# KNOWLEDGE BASED APPROACH TO SPEECH RECOGNITION

A. Samouelian

Department of Electrical and Computer Engineering
University of Wollongong

ABSTRACT-This paper presents a knowledge/rule based approach to continuous speech recognition. The proposed recognition system (Samouelian, 1994) uses a data driven methodology, where the knowledge about the structure and characteristics of the speech signal is captured *explicitly* from the database by the use of *inductive inference* (C4.5) (Quinlan, 1986). This allows the integration of features from existing signal processing techniques, that are currently used in HMM stochastic modelling, and acoustic-phonetic features, which have been the cornerstone of traditional knowledge based techniques. Phoneme recognition results on the phonetic classes of plosives, semivowels and nasals for a combination of feature sets, for speaker dependent and independent recognition, are presented.

## INTRODUCTION

The use of the knowledge/rule based approach to continuous speech recognition has been proposed by several researchers and applied to speech recognition (De Mori & Lam, 1986; Aikawa, 1986; Bulot & Nocera, 1989), spectrogram reading (Fohr et al, 1989; Zue and Lamel, 1986; Komori et al, 1989), speech understanding systems (De Mori and Kuhn, 1992), and speech verification (Diest et al, 1989). The production rules are in general of the form IF *condition* THEN *action*.

In traditional knowledge based approach, the production rules are developed heuristically, from empirical linguistic knowledge or from observations of the speech spectrogram. In this approach, the main bottleneck is in the acquisition and classification of the knowledge from the linguist, phonetician or spectrogram reading expert to formulate the appropriate production rules upon which phonetic classifications can be performed. To reduce this bottleneck, this paper proposes the use of machine learning in the form of C4.5 Induction System to capture and define the structure and characteristics of the speech signal explicitly, using classification rules in the form of decision trees generated from observations of the speech data, during the training phase. The resultant production rules are generated according to the features that provide the most information about a classification.

To induce the knowledge from the hand segmented and labelled speech database, all that is required is to nominate the set of features required for feature extraction and provide a basic structure of the classification. During the training phase, the feature extraction framework (Samouelian, 1992) extracts a set of features from the continuous speech on a frame by frame basis, and the induction system then generates automatically from the examples, a set of production rules. These rules are derived from the parameters that provide most information about a classification. The recognition is performed at the frame level, using an inference engine (Horn, 1991) to execute the decision tree and classify the firing of the rules.

The proposed system has two main advantages. Firstly, it uses the data-driven approach to phoneme classification, thus attempting to solve the problem of *inter* and *intra* speaker speech variability, by the use of a large speech database. Secondly, it has the ability to generate decision trees using any combination of features (parametric or acoustic-phonetic).

The aim of this research is to develop a flexible automatic speech recognition system that can integrate the various feature combinations, including traditional acoustic-phonetics, speech specific or spectrally based features or parameters and automate the process of generating the production rules, using a large speech database.

## TRAINING AND RECOGNITION STRATEGY

### Speech Database

The speech database consisted of 195 Australian accented English phrases and included all the permissible sounds of the language in all possible combinations by class. The phrases were collected from two females and one male speaker, each reading the phrases from prepared text, only once.

The phrases were devised and collected by National Acoustic Laboratories as part of the GLASS project. The database was hand segmented and phonetically labelled by The University of Sydney as part of the same project. Table 1 shows the classification of the speakers in the database.

| Speaker No. | Sex | Classification |
|---|---|---|
| 1 | Female | General to educated |
| 2 | Female | Broad to general |
| 3 | Male | Broad |

Table 1. The classification of the speakers in the database.

