ACOUSTIC FEATURE EXTRACTION FRAMEWORK FOR AUTOMATIC SPEECH RECOGNITION

A. Samouelian

Speech Technology Research Group
Department of Electrical Engineering,
The University of Sydney

ABSTRACT - This paper presents a feature extraction framework that allows the use of speech knowledge in training a phonetic recognition system. It can train on any combination of features that may be derived from time and/or frequency domains, parametric, acoustic-phonetic and auditory models including speech specific features. The system requires a moderate size, phonetically labeled database. During the training phase, nominated features per frame are automatically extracted and used as a set of attributes to generate a recognition decision tree, using c4.5 Induction Program. During recognition, the feature extraction framework generates the set of attributes, which are then fed through the decision tree, which assigns a phonetic label to each frame. Recognition results on the class of semi_vowels are presented.

INTRODUCTION

Recognition systems vary in their approach to speech knowledge. There are systems that use heuristic rules, which are developed from intense knowledge engineering, to develop acoustic to phonetic mapping and to emulate the spectrogram reading capabilities of a trained phonetician [O'Kane, 1983; Zue, 1985]. Feature extraction techniques, which are used to segment and label the speech signal, are developed using acoustic theory of human speech production and the ability of trained phoneticians to identify sounds or phonemes directly from spectrograms. In reality this is a fairly difficult task, since it requires the capturing of all the complex interrelationships in speech sounds. Others use template matching and stochastic modeling systems that generally ignore acoustic features or make no use of speech specific knowledge and instead rely on spectral representation of the speech signal to either create reference templates or develop stochastic models. These systems require a large database to develop good representative models of the speech signal. While others express speech knowledge within a formal framework using well defined mathematical tools, where features and decision strategies are discovered and trained automatically, using a large body of speech database [Zue et al, 1989].

Although knowledge engineering (developing specific rules to interpret the extracted features and provide the mapping to its corresponding phonetic label) is manageable for a small vocabulary and isolated words systems, for large vocabulary systems, which require phonetic recognition, a large body of rules is required (developed from examination of hundreds of speech waveforms and their spectrograms). These rules utilize enormous number of acoustic-phonetic, lexical, syntactic, semantic and prosodic facts and the subtle interaction between them makes this task truly formidable.

To parametize the speech signal, most recognizers use the speech production model, which separates excitation and vocal tract response. For each frame, excitation is typically represented by an overall amplitude or energy term. For the spectral representation 8-14 coefficients are generally used to represent the spectral parameters. These coefficients are usually derived from Linear Predictive Coding (LPC) analysis, Fourier Transform, or bank of bandpass filters. Common parameters are

LPC coefficients, Mel Frequency Cepstral Coefficients (MFCC) and energies in the filterbank. These parameters are classified as features and used to train the recognizer.

A new feature extraction framework is presented that allows any combination of features derived from time and/or frequency domains, parametric, acoustic-phonetic and auditory models including speech specific features to be automatically extracted from input speech signal. These features are used as a set of attributes to train and generate a decision tree, using c4.5 Induction Program (Quinlan, 1983).

INDUCTIVE INFERENCE

Inductive inference has been used to extract classification knowledge from large data bases and collections of examples (Quinlan, 1983; Quinlan et al, 1986). Inductive inference produces decision trees that use attributes that provide the most information about classification are chosen as discriminating attributes. In cases when all or the majority of attributes are numeric, induction can produce a very large and unnecessarily complex decision trees. To simplify the complex tree, branches which do not contribute significantly to the accuracy are pruned off (Quinlan, 1987). The problem with this approach is that a significant branch may be pruned off specially if there are no sufficient examples in the training set. The accuracy of these trees can be greatly increased by using Ripple Down Rules to maintain the tree after induction (Horn, 1991).

Some of the advantages of using inductive learning technique are:

- Examination of database containing many examples allows generalizations.
- A decision tree can be generated using any set of attributes without discriminating between rule based or parametric features.
- Parametric features such as LPC or MFCC coefficients are not well suited for rule based systems, since it is difficult to explicitly associate coefficient values with acoustic or phonetic events. The inductive system can easily examine all of the database and set up appropriate thresholds to generate a decision tree and a set of rules for phonetic classification.
- It allows the true integration of features from existing signalling processing techniques that have proven to produce good results in stochastic modeling, and at the same time allows the incorporation of speech specific knowledge into the decision tree.
- It allows the development of decision trees in planned refinement of the rules if the performance is inadequate. This is achieved through hand modification of the decision tree by changing the features or the combination of features used to classify a specific sound or phoneme class. The planned refinement of the rules and the inductive learning technique should make the task of rule based system manageable and provide a productive tool for evaluating the feature sets, assessing the performance of the recognizer and monitoring the incremental improvement in recognition accuracy as a function of the combination of features.

TRAINING AND RECOGNITION STRATEGY

The feature extraction framework allows the extraction of nominated set of features from the input speech signal and creates the appropriate "data" and "bulk" files for training and testing of the recognition system respectively. The "data" file contains all the attributes per frame with the appropriate phoneme labels appended to the end, while the "bulk" file contains only the attributes per frame.

