%			
21	81.8%	88.9%			
22	69.8%	97.7%			
23	66.7%	92.8%			
24	54.7%	84.3%			
Average	86.0%	91.3%	Average	68.3%	89.3%

The overlapping between the phoneme classification can be shown in a form of a confusion matrix as shown in Figure 3.3. (See Appendix for phoneme codes). This matrix suggests way to measure similarity in sounding between different phonemes. For example, a high degree of confusion can be seen between the nasals, but only a small amount of confusion between nasals and other phoneme classes.

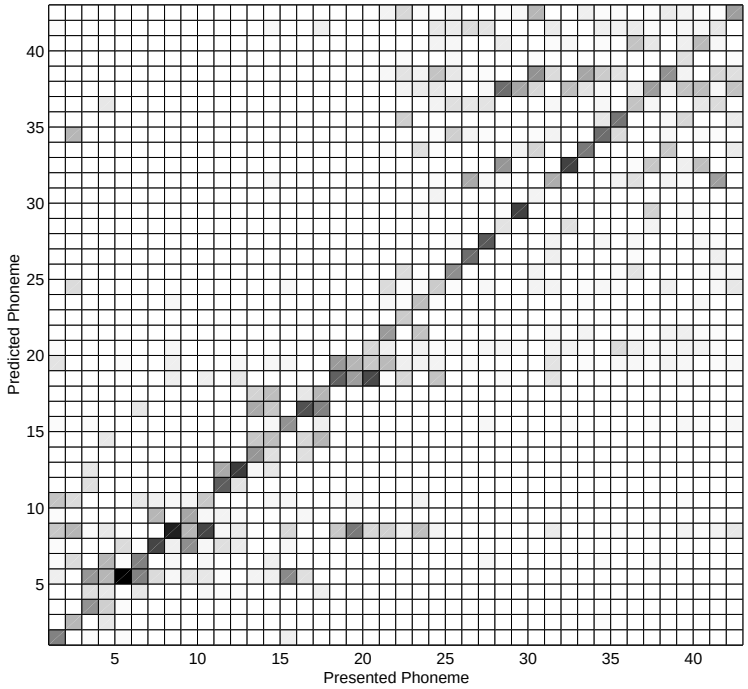

Fig. 3.3. A confusion matrix of the test classification of NZ English phonemes in the FuNN-based multi-modular all-phoneme classifier. Darker areas show higher recognition rates.

4. Fuzzy Spatial Temporal (FuST) Maps for phoneme and word data mapping

4.1 A General Introduction

FuST maps are connectionist structures which have time sequence of vectors as inputs and a topologically, spatially organised map as an output. Kohonen self-organised maps (SOM) [22] with time sequence of vectors as inputs, are examples of FuST maps. FuST maps can be trained either a supervised or unsupervised way.

FuST maps can be used to map similar sequences of elementary events over time intervals (not necessarily equal in duration) into topologically close area on the map. When used in the general framework from Figure 2.1, to realise the block of sequence of events recognition, the FuST map may recall, after new data is input, both meaningful and meaningless sequences, which is defined by the conscious decision making block through a feed-back connection.

4.2 FuST Map for Mapping Phoneme- and Word- Data

One of the first applications of the Kohonen SOM was for phoneme and word recognition. In this section Kohonen SOMs are used to map temporal sequences of linguistic features into phonemes, where the activation of all the output nodes is presented and not only the winning output neuron. Mapping phonetic representation of words into a 'sounds-like' FuST map is also demonstrated.

Choosing the appropriate features to represent and deal with speech sounds is extremely important in order to successfully map the sounds into a spatially organised map. The distance between the sound projections will help the conscious module to correctly recognise the elementary sounds and make them meaningful in terms of the ultimate task of spoken language (languages) recognition.

Based on the similarity between the sounding of the phonemes, inferred from the results produced by the FuNN-based classifier and the FuST map, and also based on linguistic knowledge, a 45-element phoneme activation vector (PAV) for each of the phonemes has been created. For shared features between the phonemes (for example 'alveolar'), see the Appendix. A PAV contains a value of activation of 1 for that phoneme, lesser activation values for similar phonemes and values of 0 for different phonemes [17, 18]. Mapping the PAVs into a SOM is shown in Figure 4.1. The SOM was trained with 45 PAV for 100,000 epochs. The phonemes are clearly distinguished on the map. Further training and adaptation of the SOM on more data is possible.