The phonemes of the consonant classes of *plosives*, *semi_vowels* and *nasals* were selected for investigation. Table 2 shows the number of phoneme tokens for each speaker. These tokens are according to the hand labelled transcription by the phonetician.

| Consonant Class | Phoneme | Number of Tokens | | |
|---|---|---|---|---|
| | | Speaker 1 | Speaker 2 | Speaker 3 |
| Plosives | /p/ | 123 | 119 | 121 |
| | /d/ | 289 | 304 | 259 |
| | /k/ | 183 | 173 | 167 |
| | /t/ | 366 | 344 | 315 |
| | /b/ | 106 | 113 | 112 |
| | /g/ | 73 | 77 | 75 |
| Semi-vowels | /l/ | 241 | 222 | 210 |
| | /w/ | 156 | 138 | 132 |
| | /r/ | 230 | 180 | 187 |
| | /j/ | 28 | 30 | 29 |
| Nasals | /m/ | 192 | 174 | 175 |
| | /n/ | 362 | 344 | 337 |
| | /ng/ | 45 | 43 | 48 |

Table 2. Number of phoneme tokens for each speaker.

It can be seen from table 2 that the number of phoneme tokens per speaker for the 195 sentences are not identical. Some of the reasons for these variations are:

1. In continuous speech, if the final phoneme is similar or the same as the initial phoneme of the following word, e.g. *soMe Meat*, then the two phonemes may merge resulting in a single phoneme.

2. Certain words that were merged by the speaker, e.g. *gunna* for *going to* were treated as one word.

3. The sounds were labelled according to what they were and not what they should have been.

4. Variations between speakers and the resultant articulations, e.g. physical differences between speakers may produce different phonetic labels (i.e. breathing phonation).

5. Conscious or precise articulation.

6. Coarticulatory effects on phonemes may vary between speakers, as does assimilation and elision, resulting in variations in number of tokens per speaker.

## Training

A block schematic of the training and recognition strategy is shown in Figure 1. Four different feature extraction modules have been used to train and test the recognition system. The description of each feature extraction module and the training methodology are detailed in a previous paper (Samouelian, 1992, 1994). Only brief details are provided here for completeness.

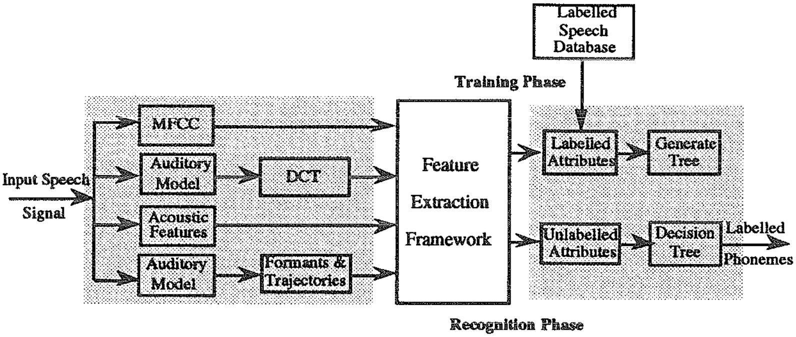

Figure 1. Block schematic of training and recognition strategy

The four feature extraction modules were:

1. MFCC.
2. DCT.
3. FEATURE.
4. TRAJ.

Modules MFCC and FEATURE extracted features in the time domain, while modules DCT and TRAJ extracted features from the auditory model in the frequency domain.

During the training phase, the feature extraction framework extracted features or parameters from the continuous speech on a frame by frame basis. The time aligned phonetically labelled files were then used to associate each frame with its corresponding label and generate a training data file. This data file was constructed on the basis of the training samples, which contained labelled examples in the form (X,b), where *X* was a feature vector and *b* was the corresponding class. This data file was then used by the C4.5 program to generate a decision tree.

Table 3 shows the feature set for the four different feature extraction modules that were used to test the recognition system.

| Parameters of Feature Extraction Modules | | | |
|---|---|---|---|
| MFCC | DCT | FEATURE | TRAJ |
| 12 mfcc coeffs | 12 DCT coeffs | rms | F1 |
| | | max_amp | F2 |
| 12 delta mfcc coeffs | 12 delta DCT coeffs | zcr | F3 |
| | | voicing | F4 |
| energy | energy | energy | F5 |
| | | local_diff_p | trajF1 |
| | | local_diff_n | trajF2 |
| | | envelope | trajF3 |
| | | auto_peak | trajF4 |
| | | ac_pp | trajF5 |

Table 3. The feature set used in the various feature extraction modules.

