KNOWLEDGE BASED APPROACH TO CONSONANT RECOGNITION

A. Samouelian

Department of Electrical and Computer Engineering, University of Wollongong,
Northfields Avenue, Wollongong, NSW 2522, Australia.
ara@uow.edu.au

ABSTRACT

This paper presents a knowledge based approach to consonant recognition. In traditional knowledge based systems, the expert is the linguist/phonetician who attempts to describe and quantify the acoustic events, in the form of *production rules* into phonetic description. This paper proposes to alter the expert's role so that the expert only needs to provide the basic structure of the phonetic classification. The knowledge itself can then be induced from examples in the agreed structure. Thus the acoustic-phonetic rules are moved from the expert's head to the machine memory via the language of examples rather than via the language of explicit articulation. Recognition results on three broad phonetic classes, namely *plosives*, *semi_vowels* and *nasals*, for a combination of feature sets, for speaker dependent and independent recognition, are presented.

1. INTRODUCTION

One of the major difficulties in automatic speech recognition (ASR) is the extreme variability of the speech signal at the speaker and acoustic-phonetic level. Pattern recognition approaches can handle this variability to some extend by being data driven, but generally ignore acoustic-phonetic features. The use of knowledge/rule based approach to continuous speech recognition has been proposed by several researchers and applied to speech recognition [1], spectrogram reading [2, 3] and speech understanding systems [4]. In general, the *production rules* are in the form of IF *condition* THEN *action.* In speech recognition, there is always the need to generate additional rules to handle exceptions. Very quickly, the number and complexity of the rules increases to such an extent that it is very difficult, for human experts, to generate *heuristically*, a large number of interrelated rules, from empirical linguistic knowledge or from the observations of the speech data. To overcome this problem, Aikawa [5] proposes an automatic rule generation approach for rule-based consonant discrimination. The rules are generated automatically from the database.

Inductive systems have been widely used for collection of classification knowledge from large databases and collections of examples. The essence of induction is to use a known set of examples to a theory that explains both these examples and, hopefully, other unseen examples as well [6]. Since the most successful recognition systems are data driven, where the structure and characteristics of the speech signal is captured *implicitly* from the training data, this paper proposes a data driven knowledge-based approach to consonant recognition in continuous speech. The system is based on automatic generation of production rules from examination of hand segmented and labelled database by the use of an induction system (C4.5) [7]. The approach is based on two assumptions. First, it assumes that data driven methodology is the way to solve the problem of *inter* and *intra* speaker speech variability. Second, it assumes that inductive learning has the ability to generalise the characteristics of the speech signal *explicitly* from the database.

The motivation here is to develop a flexible ASR system that allows the generation of *production rules* from any number of different feature combinations, including traditional acoustic-phonetics, speech specific or spectrally based features or parameters.

Section 2 introduces the training and recognition strategy. Recognition results for a relatively small database are presented for the consonant class recognition in section 3. The performance evaluation of the recognizer using four different feature extraction modules are shown in section 4. Section 5 discusses the advantages of inductive systems for speech recognition with section 6 concluding this paper.

2. TRAINING AND RECOGNITION STRATEGY

2.1 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 recordings were made in an anechoic room. The speech was sampled at 40 kHz, and digitised into 16-bit samples. The speech was down sampled to 16 kHz for final analysis by the feature extraction modules. 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.

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

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 consonant classes of *plosives*, *semi_vowels* and *nasals* were selected for investigation since these tend to be more difficult to recognize. Table 2 shows the number of phonetic class tokens for each speaker. These tokens are according to the hand labelled transcription by the phonetician.

Consonant Class	Number of Tokens		
	Speaker 1	Speaker 2	Speaker 3
Plosive	1140	1130	1049
Semi_vow	655	570	558
Nasal	599	561	560

Table 2. Number of phonetic class tokens for each speaker.

2.2 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 first feature extraction module (MFCC) generates MFCC coefficients. The second module (DCT) is an auditory front end [8] based on the Generalised Synchrony Detector (GSD) proposed by Seneff [9] and modified so that the synchrony output is transformed by Discrete Cosine Transform (DCT) function to produce a set of coefficients similar to MFCC. Both of these modules generate 12 MFCC/DCT, 12 delta MFCC/DCT and an energy term. The third modules (TRAJ) extracts formant and formant transition information from the output of the auditory model. The final module (FEATURE) extracts from the speech signal, various time domain features such as root mean square (rms), maximum amplitude, zero crossing rate, voicing, energy, envelope, AC peak to peak, difference between maximum and minimum values in the positive and negative halves of the signal and auto-correlation peak. Modules 1 and 4 extract features in the time domain, while modules 2 and 3 extract features from the auditory model in the frequency domain.

The feature extraction framework allows the collection of attributes used to describe the phonetic class. These attributes are of two kinds: those whose possible values form a small discrete set, and those whose values were real numbers. For example, the value of attribute F1 trajectory can have values in R, L, F indicating rising, level and falling F1 trajectory, while the value of F1 can be any (positive) real number. Attributes of either kind may have unknown values for a particular phonetic class, and these are designated with a question mark (?).