The FuST map from Figure 4.1 shows fuzzy, but distinguishable patterns of each of the phoneme as data over time is mapped into a topologically organised map. This is in contrast to the ambiguous mapping in the first two

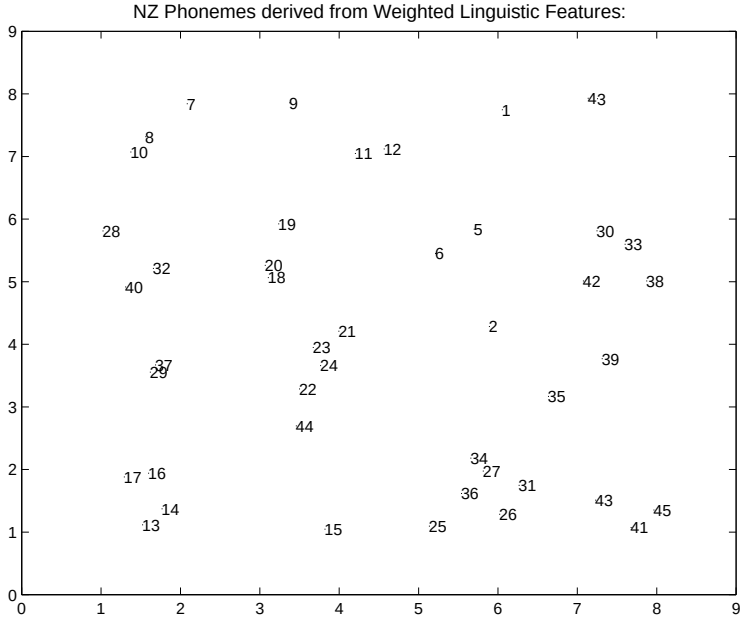

Fig. 4.1. The activation of all the output neurons of a SOM for input vectors representing the 45 phonemes in New Zealand English.

formant feature space. The phoneme data activate fuzzy areas, 'patches' on the FuST map.

Two further 45-element PAV data sets representing two separate classes of FuST maps were created, each containing 24 consonants and 13 vowels (the diphthongs were removed). The consonant labelled PAV data set was weighted with 1s with its vowels weighted to 0.25 respectively, and the vowel PAV data set was weighted with 1s and the consonants weighted at 0.25. Trained each of the two SOMs (7x7) with their respective PAV data sets for 100,000 epochs, holding the learning rate (LR) at 0.05 for the first 50,000 then releasing the LR, the neighbourhood (NB) was set to 1.

The resulting plotted maps show very good signs of distribution and clustering, as the distribution between the opposing phoneme classes is high, and the clustering of similar phoneme classes is excellent (see Figures 4.2 and 4.3)

The FuNN-based phoneme classifiers and the FuST maps described in this and the previous sections are used for a partial implementation of HySpeech/2 as explained in the next section

5. Aggregation of Phoneme Activation Vectors and Word Recognition

Phonemes segments are recognised over time. That is, one phoneme segment is produced approximately every 5.8ms. These must be consolidated into the

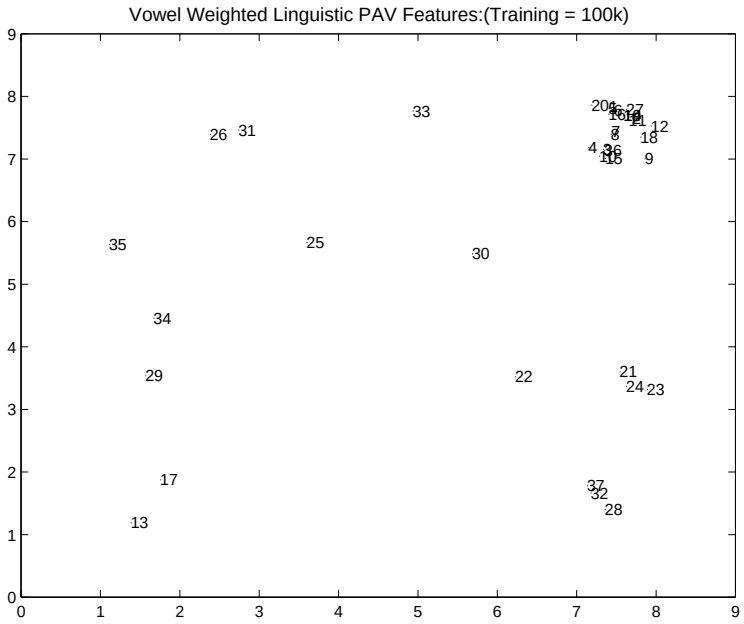

Fig. 4.2. Map of the vowel PAVs into a SOM.

