Integrating neural networks and fuzzy systems for speech recognition

N. Kasabov, C.Watson, S. Sinclair, R. Kilgour
Department of Information Science, University of Otago
P.O.Box 56, Dunedin, New Zealand, contact email: nkasabov@otago.ac.nz

The paper presents a framework of an integrated environment for speech recognition and a methodology of using such environment. The integrated environment includes a signal processing unit, neural networks and fuzzy rule-based systems. Neural networks are used for "blind" pattern recognition of the phonemic labels of the segments of the speech. Fuzzy rules are used for reducing the ambiguities of the correctly recognised phonemic labels, for final recognition of the phonemes, and for language understanding. The fuzzy system part is organised as multi-level, hierarchical structure. As an illustration, a model for phoneme recognition of New Zealand English is developed which exploits the advantages of the integrated environment. The model is illustrated on a small set of phonemes.

1. INTRODUCTION

Numerous are the speech recognition systems built so far. Different techniques have been explored at the different levels of the whole speech recognition process. These levels are signal processing, pattern matching and quantisation, syntactic and semantic analysis, and language understanding. For the pattern matching phase, the techniques of hidden Markov modelling, dynamic time warping, neural networks, and hybrid systems made from the above techniques, have been used (Morgan and Scofield, 1991). For language analysis, rule-based systems and hierarchical neural networks have been applied.

Despite the enormous efforts so far, the goal of building speech recognition system for continuous speech, speaker-independent, large or non-limited vocabulary, has not been achieved so far. This we believe is due to the absence of higher order language models in the speech recognition process. Ambiguities in the meaning of speech can occur from the sound level right through to the syntactic level. This means that some form of higher level processing (higher, that is than the level in which the ambiguity occurs) is required by the "hearer" of the utterance. Due to the ambiguities at the syntactic level, correct interpretation of a sentence can only occur if the context of the sentence is known (Ainsworth, 1988). Therefore it is necessary, that some sort of language modelling will be utilised for computer speech recognition. This is especially true for continuous, unlimited vocabulary computer speech recognition.

The approach used in this paper aims at continuous speech recognition, with unlimited vocabulary. To do this we believe it is necessary to have a well-developed language model, starting from the acoustic, phonetic level and going right through to the semantic level (this philosophy is shared by others, Roe and Wilpon (1993) for example).

The approach here treats the whole process of speech recognition and language analysis as a continuous process and there is no rigid border between the two. So, speech knowledge and language analysis knowledge should concurrently reside in one system and should be represented in a comprehensive language model.

There is currently no comprehensive language model that can suitably be applied to computer speech recognition and language analysis. The problems in applying language modelling to speech recognition have been discussed by Roe and Wilpon (1993). They said there are two

"fundamental problems". First "how can one mathematically describe the structure of a language's valid sequence of words and phonemes; that is, given a putative sequence of symbols, how can one determine the likelihood that the sequence is valid in human language". Secondly "given such a mathematical description of the language, how can one efficiently compute the optimum sequence of symbols from the acoustic pattern classifier that meets that mathematical description? ". Language modelling is "the weak link" in computer speech recognition, it is an area that requires extensive research (Roe and Wilpon, 1993).

Rule-based higher level recognition and processing require rules that reflect the ambiguity in the language. Higher level rule-based systems must have facilities to represent and process ambiguous, contradictory, incomplete, uncertain knowledge. Articulating rules for language processing, which are fuzzy by nature, is a difficult task for the language and the knowledge engineering experts. The rules, used by humans, reflect their knowledge of the language. They are not a simple "flat" set, but possibly a hierarchical structure with many layers of rules. The more we learn and practice a language, the more layers we build in our brain, the more complex speech constructions we can recognise and comprehend. Huckvale (1990) puts stress on exploiting speech knowledge in neural networks for speech recognition. Not only pattern matching, but also explicit rules are recommended by the author. Unfortunately he does not suggest any concrete ways to implement this idea.

In order to represent complex hierarchical speech and language structures in a computer system, we need an environment where one can represent speech and language structures and process them efficiently, and where one can use both speech data and speech knowledge. This is the theme of the paper.