During the training phase, the feature extraction framework extracts features or parameters from the continuous speech on a frame by frame basis. The time aligned phonetically labelled files are then used to associate each frame with its corresponding label and generate a training data file. This data file is constructed on the basis of the training samples, which contain labelled examples in the form (X,α), where X is a feature vector and α is the corresponding class.This data file is 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, namely MFCC, DCT, FEATURE and TRAJ that were used to test the recognition system.

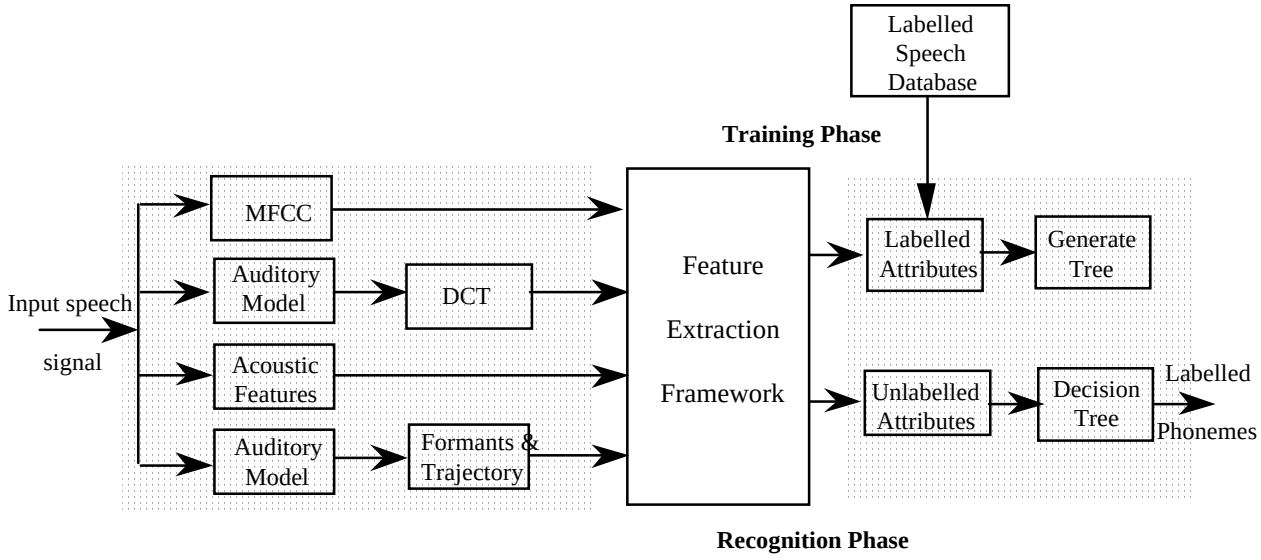

Figure 1. Block schematic of training and recognition strategy

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.

2.3 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 [10].

2.4 Combination of Features

To evaluate the significance of the features combinations in improving or degrading the recognition performance, the features initially selected for the TRAJ module were augmented with a selection of features from the FEATURE module. This resulted in four new feature modules, namely t1_TRAJ, t2_TRAJ, t3_TRAJ and t4_TRAJ. Table 4 shows the combination of features used for each variation on TRAJ.

3. RECOGNITION RESULTS

Tables 5 and 6 show the overall performance, 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.

Parameters of Feature Extraction Modules			
t1_TRAJ	t2_TRAJ	t3_TRAJ	t4_TRAJ
F1	F1	F1	F1
F2	F2	F2	F2
F3	F3	F3	F3
F4	F4		F4
F5	F5		
trajF1	trajF1	trajF1	trajF1
trajF2	trajF2	trajF2	trajF2
trajF3	trajF3	trajF3	trajF3
trajF4	trajF4		trajF4
trajF5	trajF5		
rms	rms	rms	rms
zcr	zcr	zcr	zcr
	envelope	envelope	envelope
	auto_peak	auto_peak	auto_peak

Table 4. The combination of feature set used for variations on TRAJ feature extraction module.

Speaker Dependent Recognition Results (% correct)								
Phone	MFCC		DCT		FEATURE		TRAJ	
Class	Min	Av.	Min	Av.	Min	Av.	Min	Av.
Plos	96	97	97	97.3	96	96.3	79	82
S_vow	93	93.7	93	93.3	75	76.7	69	75.3
Nasal	93	93.7	93	93.3	82	83.3	79	82.7

Table 5. Speaker dependent recognition results, for consonant class recognition, for various feature extraction modules, for speakers 1, 2 & 3 .

Speaker Independent Recognition Results (% correct)								
Phone	MFCC		DCT		FEATURE		TRAJ	
Class	Min	Av.	Min	Av.	Min	Av.	Min	Av.
Plos	84	87	83	87.3	86	90	65	78.2
S_vow	56	65.8	41	53.8	53	60.5	44	56.2
Nasal	13	37.5	33	48	47	56.7	14	48.8

Table 6. Speaker independent recognition results, for consonant class recognition, for various feature extraction modules, for speakers 1, 2 & 3.