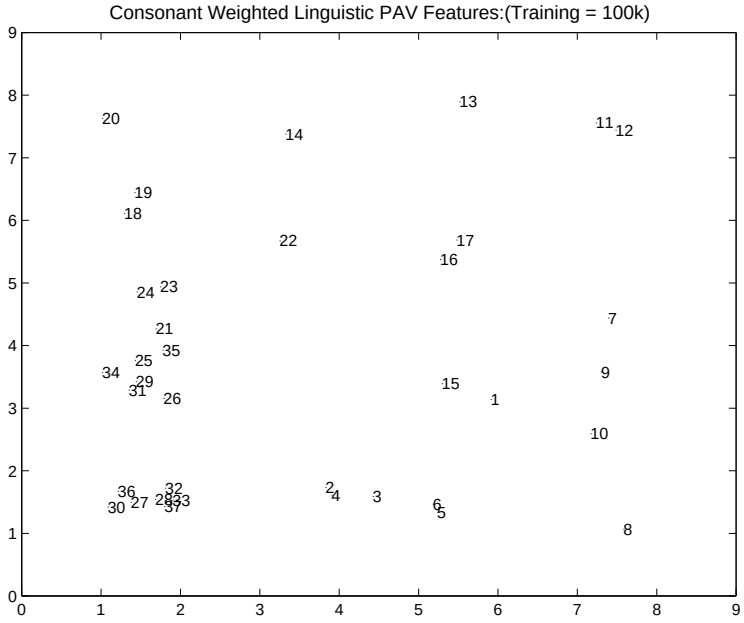

Fig. 4.3. Map of the consonant PAVs into a SOM.

abstract units (phonemes) used by the higher order processes. Several options are available for phoneme aggregation, and are currently under investigation.

The application of linguistic knowledge is the principle method for aggregation. Consecutive segments that are recognised as the same phoneme are combined into a single, more certain, unit. Furthermore, some classes of phonemes cannot occur together on some contexts. For example, within a single syllable, two nasals cannot occur together, and two consecutive nasals are rare, in such words as "government". Even here, the /n/ phoneme is often absent, as the sound is not articulated. Thus, the probability of two consecutive nasals are rare, and if recognised as consecutive, it is likely that one is simply a misrecognition. This is especially true with the phoneme segments which, as can be seen from Figure 3.3, may be highly confused.

Consonant clusters are further restricted by phonotactic restraints. These are particularly important at word onset, as many clusters are not permissible. For example, the most complex onset clusters are three phonemes long, and must begin with /s/. By disallowing consecutive phonemes that violate these restraints, the certainty of the legal phonemes can be increased.

Further, acoustic knowledge can be included in the system. The stop phonemes, for example, are always preceded by a period of silence. This silence is a reliable indication of the following stop, particularly in isolated words where it lasts considerably longer than in continuous speech.

Another source for aggregation is the expected length of the uttered segments. Here, the expected length is taken from the segments segmented manually. Currently under development is an automatic segmentation system, where phoneme boundaries are detected, as part of the adaptation (see below). This may be used for the improvement of probabilities based on duration between boundaries.

6. Current Experimental Applications of HySpeech/2

The two current applications of the HySpeech framework are discussed below. The first is designed to test possible adaptation techniques, the second, developed in parallel, is an bilingual isolated word English-Maori translation system.

6.1 The Rostam System

The Rostam system is a small experimental system, currently undergoing testing for possible automatic adaptation methods. Rostam uses a subset of the FuNN in the HySpeech system, and when a new speaker is introduced to the system the FuNN adapt to this speaker.

In order for the system to automatically adapt, the current scheme is as follows. A correct recognition needs no immediate adaptation, although

the speech may be saved for later examples of correct data. An incorrect recognition must be identified by the user, and the utterance is saved as the corrected word.

Once a word is identified as incorrect, it is marked for adaption. A reliable method for automatic phoneme segmentation is currently under investigation. The existing FuNN are trained further by the new data, and the FuNN are used for future attempts at recognition.

