AUTOMATIC SPOKEN LANGUAGE IDENTIFICATION BASED ON ANN USING FUNDAMENTAL FREQUENCY AND RELATIVE CHANGES IN SPECTRUM

Javad Sheikhzadegan, Mahmood Reza Roohani
Research Center of Intelligent Signal Processing (RCISP)
7th Golestan, Pasdaran Avenue, Tehran, Iran
Fax: +98 – 21 – 2548055
mr_rohani@hotmail.com

ABSTRACT: An automatic spoken language identification system based on artificial neural network (ANN) is described in this paper. Two different sets of statistical parameter namely prosodic and segmental features, extracted from fundamental frequency (F0) contour and frequency spectrum, were used for language classification. The procedures of feature parameter extraction are: 1) F0 contour extraction, 2) approximation of polygonal line of F0 contour, 3) determination of frequency spectrum, 4) calculation of energy relative changes of distinctive frequency bands, 5) extraction of statistical parameters of F0 contour (prosodic parameters) and relative bands energy (segmental parameters). F0 contour was extracted by one innovated version of cepstrum pitch extractor method, and frequency spectrum is determined by short time fast fourier transform. A multi-layer perceptron (MLP) type of neural network used for classification purpose. Training and testing process was performed using a multi-language speech database generated at RCISP. Identifying task correction rate for six languages is greater than 97% in closed experiment tests and about 75% in open experiment tests.

1 - INTRODUCTION

Speech translation by machine is a desirable ambitious goal for future speech communication society. Speech recogniser is one of the main units of speech machine translator. Today's, speech recognisers are designed to accept a single language, but for future, it will be necessary for speech recognisers to accept a variety of languages. The role of spoken language identification is very important for multi-language speech understanding as a pre-processing unit [1].

Either segmental or prosodic features of speech are known as the main sources of spoken language identification. Phoneme inventory was used for language identification in some earlier works and also a few works have been done using fundamental frequency contour characteristics as speech prosodic features. In the phoneme inventory methodology, the difference of language specific phonemes and the language specific co-occurrence of phonemes (phonotactics) can be used for language identification. Efficient utilization of language specific phonemes and also phonotactic constraints heavily depends on the accuracy of phoneme recognition, while acceptable phoneme recognition is not easy to be achieved. On the other hand, although the extraction of fundamental frequency contour with an acceptable accuracy is not a difficult task, it can provide relatively effective and robust features for language identification. We have realized an automatic language identification system based on speech F0 contour and frequency band energy relative changes [2].

The paper organization is as follows: In section 2 identification method is described briefly. In section 3, we presents details of our experiments and results, that contains training and test corpus, extraction of feature parameters, training and testing process and tables that show the results of experiments. In the final section (section 4) we will have the conclusion of the paper.

2 - IDENTIFICATION METHOD

In the first step for each utterance, speech signal has been segmented to continued pitch portions and then in each portion piecewise line approximation conduct spectral parameter extraction used frame by

frame. In the second step statistical features extracted for all of them in utterance have long lasted more than 20 seconds then this 30 features has been introduced to the MLP classifier that contained 60 and 30 neurons in two hidden layers. In each process values of the output layer are compared with a preset threshold, otherwise, if they are less, we will enter more than one set of features and it will be extracted and enter to the MLP inputs and finally the maximized output for further times is winner. The MLP has been trained in such a manner that comply best performance in identification task. Structure of identifier has been shown in figure 1.

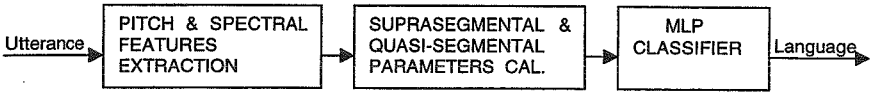

Figure 1. Method Realization

3 – EXPERIMENTS AND RESULTS

3.1 – Training And Test Corpus

The corpus for the work has been gathered from international satellite TV programs that can be received in Croatia and Iran. Number of Languages in this corpus is six, Farsi, Arabic, English, French, German and Russian. Sampling frequency is 22.05 kHz and sample resolution 16 bits. Signal has recorded from 44 speakers for each language and length of each utterance is 40 seconds. Style of speaking is spontaneous. 20 seconds of recorded data from each speaker for system training has been used and rest for closed experiment test.

3.2 – Feature Parameters

We used 21 parameters as introduced in reference [2], and 9 parameters introduced in the later subsection. These parameters are extracted by applying filter and decimation of signal to 11.025 kHz sampling frequency. Processing frame time is about 44 ms long and frame step forward time is about 11 ms long.