Tables 7 and 8 show the overall performance, for speaker dependent and independent recognitions respectively, for speakers 1, 2 & 3 in % correct, for the variations on TRAJ feature extraction module.

Speaker Dependent Recognition Results (% correct)								
Phone	t1_TRAJ		t2_TRAJ		t3_TRAJ		t4_TRAJ	
Class	Min	Av.	Min	Av.	Min	Av.	Min	Av.
Plos	89	92.7	90	92.3	91	92.7	91	92.7
S_vow	78	81.7	82	86.7	82	86.3	83	86.7
Nasal	78	84	88	89	88	88.7	88	88.7

Table 7. Speaker dependent recognition results, for consonant class recognition, for variations on TRAJ feature extraction module, for speakers 1, 2 & 3 .

Speaker Independent Recognition Results (% correct)								
Phone	t1_TRAJ		t2_TRAJ		t3_TRAJ		t4_TRAJ	
Class	Min	Av.	Min	Av.	Min	Av.	Min	Av.
Plos	74	83.5	79	85	79	85.2	79	84.2
S_vow	60	67	61	67.7	59	68.7	62	70
Nasal	13	45.8	16	48.2	15	45.3	17	45.7

Table 8. Speaker independent recognition results, for consonant class recognition, for variations on TRAJ feature extraction module, for speakers 1, 2 & 3 .

4. 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 (2 Females and 1 Male), this approach produced an average consonant class recognition accuracy, for the speaker dependent mode in the range of 80.0% to 94.8%, and for the speaker independent mode in the range of 61.0% to 69.0%, 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 94% to 95.3% between the three consonant classes. The average recognition rate between all of the feature extraction modules was between 87% to 90.1%. For the speaker independent recognition, the best performing feature was FEATURE, with an average recognition rate of between 62.0%-76.3%. The average recognition rate between all of the feature extraction modules was between 51.5% to 76%.

It can be seen from table 7 that the performance of the recognizer for the speaker dependent mode can be improved by the choice of the feature combinations. By the inclusion of features *zcr* and *rms*, a minimum recognition improvement of 10% (plosives) and 9% (semi_vowels) was achieved. By th einclusion of a further two features, *envelope* and *auto_peak*, the recognition improved by a further 1-2% (plosives), 4-5% (semi_vowels) and 9% (nasals). Similar performance improvements were obtained for speaker independent mode as can be seen from table 8.

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. The performance results also indicate that the feature combination does play a significant role in improving the recognition accuracy.

5. DISCUSSION

Traditional knowledge/rule based ASR systems rely on the *expert* to describe and quantify the acoustic events, in the form of *production rules* into phonetic description. The data driven rule based approach using inductive learning from examples helps to eliminate the *bottle-neck* in transferring knowledge from the expert to the system developer. In addition, inductive learning can quantify from the parameters a set of rules that can reliably identify a class of sound. A further advantage is that a large database can be used to extract the knowledge automatically and generate a decision tree which is derived according to the parameters or features that provide most information about a classification. Thus only those parameters are used as discriminating features. This also helps to optimise the size of the feature set that needs to be extracted.

6. CONCLUSION

This paper demonstrated a data-driven knowledge/rule based approach to broad consonant classification, namely *plosives*, *semi_vowels* and *nasals*, for a combination of feature sets using inductive learning to generate the production rules in the form of decision trees and an inference engine to classify the firing of the rules. The experimental results indicate the ability of this approach to solve the problem of *inter* and *intra* speaker speech variability by the use of a large speech database, and the ability to generate decision trees using any combination of features (parametric or acoustic-phonetic). This paves the way for 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.

7. REFERENCES

[1] Bulot, R. and Nocera, P., "Explicit Knowledge and Neural Networks for Speech Recognition", Eurospeech 89, Paris, France, Vol. 2, pp. 533-536, September 1989.

[2] Komori, Y., Hatazaki, K., Tanaka, T., Kawabata, T. and Shikano, K., "Phoneme Recognition Expert System Using Spectrogram Reading Knowledge and Neural Networks", Eurospeech 89, Paris, France, Vol. 2, pp. 549-552, September 1989.

[3] Zue, V. W. and Lamel, L. F. "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.

[4] De Mori, R. and Kuhn, R., "Speech Understanding Strategies Based on String Classification Trees", Proc. ICSLP 92, Vol. 1, pp. 441-459, Canada, 1992.

[5] Aikawa, K., "Automatic Generation of Consonant Discrimination Rules", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, pp. 2755-2758, April 1986

[6] Quinlan, J. R., "Discovering Rules by Induction from Large Collections of Examples", Machine Learning, Vol. 1, No. 1, 1979.

[7] Quinlan, J. R., "Induction of Decision Trees", in Expert Systems in Micro Electronic Age, D. Mitchie, ed. Edinburgh: Edinburgh University Press, 1986.

[8] Samouelian, A., "Speech Recognition Front-End Using Auditory Model", Int. Conf. on Signal Proc. '90, Beijing, China, pp 337-340, October, 1990.

[9] Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech processing", Journal of Phonetics, Vol. 16, No. 1, pp77-91, 1988.

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