6.2 English-Maori translation

The HySpeech system is now being used as an interface to an English-to-Maori isolated word translation system. The existing system has a vocabulary of 2000 English words, and their closest Maori translations. As an isolated word translation system, there is no context information, and the most common translation is taken, although other translations are also retrieved.

The architecture allows for replacement of various modules for the subtasks. For example, if a FuNN is adapted to a new speaker, it may be restored if the performance degrades on existing speakers. The modularity also allow parallel implementation. Thus, the higher-level processes are currently being developed independently from the existing system (see below).

7. Directions for Further Research

7.1 Agent-based Implementation

A next version of the described in this paper system based on multi-level recurrent organisation [14] is under development and its agent-based implementation [5] is being experimented with. The new system will deal with continuous speech recognition with unlimited vocabulary as the FuST map used to store the vocabulary in 'sounds-like' clusters allows for a fast recall operation to retrieve the best matched words at any time moment. At the next level another FuST map is used to store words in concept clusters. This architecture is illustrated in Figure 7.1.

A higher-level language model is used to select the most likely words from the two FuST maps according to statistical domain knowledge. Such knowledge also takes into account linguistic rules and language patterns of the expected users. The knowledge represents the probability of a transition from an ordered set of previous words to a subsequent word, for example the conditional probability of word 3 (W3) following words 1 and 2 (W1 and W2 respectively). As explained in Section 1.3 such probabilities are extremely useful in speech recognition. Assuming that the two previous words have been identified with attached certainty factors (CF1 and CF2 respectively) defined as their activation values, the revised CF3 is calculated as in equation 7.1.

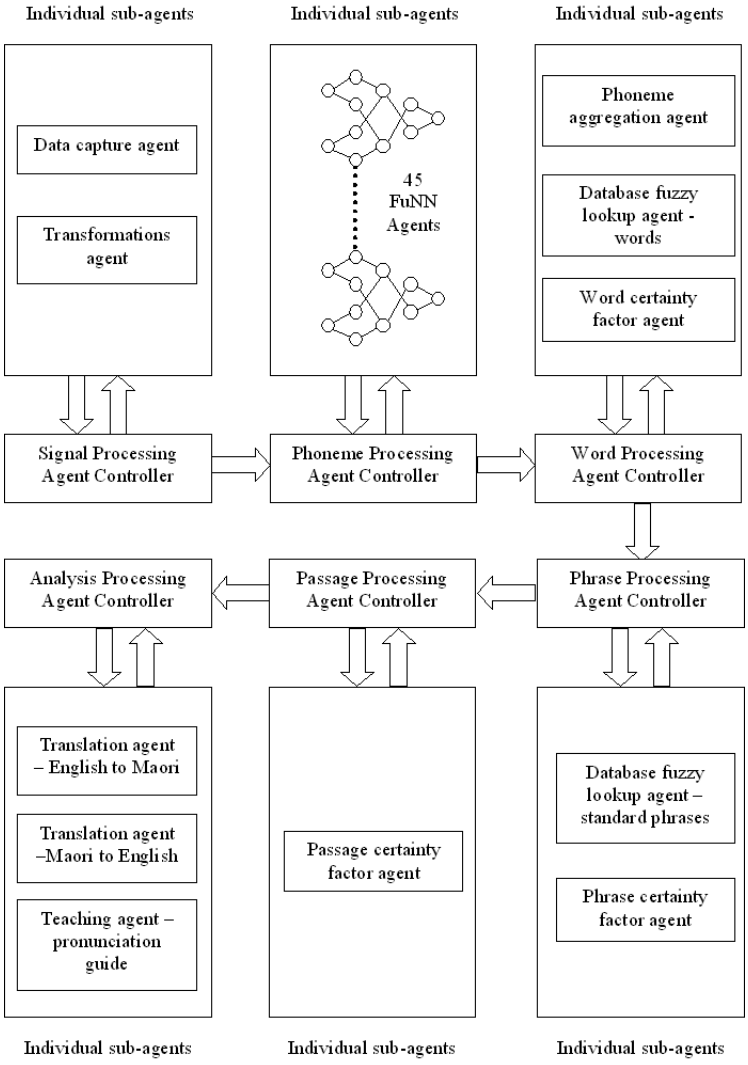

Fig. 7.1. Agent-based HySpeech system.