2. INTEGRATED SYSTEM FOR SPEECH RECOGNITION AND LANGUAGE ANALYSIS

We consider speech recognition and language understanding to be a continuous process and the borders between the two are not well defined. Where does the speech recognition process end and the language analysis process start? Are they not two overlapping processes ? We believe that by separating the two tasks in a system, one is limiting the possibilities of improving each of them.

Our philosophy on speech recognition and language analysis is the guiding direction for building integrated software environments which facilitate both speech recognition and language understanding, thus making use of both pattern matching techniques (neural networks) and knowledge-based processing (fuzzy rule systems). This approach has been suggested in (Kasabov 1994a).

The structure of a speech recognition system, which is under consideration for further experiments within the project group, is given in figure 1. It has the following four main blocks:

- A speech digitizer (in which the sampling rate can be selected).
- A signal processing module - this performs non-linear transformations over the raw speech signal. Different transformations can be applied (such as spectral analysis or cepstral analysis) these could contribute to the accuracy of the final recognition.
- A variable neural network module for low level recognition (different architectures can be selected, e.g. recurrent, back propagation, hierarchical, modular).

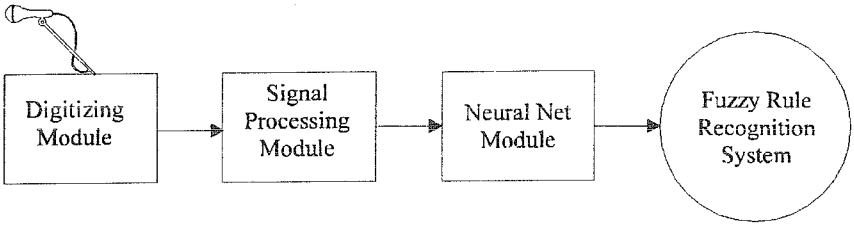

Figure 1 An integrated architecture for speech recognition and language analysis

- A hierarchical fuzzy production system (FPS) for higher level processing - this performs approximate reasoning over fuzzy production rules for phoneme, word, sentence, concept, and meaning recognition. A FPS can be realised either in a fuzzy inference machine (for example the TIL Shell, Fuzzy Clips etc.) or in a connectionist environment. An example of such connectionist fuzzy production system environment is the Neural Production System architecture (Kasabov and Shishkov, 1993).

The architecture shown in Figure 1 is adaptable in two ways:
1. additional training of the neural networks with more examples about new speech variations;
2. adding fuzzy rules to the FPS higher level module for representing the new speech peculiarities.

3. AN INTEGRATED MODEL FOR PHONEME RECOGNITION

Various experimental approaches are being investigated using the tool box outlined in figure 1. The following outlines an integrated model for phoneme recognition.

3.1 Data Segmentation

In this particular model the sampling rate is selected to be 22.05 kHz. This sampling rate was deemed large enough to give a good representation for all speech sounds. Whilst most speech has no significant energy higher than 5 kHz, some sounds, such as fricatives, have significant energy until at least 10 kHz. The digitized speech was transformed into Mel-scaled cepstrum coefficients (MSCC), using the method as described by Davis and Mermelstein (1980). MSCC are obtained from a 256 point fast Fourier Transform, calculated from Hamming windowed speech. Consecutive sets of MSCC were calculated from speech frames overlapped by 50%.

3.2 Training A Modular Neural Network Structure

The MSCC used to train the network were obtained from spectrally constant portions in the speech wave form. Since the vocal tract is continually changing in speech, so to is the spectrum of speech. However over a short duration (20ms -50ms) the spectrum can be considered to be constant. To locate the spectrally constant portions, the cleanest allophonic realization of the phonemes were isolated, the judgment was an aural one. Next, several overlapping portions of the speech signal were saved. The portions were between 20 -50 ms long, the spectrum within