Recognition

The recognition was performed at the frame level and the performance was evaluated by comparing each classified frame against the reference frame derived from the hand labelled data. This procedure allowed the correct identification of substitutions and insertions per frame. An inference engine (written in Prolog) was used to execute the decision tree.

## RECOGNITION RESULTS

Tables 4 and 5 show the overall phoneme recognition results for the consonant class of plosives, for speaker dependent and independent recognitions respectively, for speakers 1, 2 & 3 in % correct, for the various feature extraction modules. For speaker dependent recognition, the system was trained and tested on the same speaker, while for speaker independent recognition, the system was trained on one speaker and tested on the other two in turn.

| Speaker Dependent Recognition Results (% correct) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Phoneme | MFCC | | DCT | | FEATURE | | TRAJ | |
| | Min | Av. | Min | Av. | Min | Av. | Min | Av. |
| /p/ | 88.0 | 89.3 | 84.0 | 86.0 | 75.0 | 78.3 | 30.0 | 40.3 |
| /b/ | 81.0 | 82.3 | 79.0 | 80.7 | 56.0 | 64.0 | 12.0 | 24.7 |
| /t/ | 90.0 | 91.7 | 89.0 | 89.3 | 85.0 | 86.0 | 71.0 | 77.0 |
| /d/ | 82.0 | 83.3 | 81.0 | 83.7 | 64.0 | 69.7 | 44.0 | 52.7 |
| /k/ | 90.0 | 90.3 | 82.0 | 85.3 | 71.0 | 77.3 | 62.0 | 69.3 |
| /g/ | 74.0 | 76.0 | 73.0 | 75.0 | 45.0 | 51.3 | 14.0 | 17.3 |

Table 4. Speaker dependent phoneme recognition results, for speakers 1, 2 & 3, for consonant class of plosives, for various feature extraction modules.

| Speaker Independent Recognition Results (% correct) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Phoneme | MFCC | | DCT | | FEATURE | | TRAJ | |
| | Min | Av. | Min | Av. | Min | Av. | Min | Av. |
| /p/ | 24.0 | 27.5 | 26.0 | 30.7 | 7.0 | 17.3 | 6.0 | 11.2 |
| /b/ | 13.0 | 22.2 | 14.0 | 18.8 | 18.0 | 23.5 | 2.0 | 9.5 |
| /t/ | 40.0 | 45.3 | 38.0 | 42.7 | 36.0 | 56.0 | 34.0 | 48.2 |
| /d/ | 26.0 | 31.0 | 21.0 | 28.3 | 20.0 | 28.3 | 30.0 | 37.0 |
| /k/ | 28.0 | 36.8 | 27.0 | 32.8 | 10.0 | 19.2 | 38.0 | 44.0 |
| /g/ | 9.0 | 13.5 | 9.0 | 12.0 | 6.0 | 8.0 | 2.0 | 2.8 |

Table 5. Speaker independent phoneme recognition results, for speakers 1, 2 & 3, for consonant class of plosives, for various feature extraction modules.

Tables 6 and 7 show the overall phoneme recognition results for the consonant class of semi-vowels, for speaker dependent and independent recognitions respectively.

| Speaker Dependent Recognition Results (% correct) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Phoneme | MFCC | | DCT | | FEATURE | | TRAJ | |
| | Min | Av. | Min | Av. | Min | Av. | Min | Av. |
| /l/ | 95.0 | 95.7 | 94.0 | 95.3 | 79.0 | 82.3 | 78.0 | 80.7 |
| /j/ | 89.0 | 92.3 | 86.0 | 89.0 | 25.0 | 38.7 | 61.0 | 70.3 |
| /w/ | 92.0 | 94.0 | 92.0 | 93.0 | 76.0 | 78.3 | 62.0 | 67.0 |
| /r/ | 98.0 | 98.0 | 98.0 | 98.0 | 83.0 | 87.3 | 78.0 | 85.0 |