The word with the highest modified certainty factor can then be selected, subject to subsequent revision as new information becomes available.

$$CF'_{W3} = CF_3 \times P(W3|W1\&W2) \times CF_{W1} \times CF_{W2} \tag{7.1}$$

For example, based on the following case W3 would be identified as "FAST" in a robot control system giving the phrase "MOVE UP FAST".

Table 7.1. Example recognised and revised word probabilities

W1	CF_{W1}	W2	CF_{W2}	CF'_{W2}	W3	CF_{W3}	CF'_{W2}
MOVE	**0.75**	**UP**	**0.85**	**0.57**	FIRST	0.85	0.16
					FAST	**0.85**	**0.38**
		OFF	0.55	0.04	FIRST	0.85	0.07
					FAST	0.85	0.28

Table 7.2. Example word rules

Rules	
P(UP|MOVE)	=0.90
P(OFF|MOVE)	=0.10
P(FAST|MOVE UP)	=0.70
P(FIRST|MOVE UP)	=0.30
P(FIRST|MOVE OFF)	=0.20
P(FAST|MOVE OFF)	=0.80

Such a system allows for immediate revision of the speech already recognized as new speech information becomes available. This is in addition to the higher accuracy obtained by using prior information.

The agent-based approach to implement the new architecture takes into account the concepts of modularity, expandability, adaptability, gradual commitment, and the use of expert knowledge, to provide the greatest accuracy and flexibility possible. By using a series of agents (modules that communicate amongst themselves) for various phases of the recognition process, the system uses the process of negotiation and information requests to arrive at the final classification for the utterance. In the proposed architecture there are six levels of agents: signal processing, phoneme recognition, word formation, phrase construction, passage production, and pragmatic analysis. Each level contains sub-agents, for example phoneme recognition consists of a voting agent and 45 separate phoneme recognisers. The levels communicate only with adjacent levels, passing on probabilities associated with various outputs, requesting information on new outputs, and requesting changes to processing. At any time, as new information becomes available, the agents can collectively revise their outputs.

7.2 Using GA for optimisation and adaptation in speech recognition systems

Genetic algorithms (GAs) are a computational representation of the processes of biological evolution [4]. By iteratively combining previous, partially successful attempts at solving a problem, they are able to efficiently find approximate solutions to combinatorial problems. Employing GAs in speech recognition systems utilises this ability to optimise neural networks.

One of the most promising application areas of GAs in speech systems is the automated design of FuNNs [21]. Issues such as the selection of input features, and the number of membership function and rule nodes selected are more easily solved by GA than other approaches.

GAs may also be used to tune already trained FuNNs. This may be done by either using a GA to tune the membership functions of the FuNN [16], or, alternatively, using a GA to select the backpropagation training parameters for each layer of weights [21]. GAs could also be used to select which examples within a new data set are the most relevant for adaptation.

Further development of this project have already started into the direction of building a multi-lingual 'conscious' machine [14].

8. Conclusions

The paper describes a current research project on speech data analysis and speaker independent isolated word recognition which explores new techniques, such as fuzzy neural networks and spatial temporal maps organised in a novel architecture.

Acknowledgement. This work was done as part of the UOO606 project funded by the New Zealand PGSF of the Foundation for Research Science and Technology.

References

1. Amari, S. and Kasabov, N. eds (1997) Brain-like Computing and Intelligent Information Systems, Springer Verlag
2. Benton, R. A. (1992) Māori English: A New Zealand Myth?, New Zealand English Newsletter, Number 6, Department of English, University of Canterbury, Christchurch. (pp. 27-35).
3. Cole, R. et al (1995) The Challenge of Spoken Language Systems: Research Directions for the Nineties, IEEE Transactions on Speech and Audio Processing, vol.3, No.1, January 1995, 1-21.
4. Goldberg, David E, 1989. Genetic Algorithms in Search, Optimisation and Machine Learning, Addison-Wesley Press, 1989

