Comparing Gaussian Mixture and Neural Network Modelling Approaches to Automatic Language Identification of Speech.

J.P Willmore, R.C. Price and W.J.J Roberts

Information Technology Division
Defence Science and Technology Organisation
P.O. Box 1500, Salisbury 5108, South Australia
jonathan.willmore@dsto.defence.gov.au, richard.price@dsto.defence.gov.au,
william.roberts@dsto.defence.gov.au

ABSTRACT

In this paper we compare the performance of two well-known approaches to automatic Language Identification: Gaussian Mixture Modelling and Neural Network modelling. The systems were evaluated with the Oregon Graduate Institute Multi Language Telephone Speech Corpus. In a comparison of the two systems using identical training and testing data, similar performance was obtained.

INTRODUCTION

Automatic language identification refers to the process of recognizing the language spoken from a sample of speech by an unknown speaker. This process may be performed by comparing the utterance from a language of unknown identity with templates or models of various languages of interest. The degree of similarity between the models and the utterance is then used to make a decision. This paper compares two approaches to language identification, and shows the comparative performance of the two systems when trained and tested with identical data.

SPEECH CORPUS

The systems were evaluated with the Oregon Graduate Institute Multi Language Telephone Speech (OGI-TS) Corpus (Muthusamy et. al). This database consists of utterances in English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. For each language, 90 native speakers were required to each speak six spontaneous and four fixed-vocabulary utterances, producing approximately 2 minutes of speech. Test utterances were extracted from the development test set according to the April 1993 National Institute of Standards and Technology (NIST) Specification (Martin).

"45-sec" utterance testing refers to the set of 45-second utterances spoken by the development test speakers. OGI refer to these utterances as "stories before the tone".
"10-sec" utterance testing refers to the same set of 45-second utterances spoken by the development test speakers, only the utterances have been segmented into 10-second segments (Zissman 1996).

GAUSSIAN MIXTURE MODELS

The Gaussian Mixture Models (GMMs) were trained by parameter estimation using the maximum likelihood (ML) criterion. Once trained, classification was performed using the maximum a-posteriori (MAP) decision rule (Roberts and Willmore 1999). The target languages were each modeled by GMMs containing 250 states. The Gaussian mixtures are non-zero mean with diagonal covariance matrices.

NEURAL NETWORKS

Two neural networks were constructed, one to distinguish between English and Japanese and another to distinguish between the complete set of 10 languages of the OGI-TS corpus (excluding Hindi).

NN paradigms such as probabilistic NNs and generalised NNs are known to be unsuitable for problems that have a large number of vectors in the training set, as they attempt to represent each pattern with a node in the network. This causes these techniques to converge slowly and have very large memory requirements. This was also experienced first hand. Back propagation (BP) networks were selected as the most appropriate to use as they trained well, and tested quickly, especially as the training data exceeded 5000 thousand vectors.

Neural Networks are trained on a vector by vector basis and testing is performed in the same manner. Each test utterance is split up into a sequence of vectors, each of which is tested against the neural network. An individual vector will "score" according to how close the vector is to each language. The language with the highest score is chosen as the language identity of the unknown utterance.

All experiments were conducted using the commercially available MATLAB neural network toolkit.

IMPLEMENTATION

Both the GMMs and NNs were trained and tested with feature vectors obtained from the speech waveform. The feature vectors used consist of the first 20 cepstral coefficients (Roberts and Willmore 1999) plus 20 Delta Cepstral coefficients with a delta window size of 3.

A comparison of the performance of GMM and NN Language ID systems was performed using a maximum of 2000 vectors per language. Specifically, for English vs. Japanese we used a total of 4000 training vectors, and for the 10 language NN tests we used a total of 20000 training vectors. The evaluations compared English vs. Japanese, as well as a 10 language forced choice comparison.

The number of vectors being selected at 2000 per language was decided based upon trial and error. For example an initial attempt at training a 10 language NN on 50000 vectors failed to converge after 3 days training on a computer with a 400 MHz CPU and 256MB RAM.