During the training phase, using c4.5 Induction program, a decision tree is generated from the "data" file. During the recognition phase, the "bulk" file is classified by the decision tree, which involves

appending a phonetic label to each frame. The recognition performance is evaluated by comparing the data and classified files per frame, using "perl" programming language. The program corrects single, double or triple consecutive errors in a phoneme segment, and produces the appropriate confusion matrix. The c4.5 induction program can also generate a set of rules in the form of IF...THEN statements, which allows for the manual examination of the attributes that are used to discriminate the phonemes and identify the ones that are acting as "noise". This allows for an informed reduction in the number of features that need to be extracted. These rules can be hand modified and/or new rules can be added to allow for cases that may not be covered in the database or to incorporate speech specific knowledge.

A block schematic of the training and recognition strategy is shown in Figure 1. Four different types of feature extraction modules have been used to test the feature extraction framework and evaluate the recognition accuracy of the decision tree.

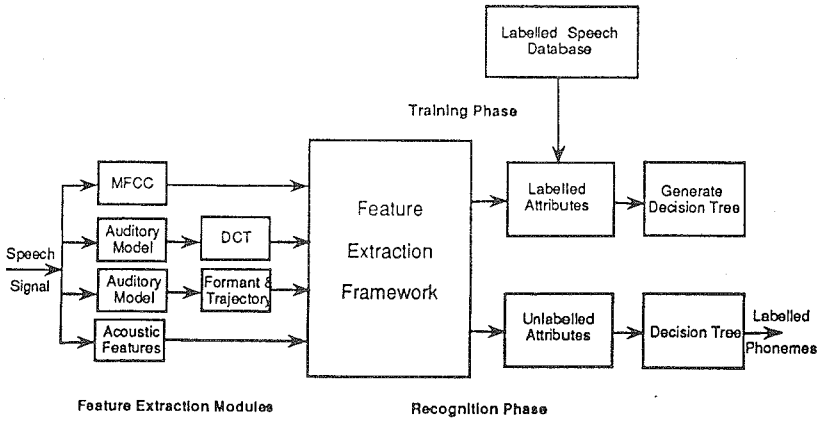

Figure 1: Block schematic of training and recognition strategy

The first feature extraction module is MFCC coefficients. MFCC and LPC coefficients are the most commonly used parametric features used in template based and stochastic modeling systems. The second module is an auditory front end (Samouelian & Summerfield, 1989; Samouelian, 1990) modified so that the synchrony output instead of generating a pseudo spectrogram, it is transformed by Discrete Cosine Transform (DCT) function to produce a set of coefficients similar to MFCC coefficients. Both of these modules generate 12 MFCCC/DCT, 12 delta MFCC/DCT and an energy term. The third modules extracts formant and formant transition information from the output of the auditory model. The final module extracts various features from the speech signal. These features are Root Mean Square (RMS) value, 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 autocorrelation 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.

DATABASE

The class of sounds known as semi_vowels /l,j,w,r/ has been selected to test the feature extraction framework since their reliable identification and fine phonetic classification have been particularly difficult to achieve because of the high degree of spectral and temporal variability of these phonemes (Samouelian &Vonwiller, 1990).

The training and test data were collected from two females and one male. For each speaker, the database consisted of 195 phonetically balanced Australian accented English sentences devised and collected by National Acoustic Laboratories as part of the GLASS projects. The sentences were phonetically segmented by The University of Sydney as part of the GLASS project.

PERFORMANCE EVALUATION

For each feature extraction module, three different decision trees were generated for each speaker, namely:

1. Trained and tested on 100% of data.
2. Trained on top 75% of data and tested on remaining 25% of data.
3. Trained on top 50% of data and tested on remaining 50% of data.

Furthermore, intra-speaker tests were performed such that a decision tree trained on 100% of data of speaker1 was used to test 100% of speaker2 and speaker3 data.

Figure 2 and 3 show the recognition results for each feature extraction module. The decision tree was trained on the top 75% of speaker2 data and tested on the remianing 25%. Since the recognition is at the frame level instead of phonetic segment, simple error correction was introduced to eliminate errors of up to 3 consecutive frames.

Table 1 shows the recognition results for the DCT feature extraction module, before and after the inclusion of one additional rule for identification of phoneme /j/. The results show better recognition scores for /l/ and /j/ and worse scores for /w/ and /r/. This highlights the difficulty in hand modifying rules without using Ripple Down Rules to maintain the tree after induction.

Phoneme	*Recognition Results*			
	No error correction		*3 errors corrected*	
	Original rules	Rules hand modified	Original rules	Rules hand modified
/l/	42%	77%	51%	89%
/j/	0%	95%	0%	100%
/w/	73%	43%	80%	62%
/r/	57%	52%	61%	54%

Table 1: Recognition results for DCT feature extraction module

CONCLUSION

This paper has shown the possibility of using various parametric and/or acoustic-phonetic features to generate a decision tree using c4.5 Induction program for the recognition of fine phonetic classification of semi_vowels. Preliminary results indicate that it is possible to improve the recognition accuracy by hand correcting the decision tree. Future work will concentrate on selecting an optimum set of features to optimize the recognition of semi_vowels.

