AUTOMATIC LANGUAGE IDENTIFICATION USING INDUCTIVE INFERENCE

A. Samouelian

Speech Technology Research Laboratory
Department of Electrical and Computer Engineering
University of Wollongong

ABSTRACT- Automatic spoken language identification (LID) plays an important part in routing foreign callers to operators who speak the caller's language, or as a front-end to a multi-lingual translation system to route the call to the appropriate translation system. A common approach to spoken language ID is adopted from current speaker independent recognition techniques. These generally involve the development of a phonetic recogniser for each language and then combining the acoustic likelihood scores to determine the highest scoring language. The models are trained using hidden Markov modelling (HMM) or neural networks (NN).

This paper proposes a novel approach to spoken language identification by the use of inductive inference "decision trees". To develop the production rules, the classification models are generated inductively by examining a large speech database and then generalising the pattern from the specific examples. This approach has already been successfully used for isolated digit recognition (Samouelian, 1996). The aim of this research is to demonstrate that inductive learning can provide a viable alternative approach to existing automatic spoken language identification techniques.

The proposed LID is based on automatic speech recognition (ASR) system using inductive inference (Samouelian, 1994a, 1994b). It uses a single decision tree to capture all the complexities of each language, using mel-scaled cepstral coefficients (MFCC) as input. The training database is labelled at the language level. The LID classification is performed at the frame level, using an inference engine to execute the decision tree and classify the firing of the rules. A simple sorting routine is then used to identify the spoken language.

Spoken language identification results using the OGI Multi-language Telephone Speech Corpus (OGI_TS) on the three language task (English, German and Japanese), are presented.

INTRODUCTION

Automatic spoken language identification (LID) plays an important part in routing foreign callers to operators who speak the caller's language, or as a front-end to a multi-lingual translation system to route the call to the appropriate translation system. They can also form an important part of a an automatic system for spoken data retrieval system.

A common approach to spoken language LID is adopted from current speaker independent recognition techniques. These generally involve the development of a phonetic recogniser for each language and then combining the acoustic likelihood scores to determine the highest scoring language. The models are trained using hidden Markov modelling (HMM) (Lamel and Gauvain, 1994; Zissman and Singer, 1994) or neural networks (NN) (Berkling and Barnard, 1994).

Current approaches to LID use phonemes (Zissman & Singer, 1994; Lamel & Gauvain, 1994) and broad phoneme classes (Muthusamy *et al*, 1993). Some of these approached also use phonotactic information such as phoneme bigram (Lamel & Gauvain, 1994) or trigram (Reyes *et al*, 1994). Others use phonemes to develop a word level recogniser (Schults *et al*, 1996). However, most common LID systems use some form of phoneme based approach.

There are also several different architectures for LID systems such as a single integrated architecture consisting of a single global language identifier (Muthusamy *et al*, 1993) or parallel architecture (Zissman, 1996; Yan & Barnard, 1995a), where a separate language dependent recogniser is trained for each language. During the identification phase, the unknown utterance is evaluated in parallel by all the language dependent recognisers to determine the best score.

This paper proposes a novel approach to automatic language identification by the use of inductive inference "decision trees". To develop the production rules the classification models are generated inductively by examining the speech database and then generalising the pattern from the specific examples. This approach has already been successfully used for isolated digit recognition [Samouelian, 1996]. The aim of this research is to demonstrate that inductive learning can provide a viable alternative approach to existing automatic language identification (LID) techniques.

The proposed LID system uses short term spectral parameters in the form of mel frequency cepstral coefficients (MFCC) . The C4.5 inductive system (Quinlan, 1993) generates the decision tree automatically from labelled examples in the training database. The language classification is performed at the frame level, using an inference engine to execute the decision tree and classify the firing of the rules. A simple sorting routine is then used to identify the language of the unknown utterance.

This approach has the ability to generate decision trees using any combination of features (parametric or acoustic-phonetic). 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. Language identification results on OGI multi-language telephone speech corpus for closed and open set identification, on the three language task (English, German and Japanese), are presented.