Whilst techniques are available to determine optimal network topologies for the back propagation paradigm, these techniques were not employed due to the overhead in training times this would incur. The network topologies and type of each NN were also established by trial and error. Two BP networks consisting of 40 inputs (20 cepstral coefficients and 20 delta cepstral coefficients), a hidden layer of 150 neurons and an output layer of either 2 or 10 were selected.

The mean square error (MSE) convergence criterion was set at 0.02, but after more than 72 hours neither NNs converged below 0.09 and the NNs ceased training due to the maximum number of epochs being reached. This was commonly around 500 to 1000 for the 10 language NN, and 5000 to 10000 for the 2 language NN. This slow and limited convergence during training proved to be the case regardless of the neural network topology adopted. Multiple hidden layers and different numbers of nodes in the hidden layer were also tried, but failed to solve the problem.

In comparison, ten GMM language models each comprising 250 states were able to be trained with two hours of training data (approximately one million vectors per language) within 24 hours on a 200MHz, 256KB computer.

RESULTS

Table I below shows the results of both the GMM and the NN systems where only 2000 vectors of training data per language have been used. Table II shows the performance of the GMM system where all of the available training data was used.

	Eng/Jap	Eng/Jap	10 lang	10 lang
	45-sec	10-sec	45-sec	10-sec
GMM	16.0	23.1	78.6	78
NN	15.4	17.7	85.8	85.5

TABLE I. Results comparing the GMM and NN systems using only 2000 training vectors per language (% ERROR).

	Eng/Jap	Eng/Jap	10 lang	10 lang
	45-sec	10-sec	45-sec	10-sec
GMM	10.5	9.2	52	55

TABLE II. Results showing the GMM system using the full OGI-TS training data set. (% ERROR).

For the case of English vs. Japanese, it can be seen from Tables I and II that the error rates obtained by using only 2000 training vectors per language are not too dissimilar to those results obtained by using the full training data set. Future research will concentrate on how this result could be used to reduce the training times involved in language identification systems.

CONCLUSIONS

We conclude that the GMM and NN systems have similar performance characteristics when only 2000 training vectors per language are used. It was established with the computing resources available, that NNs with greater than 20000 vectors became wholly infeasible due to the intense consumption of computer resources, especially memory.

In contrast, with the same computer resources, GMMs are capable of training on significantly more data and can take advantage of the full OGI-TS training data set. This is not to say that NNs would not perform as well when given the full training data set, but rather that it was not possible to determine this on our current computing resources. In the very near future however we will have access to a considerably faster machine with at least 1GB RAM and will be able to perform a more complete and comprehensive set of experiments. We will also examine other avenues of approach with NNs whereby such large amounts of memory are unnecessary.

REFERENCES

Muthusamy, Y.K., Cole, R.A. & Oshika, B.T. (1992) "The OGI multi-language telephone speech corpus", *Proc. ICSLP '92,* vol. 2, Oct. 1992, pp 895-898.

Martin, A.F., *LanguageID Guidelines and Results*. Gaithersburg, MD: Nat. Inst. Std. Technol. (NIST), Spoken Language Processing Group.

Zissman, M.A. (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, pp. 91-108.

Roberts, W.J.J & Willmore, J.P. (1999) "Automatic Speaker Recognition Using Gaussian Mixture models," in *Information Decision and Control Conference*, Adelaide, Feb. 1999.

LANGUAGE IDENTIFICATION USING EFFICIENT GAUSSIAN MIXTURE MODEL ANALYSIS

E. Wong, J. Pelecanos, S. Myers and S. Sridharan
Speech Research Lab, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
ee.wong@qut.edu.au, j.pelecanos@qut.edu.au
sd.myers@qut.edu.au, s.sridharan@qut.edu.au

ABSTRACT: Automatic Language Identification (LID) is the automated process of identifying the language of a speech utterance. In this paper, we will describe a language identification system that utilises Mel-Frequency Cepstral Coefficients (MFCCs) and Gaussian mixture models (GMMs) to model the short-term characteristics of a language. We also compare this standard GMM language model to the models that are adapted from a universal, language-independent background model (UBM). Experiments show that model adaptation gave comparable performance. In addition, a computation speed-up approach was tested on the adapted language models. The accuracy of the system remained comparable while the computation time was reduced significantly.