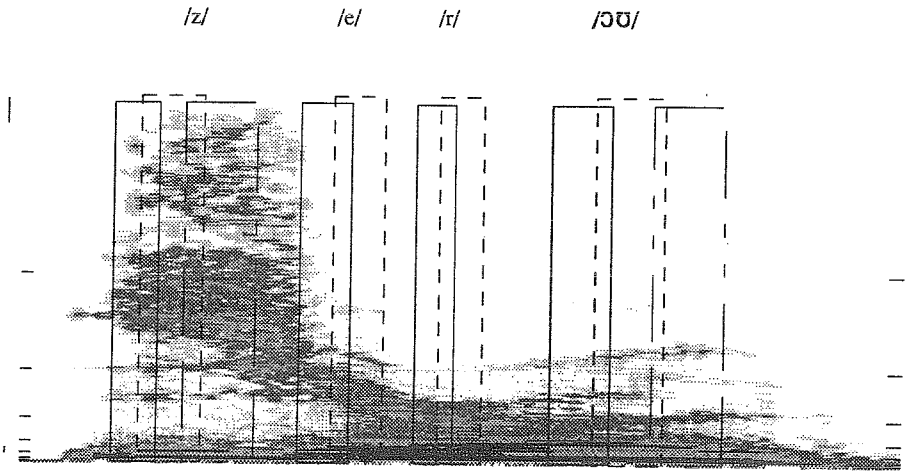

Figure 2 Demonstrating how an utterance of the word "zero" is partitioned from a spectrogram.

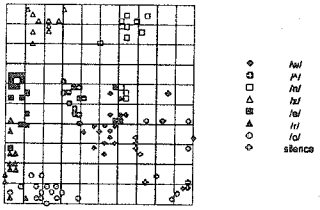

Figure 3 A Kohonen map of a number of phonemes of New Zealand English and silence.

these portions was virtually unchanging. Figure 2 is an example of how an utterance of the word "zero" was portioned. First the portions within the word, which were the cleanest allophonic realizations of the phonemes /z/, /e/, / ɪ/, /ɔʊ/, are found. Then, for each allophonic realization of a phoneme, 20-50ms portions were saved. For long sustained sounds (e.g. [z]) usually three overlapping portions were saved; for short sounds (e.g. [e]) usually two overlapped portions were saved. It can be seen from Figure 2 that some parts of the word "zero" were not saved at all. These parts, if listened to, could not be classified to a particular phone.

By looking at a Kohonen map of the phonemes from the words "zero" and "one", obtained from 4 different male speakers, the method of segmentation appears valid. It can be seen in Figure 3

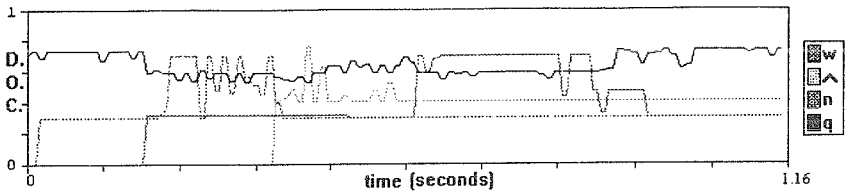

Figure 4 The degrees of certainty (D.O.C.) for silence (q) and the phonemes /w/,/ʌ /,/n/, for an utterance of the word "one".

that the clusters of the different phonemes, and of silence are all well separated on the map.

The MSCC obtained from the saved portions were passed to a modular back propagation neural network system. For each phoneme, a small specialized network was trained, the output of the network being high if the phoneme is recognised and low if it was not. A network was also trained to recognise silence.

3.3 Fuzzy Rules

The outputs of all the networks were then passed into a hierarchical fuzzy system. For phoneme recognition there was only two levels. The first level reduced the data from the neural networks to a stream of probable phonemes, these phonemes mostly had high degrees of certainties associated with them. The same phoneme could occur consecutively. The rules used to do this reduction were linguistically based. For example the phoneme /w/ cannot precede the phoneme /n/ in New Zealand English. Therefore if the /w/ phoneme had a high degree of certainty and the surrounding /n/ phonemes also had a high degree of certainty then the /w/ is disregarded. The second level in the fuzzy system reduced the phoneme stream into a machine deduced phonemic transcription of the input speech. A second set of linguistic rules were used to perform this reduction.