TRAINING AND RECOGNITION STRATEGY

Speech database

The Oregon Graduate Institute Multi-language Telephone Speech Corpus (OGI_TS) (Muthusamy *et al*, 1992) is used to evaluate the LID system. The corpus contains 90 speakers for each of the 11 languages divided into training (50 speakers), development (20 speakers) and test (20 speakers). Each utterance was collected over a long distance telephone line and comprised of responses to ten prompts.

Three languages: American English, German and Japanese were selected for investigation. These are all stress languages. English and German belong to the same Germanic group of Indo-European languages, while Japanese is a member of the Altaic family.

A single utterance from each speaker for each language was used from the database. These were "story-before-the-tone" (story-bt) utterances. These are natural continuous speech (text independent) in response to the prompt asking the speaker to speak about any topic of his/her choice of about 45 s in duration.

The database was divided into training and test as shown in Table 1. The first 10 speakers of each of the languages in the database was selected for training. The remainder of the speakers in the Training Database (Test1) and all the speakers in the Development Database (Test2) were selected for testing. The "45 s" spontaneous speech utterances of the Development database were further divided into "10 s" chunks for testing. None of the train and test sets overlapped.

Language	Training (Training Data)		Test1 (Training Data)		Test2 (Development Data)	
	Male	Female	Male	Female	Male	Female
English (EN)	7	3	28	12	16	4
German (GE)	4	6	20	20	16	4
Japanese (JA)	7	3	21	19	13	7

Table 1. Training and testing subsets of the database

Thus the database was subdivided into three "45 s" sets and one "10 s set":

- Training set. There were 30 "45 s" utterances in this subset.
- Training Database (Test1). There were 120 "45 s" utterances in this subset.
- Development Database (Test2). There were 60 "45 s" utterances in this subset.
- Development Database (Test2). There were 251 "10 s" utterances in this subset.

For training, the speech files were not hand segmented at either the phoneme or word levels. instead the whole utterance was labelled as English, German or Japanese. Note that silences were included as part of each language and were not identified either explicitly or implicitly. The aim here was to see if there is sufficient short term spectral differences between the three languages at the language level.

Front-end signal processing

The speech was downsampled to 8 kHz, pre-emphasised by a first-order filter ($1-0.95z^{-1}$) (Song and Samouelian, 1993). The signal was then processed by a 256 point (32 ms) Hamming window with a frame shift of 128 points (16.0 ms). A 256 point FFT was performed on each windowed speech portion. A set of 19 mel scaled triangular filter bank was applied to the FFT power spectrum. The log-energy of the 19 outputs was calculated and transformed to 12 mel frequency cepstral coefficients (MFCC) by discrete cosine transform (Davis & Mermelstein). In order to enhance the distortion measure used in the identification system, raised cosine function was used to weight the acoustic features. Twelve delta MFCCs were also computed to represent the transitional information. Finally, a delta energy term (time derivatives of the normalised energy) was included taking the total number of coefficients to 25.

Training

A block schematic of the training and recognition strategy is shown in Figure 2. During the training phase, the feature extraction framework extracted the MFCC parameters from the speech signal on a

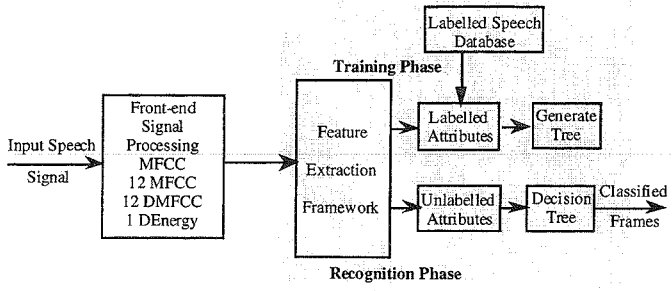

Figure 1. Block schematic of training and recognition strategy

frame by frame basis. Each frame was then automatically labelled as one of the three languages to generate the training data file. The data file contained labelled examples in the form (X,a), where X is the feature vector and a is the corresponding class. This data file was then used by the C4.5 program to generate the decision tree.

RECOGNITION RESULTS

Recognition strategy

The LID system was tested on the training and two test corpora. Each test utterance came from a different speaker and it was of about 45 s in duration. These "45 s" utterances of the Development Data were further subdivided into 10 s portions to generate the "10 s" utterances. A one three-way closed-set and open set tests were performed on the training and test data.