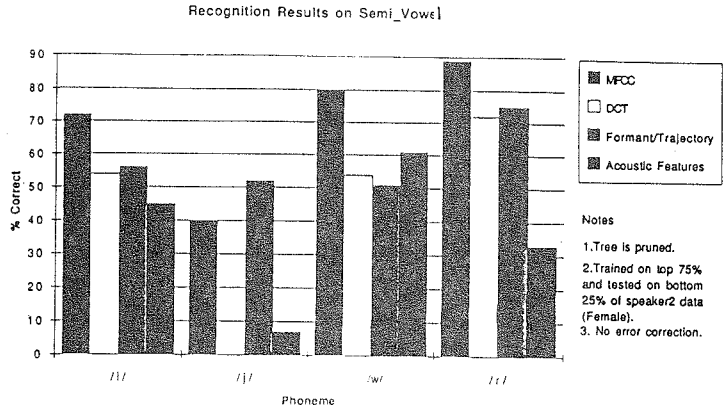

Figure 2: Recognition results for each feature extraction module for semi_vowels using top 75% of speaker2 data to train and tested on remainig 25% and no error correction.

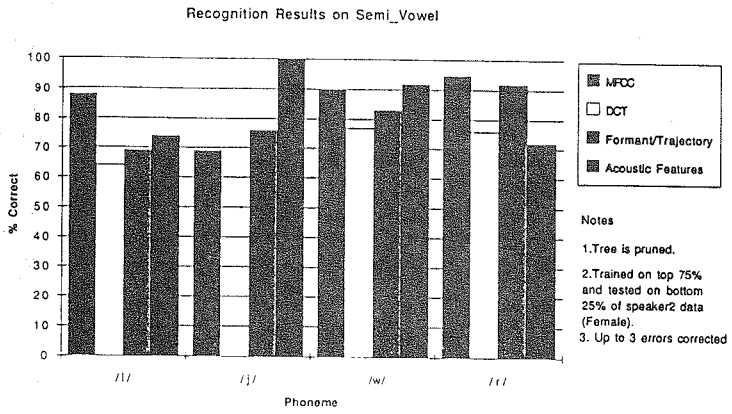

Figure 3: Recognition results for each feature extraction module for semi_vowels using top 75% of speaker2 data to train and tested on remainig 25% and up to 3 consecutive errors are corrected.

ACKNOWLEDGMENT

This work has been carried out within the framework of the GLASS Consortium. I like to thank my employer, OTC Australia, for its supprort in terms of equipment and time for my participation in the GLASS Consortium. I like to thank Mr. Kim Horn from Artificial Intelligence group of OTC R&D for suggesting the possible use of Induction technique to speech database and making available the c4.5 Induction program and developing appropriate programming tools for hand modification of the rules.

REFERENCES

Horn K. A. (1991) "RD-ID3: A system for Knowledge Acquisition and Maintenece Employing Induction with ripple down Rules", OTC Technical report, OTC R&D, December, 1991.

O'Kane M. (1983) "The FOPHO speech recognition project", Proceedings of the Eight International Joint Conference on Artificial Intelligence, Karlsruhe, pp 630-632.

Quinlan J. R. (1983) "Learning Efficient Classification Procedures and Their Applications to Chess End Games",in Machine Learning: An Artificial Intelligence Approach, R. S. Michalski et al., eds, Palo Alto. CA; Tiago Publishing Co., 1983.

Quinlan J. R., Compton P. J., Horn, K. A. and Lazarus, L. (1986) "Inductive Knowledge Acquisition: a Case Study", in Applications of Expert Systems, Quinlan, J. R., ed, Addison Wesley, 1986.

Quinlan J. R. (1987) "Simplifying Decision Trees", in International Journal of Man Machine Studies, 1987.

Samouelian A. & Summerfield C. D. (1989) "Front-end speech signal processor for speech recognition", Proc. IREECON89 Int. Conf., September 1989, Melbourne, Australia, pp 112-115.

Samouelian A. (1990) "Speech recognition front-end using auditory model", Proc. of Int. Conf. on Signal Proc., 22-26 Oct., 1990, Beijing, China.

Samouelian A. and Vonwiller J. (1990) "Performance of A Peripheral Auditory Model on Phonemes in Combination", Third Australian Int. Conf. on Speech, Science and Technology, November 1990, Melbourne, Australia, pp 154-159.

Zue, V. W. (1985) "The use of speech knowledge in automatic speech recognition", IEEE Proc., Vol. 73, No. 11, november, 1985.

Zue, V., Glass J., Phillips M. and Seneff S. (1989) "Acoustic segmentation and phonetic classification in the SUMMIT system", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, May 1989.