5. Gray, A., Kilgour, R. and Kasabov, N. (1997) A Framework for agent-based implementation of adaptive speech recognition systems, Proceedings of ICONIP/ANZIIS/ANNES'97, Springer Verlag, Singapore
6. Hashiyama, T., Furuhashi, T., Uchikawa, Y. (1992) A Decision Making Model Using a Fuzzy Neural Network, in: Proceedings of the 2nd International Conference on Fuzzy Logic & Neural Networks, Iizuka, Japan, 1057-1060.
7. Ishikawa, M. (1996) Structural Learning with Forgetting, Neural Networks, 9, 501-521.
8. Jang, R. (1993) ANFIS: adaptive network-based fuzzy inference system, IEEE Trans. on Syst.,Man, Cybernetics, 23(3), May-June 1993, 665-685
9. Jusczyk, P. (1997) The Discovery of Spoken Language, MIT Press
10. Kasabov, N. (1995) Hybrid Connectionist Fuzzy Production Systems - Towards Building Comprehensive AI, Intelligent Automation and Soft Computing, 1:4, 351-360
11. Kasabov N. (1995) Hybrid Connectionist Fuzzy Rule-based Systems for Speech Recognition, Lecture Notes in Computer Science/Artificial Intelligence, Springer Verlag, No.1011, 20-33
12. Kasabov, N. Adaptable connectionist production systems, Neurocomputing 13 (2-4), 1996, 95-117
13. Kasabov, N.(1996) Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering, The MIT Press, CA, MA.
14. Kasabov, N. (1997) A Framework for Intelligent 'Conscious' Machines Utilising Fuzzy Neural Networks and Spatial temporal Maps and a Case Study of Multilingual Speech Recognition, in: S.Amari and N.Kasabov (eds) Brain-like Computing and Intelligent Information Systems, Springer Verlag, Singapore
15. Kasabov, N., Kilgour, R., Sinclair, S. (1997) From hybrid adjustable neuro-fuzzy systems towards connectionist-based adaptive systems for phoneme-, word-, and language recognition. Fuzzy Sets and Systems, to appear
16. Kasabov, N., Kim J S, Watts, M., Gray, A (1997) FuNN/2- A Fuzzy Neural Network Architecture for Adaptive Learning and Knowledge Acquisition, Information Sciences - Applications, 1997, in print
17. Kasabov, N. and Kozma, R. (1997) Adaptive fuzzy neural networks and applications for chaotic time-series analysis and phoneme-based speech recognition, IEEE Transactions on Neural Networks, to appear
18. Kasabov, N., Kozma, R. and Watts, M. (1997) Optimisation and Adaptation of Fuzzy Neural Networks Through Genetic Algorithms and Structural Learning , Information Sciences - Applications, in print
19. Kasabov, N., Kozma, R., Kilgour, R., Laws, M., Taylor, J., Watts, M., Gray, A. (1997a) HySpeech/2: A Hybrid Speech Recognition System, Proceedings of the ICONIP/ANZIIS/ANNES'97, Dunedin, 24-28 November 1997, Springer Verlag, Singapore
20. Kasabov, N., Sinclair, S., Kilgour, R., Watson, C., Laws, M. and Kassabova, D. (1995) Intelligent Human Computer Interfaces and the Case Study of Building English-to-Māori Talking Dictionary, in: Proceedings of ANNES'95, Dunedin, IEEE Computer Society Press, Los Alamitos, 294-297
21. Kasabov, N., Watts, M., Genetic Algorithms for Structural Optimization, Dynamic Adaptation and Automated Design of Fuzzy Neural Networks in Proceedings of ICNN '97 Conference, Houston, Texas
22. Kohonen, T. (1990) The Self-Organizing Map. Proceedings of the IEEE, vol.78, N-9, pp.1464-1497.
23. Kosko, B. (1992) Neural Networks and Fuzzy Systems: A Dynamical Approach to Machine Intelligence. Prentice Hall.

24. Kozma, R., Sakuma, M., Yokoyama, Y., Kitamura, M. (1996) On the accuracy of mapping by neural networks trained with backpropagation with forgetting, Neurocomputing, 13, 295-311
25. Mitra, S., Pal, S. (1995) Fuzzy Multi-Layer Perceptron, Inferencing and Rule Generation, IEEE Transactions on Neural Networks, vol.6, No.1, 51-63
26. Ray, K., Ghoshal, J. (1997) Neuro-Fuzzy Approach to Pattern Recognition, Neural Networks, vol.10, No.1, 161-182
27. Sinclair, S., Watson, C. (1995) The Development of the Otago Speech Database. Proceedings ANNES '95, University of Otago, Dunedin.
28. Takagi, H. (1990) Fusion Technology of Fuzzy Theory and Neural Networks - Survey and Future Directions, in: Proc. First Int. Conf. on Fuzzy Logic and Neural Networks, Iizuka, Japan, July 20-24, pp.13-26.
29. Yamakawa, T., Kusanagi, H., Uchino, E. and Miki, T.(1993) A new Effective Algorithm for Neo Fuzzy Neuron Model, in: Proceedings of Fifth IFSA World Congress, (1993) 1017-1020.
30. Zadeh, L. 1965. Fuzzy Sets, Information and Control, vol.8, 338-353.
31. Zadeh L. (1984) Making Computers Think Like People, IEEE Spectrum, Aug 1984, pp.26-32