The test messages spoken in English, German and Japanese were classified at the frame level by the decision tree. A simple sorting routine was then used to sort and rank the number of frames according to their classification. The language with the highest frame score, was chosen as the recognised language. Table 2 shows the global accuracy for "45 s" and "10 s" utterances.

	"45-s" Test			"10-s" Test
Language	Training	Test1	Test2	Test2
English	100%	62.5	80.0	71.6
German	100%	55.0	40.0	39.5
Japanese	100%	30.0	40.0	35.7
Average	100%	49.2	53.3	48.6

Table 2. Global LID accuracy for "45-s" and "10-sec" utterances

The error rate for the closed-set was 0%, while for open-set the error rate varied by 5% between Test1 and Test2, probably due to the different data size. The average LID error rate decreased by 5.3% when tested on the "10 s" segments compared to the "45 s" segments. Tables 3 and 4 show the confusion matrices for the "45 s" and "10 s" utterances of the Development Data for the three languages.

	EN	GE	JA	% Correct
English	16	4	0	80.0
German	9	8	3	40.0
Japanese	2	10	8	40.0
Total	27	22	11	53.3

Table 3. Confusion matrix for "45-s" utterances

	EN	GE	JA	% Correct
English	58	18	5	71.6
German	37	34	15	39.5
Japanese	13	41	30	35.7
Total	108	93	50	48.6

Table 4. Confusion matrix for "10-s" utterances

PERFORMANCE EVALUATION

Using only acoustic parameters, the proposed approach, for a three language identification task produced an average recognition of 53.3% for "45 s" utterances and 48.6% for "10 s" utterances. To compare the performance of this LID system against other systems, it is important that the size and composition of the training and test data are the same. Berkling *et al* (1994) have reported an average identification rate of 74.2% using a two-stage system on a three language task (English, German and Japanese). The first stage uses an NN trained at phoneme level to classify the languages at the frame level. This is followed by a language classification stage using unigram and bigram features. However, the training and test data sizes are different from the one used in this paper.

On a six language LID task (English, German, Japanese Mandarin, Hindi and Spanish), Berkling & Barnard (1994) using the technique of phoneme clustering across all the six languages, and building a recogniser to classify broad phonemes, reported an average phoneme recognition rate of 59% across the six languages. For the same six languages LID task, Yan & Barnard (1995b) reported an average phoneme recognition rate of 48.1% using language dependent phone recognition. Using a NN classifier and all sets of features, they reported an average language recognition rate across the six languages of 91.96% for "45 s" utterances and 81.82% for the "10 s" utterances.

Zissman (1996) has performed LID experiments on a three language task (English, Japanese and Spanish) from the OGI_TS database. He used a language dependent parallel phone recognition (PPR) system. The average identification rate across the three languages was 49.2%.

DISCUSSION

The performance of the proposed system falls short of current stochastic techniques (HMM and NN), using phoneme recognition and some form of language model, to perform language identification. However the performance results indicate that it is possible to use inductive learning for language identification and it provides a baseline measure for future system enhancements.

In this experiment, a single decision tree was developed to capture all the acoustic complexities of the three languages, using mel-scaled cepstral coefficients as input. The identification was at the language level instead of the phoneme level . No language models or duration information were used to place a sequential constraint on the sequence of phonemes and to capture certain prosodic

information of the different languages. Muthusamy *et al* (1994) claim that the approach of using a single HMM model to represent the entire language has not been very successful for LID. A better approach would be to develop a separate stochastic model for each phoneme of the language. The main problem here is the need for phonemically labelled speech database for each of the target languages. Furthermore, using text independent speech, it is difficult to incorporate stochastic grammars to correct the many errors that an unconstrained recogniser would make.

The frame level classification rate should improve further, resulting in improved language level recognition, by using the whole of the training data (50 speakers) instead of a subset of 10 speakers per language. Also silence portions should be included as a classification label and be identified implicitly during the classification process. Future work will also utilise the ability of the c4.5 induction system to integrate the short term spectral parameters (MFCC) with prosodic information such as speech rate, pitch contour and pitch duration in an attempt to enhance the baseline identification performance of the LID.