1. INTRODUCTION

Automatic Language Identification (LID) is the process of identifying the language of a speech utterance using a computer. There are several important applications for Language Identification (Muthusamy et al., 1994). For example, telephone companies can handle foreign language calls with a LID system that routes each call to the operator that is fluent in the caller's language. This application can even extend to the handling of emergency call services. A LID system can also serve as a front-end for a multi-language translation system.

To accomplish the task of Language Identification, a variety of methods have been proposed (Muthusamy et al., 1994). These include Hidden Markov models (HMMs), expert systems, clustering algorithms, quadratic classifiers, and artificial neural networks. Our system uses the Gaussian mixture modelling (GMM) (i.e. a single state HMM) approach. This system operates in 2 phases: training and recognition. During the training phase, the system takes the speech utterances for a single language and converts them into feature vectors. A GMM is trained on the feature vectors for each language. During recognition, an unknown utterance is compared to each of the GMMs. The likelihood that the unknown utterance was spoken in the same language as the speech used to train each model is computed, and the most likely model is determined as the hypothesised language.

The GMM LID approach performs classifications using information from single observations while an LID system using HMMs has the ability to model sequential events of speech (Zissman, 1993). However, Zissman has reported that the performance of GMMs was comparable to that of HMMs and this is one of the reasons that we utilise GMMs. Note that with post-processing the performance of a phonetically based HMM system can be improved. The main reason that motivated us to utilise GMMs with Universal Background Modelling (UBM) was that this technique was successfully applied to speaker verification in a highly computationlly efficient manner (Reynolds, 1997). This paper begins with an overview of our basic GMM LID system. The approach for creating models by adapting the model from a universal, language-independent background model (UBM) is then described, followed by an approach that speeds up classifications during the recognition phase. Finally the results of the experiments are presented.

2. LANGUAGE IDENTIFICATION SYSTEM

2.1 Parameterisation

The feature vectors used for modelling languages comprised of 12 Mel-Frequency Cepstral Coefficients (MFCCs) (Rabiner et al., 1993) derived from 20 filterbanks. Each feature vector is extracted at 10 ms intervals using a 32ms window of bandlimited (300-3400 Hz) speech. Since the experiment involved telephone speech, cepstral mean subtraction was applied to the MFCCs to reduce the linear channel effects. The corresponding delta coefficients were computed over a window length of 15 frames. Initially a shorter delta coefficient window length was trialled. Preliminary experiments indicated an improvement by extending this window length. A longer delta window length may be able to encapsulate more of the temporal information that is specific to language discrimination, particularly when the GMM does not use information across frames. Finally the delta coefficient of the frame energies (over the same window size) was appended to the features.

2.2 Gaussian Mixture Model (GMM) Classification

The GMM approach attempts to model the probability density function of a feature vector, $\vec{x}$, by the weighted combination of multi-variate Gaussian densities:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (1)$$

with

$$b_i = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_i)} \qquad (2)$$

where λ is the model described by

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \qquad (3)$$

In equation 1, i is the mixture index ($1 \leq i \leq M$), p_i is the mixture weight such that $\sum_{i=1}^{M} p_i = 1$, and $b_i(\vec{x})$ is a multi-variate Gaussian distribution defined by the corresponding means $\vec{\mu}_i$ and diagonal covariance matrices, Σ_i.

The estimation of the GMM parameters is accomplished by an iterative process, termed the Expectation-Maximisation (EM) (Reynolds et al., 1995). For more rapid GMM convergence, the mixture means, weights and variances are seeded by statistics determined by a K-means (Schalkoff, 1989) vector quantisation estimate of the feature vectors (Pelecanos et al., 2000).

During recognition, an unknown speech utterance, X, comprising of observations $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T$, is classified by first calculating the average log likelihood that the language model produced the unknown speech utterance. This is given as

$$p(X \mid \lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda) \qquad (4)$$

where λ is the model for the corresponding language. The maximum-likelihood classifier hypothesis, H can be calculated as

$$H = \arg \max_{l=1}^{L} p(X \mid \lambda_l) \qquad (5)$$

where the language index l = 1, 2, ..., L for L languages. Figure 1 shows the block diagram of the two phases of the LID system.