3.2.1 – Fundamental Frequency Extraction

For F0 contour extraction we used one modified algorithm of Cepstrum Pitch Extractor (CPE), it was developed by us some years ago. In the standard CPE algorithm, one fixed threshold level has been used for determining the pitch period in the quefrency domain. Fixed thresholds, causes relatively high error in the F0 extraction accuracy, because there are some abnormal vacillations in the cepstrum curves of the voiced speech. For overcoming to this abnormalities and improving the performance of the CPE, we have modified the threshold and used it in companion with some simple rules. In this modification, instead of one fixed threshold, two adaptive thresholds were determined for every process frame. We named this modified algorithm as Adaptive Double Threshold CPE (ADTCPE) [3].

3.2.2 – Frequency Bands Parameters

These parameters are reflection of quasi-segmental features of speech. Base of them are relations:

R1=(energy of band 800-1200Hz)/(energy of band 400-800Hz)

R2=(energy of band 1200-2000Hz)/(energy of band 400-800Hz)

Parameters that used from them are:

STD(R1), STD(R2)

SKEW(R1), SKEW(R2)

KURT(R1), KURT(R2)

CORR(R1,R2)

MEAN(R1), MEAN(R2)

By using these 9 parameters, correction rate has been improved by 4%. Frequency bands about 1000Hz have the most important information that contained in utterance. In R1 and R2 computation we normalized the energy of bands to the energy of a band that has energy in all the times. Although poor identification rate reported for amplitude based features, using of them with other parameters is in order of importance [4]. Similarly we used these parameters.

3.3 – Training and Testing Process

We have trained MLP by parameters of 20 native speakers of each language. To prevent from local minima first it was trained with well-matched data and continued training by the others. Training has been used first 20 seconds of each person and the rest for matching. Also closed experiment on data has been developed for each person of training set on the 20 seconds in progress. We presented results of open experiment on parameters of 24 remaining speakers as test results in confusion table 2.

3.4 – Results

Table 1 and 2 show the results on confusion matrix among 6 languages. Test results show that the lowest accuracy is for Russian language we can explain in another way recording of this language signal is from analogue receiver output of device saturated and very distorted. For Arabic we used dialect from eastern and western Arab countries and as presented in reference [5] prosody of these dialects is in difference. Furthermore Farsi and Arabic languages are very similar and this affected classification accuracy. French identification accuracy affected by background music on broadcasting.

Language	Farsi	Arabic	English	French	German	Russian
Farsi	97.5	0.0	2.5	0.0	0.0	0.0
Arabic	0.0	100	0.0	0.0	0.0	0.0
English	0.0	0.0	100	0.0	0.0	0.0
French	0.0	0.0	0.0	97.5	2.5	0.0
German	0.0	0.0	0.0	2.5	97.5	0.0
Russian	2.5	2.5	0.0	0.0	0.0	95

Table 1 – Confusion for training data

Language	Farsi	Arabic	English	French	German	Russian
Farsi	69	15	0.0	3.0	10	3.0
Arabic	8.0	55	10	7.0	10	10
English	0.0	3.0	79	7.0	7.0	4.0
French	3.0	3.0	0.0	72	12	10
German	0.0	0.0	0.0	1.0	96	3.0
Russian	3.0	3.0	7.0	17	20	50

Table 2 – Confusion for test data

4 – CONCLUSION

The results reported here show that new 9 parameters have improved accuracy but we can not compare our work with the others because of inaccessibility of OGI-MLTSC corpus. We want to have one copy of it in future. The MLP can train very good but in test degraded, this may be as a result of training data and its generality.

REFERENCES

[1] Y. K. Muthusamy and Ronald A. Cole, "Automatic Segmentation and Identification of Ten Languages Using Telephone Speech", Proc. ICSLP92, Banff, pp. 1007-1018 (1992)

[2] S. Itahashi and L. Du, "Language Identification Based on Speech Fundamental Frequency", Proc. Eurospeech95, Madrid, pp. 1359-1362 (1995).

[3] J. Sheikhzadegan and M.R. Roohani, "Automatic pitch frequency extraction with high accuracy and using it for group speaker classification", Proceeding Third Electronics Conference, TEC – 95, Shiraz University, Oct. 1995 (in Persian).

[4] F. Cummins, F. Gers, J. Schmidhuber, "Automatic discrimination among languages based on prosody",IDSIA Technical Report IDSIA-03-99

[5] M. Barkat, J. Ohala annd F. Pellegrind, "Prosody as a distinctive feature for the discrimination of Arabic dialects". Proc. Eurospeech99, Budapest (1999).

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.