Table .1. Linguistic features for phoneme codes 01-15

Code Phoneme	1 p	2 b	3 t	4 d	5 k	6 g	7 f	8 v	9 T	10 D	11 s	12 z	13 S	14 Z	15 h
Glottal															+
Post-alveolar															
Palatal															
Lateral															
Labiodental							+	+							
Dental									+	+					
Affricate															
Nasal															
Central															
Bilabial	+	+													
Centering															
Velar					+	+									
Palato-alveolar													+	+	
Closing															
Mid															
Open															
Strident							+	+			+	+	+	+	
Front															
Low															+
Unrounded															
Close															
Alveolar			+	+							+	+			
Diphthong															
Rounded															
Unvoiced	+		+		+		+		+		+		+		+
Monophthong															
High					+	+							+	+	
Coronal			+	+					+	+	+	+	+	+	
Back					+	+									
Anterior	+	+	+	+			+	+	+	+	+	+			
Tense	+		+		+		+		+		+		+		
Consonant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Sonorant															
Continuant							+	+	+	+	+	+	+	+	+
Voiced		+		+		+		+		+		+		+	

Table .2. Linguistic features for phoneme codes 16-30

Code Phoneme	16 tS	17 dZ	18 m	19 n	20 N	21 l	22 r	23 w	24 j	25 I	26 e	27 &	28 V	29 A	30 U
Glottal															
Post-alveolar							+								
Palatal									+						
Lateral						+									
Labiodental															
Dental															
Affricate	+	+													
Nasal			+	+	+										
Central													+		
Bilabial			+												
Centering															
Velar					+			+							
Palato-alveolar	+	+													
Closing															
Mid											+				+
Open												+	+	+	
Strident															
Front										+	+	+			
Low												+		+	
Unrounded										+	+	+	+		
Close										+					
Alveolar	+	+		+		+									
Diphthong															
Rounded								+						+	+
Unvoiced	+														
Monophthong										+	+	+	+	+	+
High					+			+		+					+
Coronal	+	+		+		+	+								
Back					+									+	+
Anterior	+	+	+	+		+									
Tense	+	+												+	
Consonant	+	+	+	+	+	+	+	+	+						
Sonorant			+	+	+	+	+	+	+	+	+	+	+	+	+
Continuant						+	+	+	+	+	+	+	+	+	+
Voiced		+	+	+	+	+	+	+	+	+	+	+	+	+	+

Table .3. Linguistic features for phoneme codes 31-45

Code Phoneme	31 i	32 a	33 O	34 3	35 u	36 eI	37 ai	38 Oi	39 OU	40 aU	41 i@	42 U@	43 e@	44 Q	45 @
Glottal															
Post-alveolar															
Palatal															
Lateral															
Labiodental															
Dental															
Affricate															
Nasal															
Central				+											+
Bilabial															
Centering											+	+	+		
Velar															
Palato-alveolar															
Closing						+	+	+	+	+					
Mid				+								+	+		+
Open		+					+			+					
Strident															
Front	+					+					+		+		
Low		+	+				+		+						
Unrounded	+	+									+		+		
Close	+		+		+			+	+	+	+				
Alveolar															
Diphthong						+	+	+	+	+	+	+	+		
Rounded			+	+	+				+	+		+			
Unvoiced															
Monophthong	+	+	+	+	+										+
High	+				+						+	+			
Coronal															
Back		+	+		+		+	+	+	+		+			
Anterior															
Tense	+		+	+	+	+	+	+	+	+	+	+	+		+
Consonant															
Sonorant	+	+	+	+	+	+	+	+	+	+	+	+	+		+
Continuant	+	+	+	+	+	+	+	+	+	+	+	+	+		+
Voiced	+	+	+	+	+	+	+	+	+	+	+	+	+		+