CONCLUSION

This paper reported on initial results on the use of a novel inductive learning, language identification technique using mel-scale cepstral coefficients as input. The experimental results on the identification of three language task (English, German and Japanese) indicate the ability of this approach to perform language identification using spontaneous speech (text independent) using a single decision tree to represent all the languages.

The experimental results indicate also that inductive learning can generate decision trees using parametric features instead of the traditional acoustic-phonetic feature set and generalise sufficiently to be able to also classify languages at the frame level. The challenge ahead is to enhance the baseline identification performance of the LID, at the language level, by adding prosodic information to the MFCC feature set.

REFERENCES

Berkling, K. M., Arai, T. and Barnard E. (1994). *Analysis of phoneme-based features for language identification,* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Adelaide Australia, April 1994, pp I289-I292.

Berkling, K. M. and Barnard, E. (1994). *Language identification of six languages based on a common set of broad phonemes.* Proc. of the 1994 International Conference on Spoken Language Processing, Vol. 4, pp. 1891-1894, Yokohama, Japan.

Davis, S. and Mermelstein P. (1980). *comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,* IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, No. 4, pp. 357-366, Aug. 1980.

Lamel, L. F. and Gauvain, J. L. S. (1994). *Language identification using phone-based acoustic likelihoods.* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Adelaide, Australia, April 1994, pp I293-I296.

Muthusamy, Y. K., Cole, R. A. and Oshika, B. T. (1992). *The OGI multi-language telephone speech corpus.* Proc. of the 1992 International Conference on Spoken Language Processing, Vol. 2, pp. 895-898, Alberta, Canada.

Muthusamy Y. K., Berkling, K., Arai, T., Cole, R. A. and Barnard, E. (1993). *A comparison of approaches to automatic language identification using telephone speech.* Proc. Eurospeech '93, Vol. 2, Sept. 1993, pp. 1307-1310.

Muthusamy, Y. K., Barnard, E. and Cole, R. A. (1994). *Automatic language identification: a review/tutorial;.* IEEE Signal Processing Magazine, October 1994.

Quinlan, J. R. (1993). *C4.5 programs for machine learning,* Morgan Kaufmann series in machine learning, Morgan Kaufmann publishers, USA, 1993.

Reyes, A. A. (1994). *Three language identification methods based on HMMs*, Proc. of the 1994 International Conference on Spoken Language Processing, Vol. 4, pp. 1899-1898, Yokohama, Japan.

Samouelian, A. (1994a). *Knowledge based approach to consonant recognition*, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Adelaide Australia, April 1994, pp I77-I80.

Samouelian A. (1994b). *Knowledge based approach to speech recognition*, Fifth Australian Int. Conf. on Speech, science and Technology, Perth, Australia, pp 479-484, Dec 6-8., 1994.

Samouelian A. (1996). *Isolated voiced digit recognition using inductive inference,* to be published in Proc. IEEE Region Ten Conference, to be held in Perth, Australia, Nov 27-29, 1996.

Schultz, T. *et al.* (1995). *Experiments with LVCSR based language identification*, Proc. SRS, 1995, pp. 89-92.

Song J. M and A. Samouelian (1993). *A Robust Speaker-Independent Isolated Word HMM Recognizer for Operation Over the Telephone network*, Speech Communication, Vol 13, 1993, pp 287-295.

Yan, Y, and Barnard, E. (1995a). *An approach to automatic language identification based on language-dependent phone recognition.* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Vol. 5, May, 1995, pp. 1883-1886.

Yan, Y, and Barnard, E. (1995b). *Neural networks and linear classifiers for automatic language identification,* Int. Conf. on Neural Networks and Signal Processing, Nanjing, China, Dec. 10-13, pp. 800-803, 1995.

Zissman, M. A. and Singer, E. (1994). *Automatic Language identification of telephone speech messages using phoneme recognition and n-gram modelling.* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Adelaide, Australia, April 1994, pp I305-I308.

Zissman, M (1996). *Comparison of four approaches to automatic language identification of telephone speech.* IEEE Transactions on speech and Audio Processing, Vol. 4, No. 1, January 1996.