Using a set of language rules, it would be possible to convert this machine deduced phonemic transcription into a word. We recognise that in English a string of phonemes will not necessarily render a unique spelling (for example the string of phonemes /tu/ can be interpreted as the word "two", "to", or "too"). However a higher order language model would be able to account for the ambiguities (for it is only when the context is established that "two", "too", or "to" could be discriminated). These higher order models are currently being investigated.

3.4 Results

Figure 4 is a example of the output obtained from the first level in the fuzzy system, it is for the spoken word "one". The graph plots the degrees of certainty for each of the three phonemes in the word "one" (/w/,/ʌ /, /n/) as a function of time. The degree of certainty for silence (given the label q in figure 4) is also plotted. The likelihood of existence of a phoneme or silence is determined by the degrees of certainty. A particular phoneme is assumed to have occurred if its degree of certainty is both close to 1 and greater than the degree of certainty for silence.

It can be seen from figure 4. that before the word is spoken the degree of certainty for silence is high and those for the phonemes /w/, /ʌ/ and /n/ are low. Then the degree of certainty for silence drops and that for the phoneme/w/ becomes high. The degrees of certainty for the phonemes /ʌ/

and /n/ remain low. In time, the certainty of /w/ decreases and the certainty of /ʌ/ increase. The process repeats for /ʌ/ and /n/ and then for /n/ and silence. It can be seen that there are periods of ambiguity between consecutive phonemes which have high degrees of certainty. Ambiguity occurs when all phonemes and silence have low degrees of certainty. These periods of ambiguity are partly due to the fuzzy rules developed and partly because of the way the neural networks were trained. In the recognition phase, since the networks were not trained on portions of the words which are predominantly transitions, there are instances when the outputs of the neural networks were all low.

4. CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

An integrated architecture for speech recognition and language analysis has been presented. It consists of four main blocks, the last two being a neural network module and a fuzzy rules module. The system is expected to be a very powerful experimental tool. Many parameters can be varied in each module. In a preliminary experiment using back-propagation neural networks and a two level fuzzy system, a limited phoneme recognition experiment was performed (the system was trained on the allophonic realizations of 7 phonemes). Clearly, the speech data base from which the training of the networks took place, must be expanded, and the number of speakers in the speech data base needs to be increased (currently there are only four male speakers), and higher level language modelling should be included. However, the results are very encouraging.

ACKNOWLEDGEMENTS

We acknowledge Telecom New Zealand Ltd for providing funding for conducting this research.

REFERENCES

AINSWORTH, W. A. (1988) *Speech Recognition By Machine,* Peter Peregrinus Ltd, London.

DAVIS, S. B. and MERLMELSTEIN, P (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions of Acoustics, Speech and Signal Processing*, Vol ASSP-28, No. 4, Aug.

HUCKVALE, M. (1990)," Exploiting Speech Knowledge in Neural Nets for Recognition", *Speech Communication*, No 9, North Holland, pp.1-13.

KASABOV, N (1994a), ``Towards Using Fuzzy Connectionist Production Systems for Speech Recognition", *Proceeding of the WWW'94,* Nagoya University, Japan, pp9-13.

KASABOV, N. and PEEV E. (1994b), `Phoneme Recognition with Hierarchial Self Organised Neural Networks and Fuzzy Systems - A Case Study', Editors Mannaro, M. and Morassoi, R. *Proceedings of ICANN'94*, May 1994, Vol. 1 , pp 201-204. Sorento, Italy: Springer Verlag.

KASABOV, N. and SHISHKOV, S,(1993), ``A Connectionist Production System with a Partial Match and Its Use for Approximate Reasoning", *Connection Science,* vol.5, 3&4, pp.275-305.

MORGAN, D.P and SCOFIELD, C. L. (1991), *Neural Networks and Speech Processing,* Kluwer Academic Publishers.

ROE, D.B. and WILPON, J. G (1993). `Whither Speech Recognition: The Next 25 Years',*IEEE Communications Magazine,* Nov, p54-62.

TERANO, T. (1993)," Long-term view on Fuzzy Technology and LIFE Projects", *LIFE Technical News*, Vol. 4, No.1.