Table 6. Speaker dependent phoneme recognition results, for speakers 1, 2 & 3, for consonant class of semi-vowels, for various feature extraction modules.

| Speaker Independent Recognition Results (% correct) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Phoneme | MFCC | | DCT | | FEATURE | | TRAJ | |
| | Min | Av. | Min | Av. | Min | Av. | Min | Av. |
| /l/ | 48.0 | 59.2 | 45.0 | 55.5 | 38.0 | 46.8 | 35.0 | 45.8 |
| /j/ | 8.0 | 33.0 | 3.0 | 24.3 | 0.0 | 5.3 | 10.0 | 32.5 |
| /w/ | 38.0 | 46.2 | 39.0 | 47.7 | 43.0 | 46.5 | 22.0 | 43.8 |
| /r/ | 53.0 | 70.2 | 44.0 | 64.2 | 39.0 | 49.2 | 58.0 | 69.0 |

Table 7. Speaker independent phoneme recognition results, for speakers 1, 2 & 3, for consonant class of semi-vowels, for various feature extraction modules.

Tables 8 and 9 show the overall phoneme recognition results for the consonant class of nasals, for speaker dependent and independent recognitions respectively

| Speaker Dependent Recognition Results (% correct) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Phoneme | MFCC | | DCT | | FEATURE | | TRAJ | |
| | Min | Av. | Min | Av. | Min | Av. | Min | Av. |
| /m/ | 88.0 | 90.3 | 91.0 | 92.7 | 69.0 | 71.0 | 49.0 | 51.7 |
| /n/ | 96.0 | 96.7 | 97.0 | 97.0 | 94.0 | 95.7 | 87.0 | 88.7 |
| /ng/ | 76.0 | 77.3 | 78.0 | 81.3 | 32.0 | 45.7 | 9.0 | 17.7 |

Table 8. Speaker dependent phoneme recognition results, for speakers 1, 2 & 3, for consonant class of nasals, for various feature extraction modules.

| Speaker Independent Recognition Results (% correct) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Phoneme | MFCC | | DCT | | FEATURE | | TRAJ | |
| | Min | Av. | Min | Av. | Min | Av. | Min | Av. |
| /m/ | 24.0 | 39.8 | 36.0 | 45.7 | 20.0 | 29.3 | 22.0 | 42.2 |
| /n/ | 41.0 | 54.2 | 38.0 | 51.7 | 63.0 | 67.3 | 35.0 | 56.3 |
| /ng/ | 4.0 | 7.5 | 7.0 | 7.5 | 0.0 | 3.5 | 0.0 | 2.2 |

Table 9 Speaker independent phoneme recognition results, for speakers 1, 2 & 3, for consonant class of nasals, for various feature extraction modules.

## PERFORMANCE EVALUATION

The implementation of the proposed approach was evaluated at the speech frame level, on a relatively small corpora of Australian accented English database. Across the three speakers (two females and one male), this approach produced an average consonant phoneme recognition accuracy, for the speaker dependent mode in the range of 46.9% to 95.0%, and for the speaker independent mode in the range of 25.4% to 57.0%, depending on the feature extraction module.

Since it is more common to quote phoneme recognition results at the phoneme segment level, rather than the frame level, a simple error correction technique was implemented to correct up to two consecutive errors within a phoneme segment. The corresponding phoneme recognition accuracy, for the speaker dependent mode was in the range of 61.0% to 98.2%, and for the speaker independent mode in the range of 37.8% to 62.0%, depending on the feature extraction module.

For speaker dependent recognition, the best performing features were MFCC and DCT, with an average recognition rate of between 93.4% to 98.2%. For the speaker independent recognition, the best performing feature was FEATURE, with an average recognition rate of between 43.3%-60.3%.

The performance evaluation indicates that for speaker dependent recognition, spectrally based features such as MFCC and DCT coefficients perform better than the features based on acoustic-phonetics. This may be due to the fact that these features were probably not optimum for the task on hand. For speaker independent recognition, the performance of all the feature modules were similar. One possible explanation may be that since each decision tree was trained on a single speaker and tested on the other two, the decision tree could not be classified to be truly speaker independent.

## DISCUSSION

The proposed recognition strategy helps to reduce the *bottleneck* in the acquisition of knowledge from the expert. It also automatically generates production rules, which are derived from examples in the training database instead of being generated *heuristically* by the expert. This process also eliminates the traditional problem in rule base systems, where the number and complexity of the production rules increases to such an extend that it is very difficult for *experts* to manage the large number of interrelated rules. This method also helps to overcome one of the major difficulties in rule based systems, that is the ability to quantify, from a given speech parameters, a set of rules that can reliably identify a class of sound, and still manage to take into account *inter* and *intra* speaker speech variability.

## CONCLUSION

This paper demonstrated a data-driven knowledge/rule based approach to phoneme recognition of the phonetic consonant classes of plosives, semi-vowels and nasals. for a combination of feature sets using C4.5 Inductive System. The experimental results indicate the ability of this approach to generate reliable production rules that can be used to perform the recognition task on the unknown utterances. Although the reported performance still fails short of other recognition techniques, this may be greatly attributed to the very small number of speakers used in these performance evaluations. The proposed approach has the potenetial to allow the true integration of features from existing signal processing techniques that have proven to produce good results in stochastic modelling with acoustic-phonetic features, including the incorporation of speech specific knowledge into the decision tree.

## NOTES

(1) This work was carried out at the Speech Technology Research Laboratory, Department of Electrical Engineering, The University of Sydney.

## REFERENCES

Aikawa, K. (1986). *Automatic Generation of Consonant Discrimination Rules*, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, pp. 2755-2758, April 1986

Bulot, R. and Nocera, P. (1989). *Explicit knowledge and neural networks for speech recognition*, Eurospeech 89, Paris, France, Vol. 2, September 1989, pp 533-536.

De Mori, R. and Kuhn, R. (1992). *Speech Understanding Strategies Based on String Classification Trees*, Proc. ICSLP 92, Vol. 1, pp. 441-459, Canada, 1992.

De Mori, R. and Lam, L. (1986). *Plan refinement in a knowledge-based system for automatic speech recognition*, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, April 1986, pp 1217-1220.

Diest, M. V., Compernolle, D. V. and Oosterlinck, A. (1989). *A rule based system for speech verification*, Eurospeech 89, Paris, France, Vol. 2, September 1989, pp 248-251.

Fohr, D., Carbonell, N. and D., Haton (1989). *Phonetic decoding of continuous speech with APHODEX expert system*, Eurospeech 89, Paris, France, Vol. 2, September 1989, pp 609-612.

Horn K. A. (1991). *RD-ID3: A System for Knowledge Acquisition and maintenance Employing Induction with Ripple Down Rules*, OTC Technical Report, OTC R&D, December, 1991.

Komori, Y., Hatazaki, K., Tanaka, T., Kawabata, T. and Shikano, K. (1989). *Phoneme recognition expert system using spectrogram reading knowledge and neural networks*, Eurospeech 89, Paris, France, Vol. 2, September 1989, pp 549-552.

Quinlan, J. R. (1986). *Induction of decision trees*, Machine Learning, Vol. 1, No. 1.

Samouelian, A. (1992). *Acoustic Feature Extraction Framework for Automatic Speech Recognition*, Fourth Australian Int. Conf. on Speech, Science and Technology, December 1992, Brisbane, Australia, pp 629-634.

Samouelian, A. (1994). *Knowledge based approach to consonant recognition*, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Adelaide Australia, April 1994, pp 177-180.

Zue, V. W. and Lamel, L. F. (1986). *An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition*, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, pp. 1197-1200, April 1986.