A Feature Selection based Approach for Speaker and Language Identification in VoIP Networks

Synopsis of the Thesis to be submitted in Partial Fulfillment of the Requirements for the Award of the Degree of

Master of Technology

in

Computer and Information Technology

by

Major D J J Ranade

Under the supervision of

Prof. Indranil Sengupta

and

Dr. S K Ghosh

Department of Computer Science and Engineering

Indian Institute of Technology, Kharagpur

May 2008
# TABLE OF CONTENTS

1. Introduction .......................................................... 1-2
2. Literature Survey ..................................................... 2-3
3. Motivation and Objectives ........................................... 3
4. Work done ............................................................. 4-10
   4.1 Speaker Identification ............................................ 4
   4.1.1 Coding of Speech and Discriminative Properties of Codec Parameters 4
   4.1.2 Speaker Identification based on Direct Codec Parameters and Extraction of Statistical Features 5-6
   4.1.3 Significant Features Selection .................................. 6-7
   4.1.4 Experimental Results ........................................... 8
   4.2 Language Identification ........................................... 8-9
   4.3 Proposed Architecture of Speaker & Language Identification Framework 9-10
5. Conclusion ............................................................. 10
References ............................................................... 11
1. **Introduction**

An online speaker and language identification system can add a new dimension to the modern VoIP based telecommunication networks. There are scenarios like routing a distant voice call to a particular personal relationship manager in a call centre based on linguistic and speech characteristics, homing to a rogue based on his speech signatures, detecting a friend or foe on a tele-network, allotting resources and value added services based on subscriber’s voice signatures etc., which can be efficiently handled by an efficient speaker and language identification system with a near real-time decision making capability. However, the real-time speaker or language identification is not practically achievable using a typical *Automatic Speaker Recognition (ASR)*[1] or *Automatic Language Identification (LID)*[2] systems because ideally, they are designed to accept input speech waveforms without any signal transformation, whereas in case of VoIP the speech is generally transmitted in the compressed format. Hence, if ASR/LID system is to be employed for speaker/language identification over VoIP networks the input compressed speech is required to be re-synthesized to its original wave-form for onward processing through the system, which definitely requires considerable time. Hence a different approach is need to evolved which is capable of performing multiple identification in near real-time for VoIP networks.

We are presenting a novel approach based on *Dimensionally Reduced Statistically Significant (DRSS)* features extracted directly from the compressed speech representing the aggregate acoustic behavior associated with an individual speaker or a language. The DRSS features are extracted by forming the statistical feature vectors by calculating few statistic from the compressed speech and subsequently optimal statistical features are selected using ‘new correlation’ based criteria of ‘maximization of significant difference and independence’(MSDI). The independence is measured based on Bergsma’s ‘new correlation’[19], which is capable of measuring arbitrary dependencies between the random variables. The DRSS features are applied to Artificial Neural Network – Multi Layer Perceptron (ANN-MLP)[12] for subsequent classification job.

In our work major trust is on the speaker identification problem. However an effort has been made in the area of language identification to augment the speaker identification. Other than the possible scenario quoted above, one may consider that language identification can also be used to categorize a large speaker set into various linguistic domains which will arguably enhance the efficiency of the speaker identification approach. We considered four Indian languages (Hindi, Bengali, Tamil and English) for our work on language identification. A considerable reduction in the data required for the training and the testing of the classifier, higher speech for classification(in µsec) and the accuracy (>99%) obtained during our work may usher a real-time speaker and language identification in the live network environment.

The rest of the synopsis is organized as follows: The study of existing ASR/LID systems, advances in reorganization systems in compressed domain, speech coding mechanism for VoIP applications and the idea behind DRSS feature selection are discussed in brief in section 2. A brief motivation for the selection
of the problem and the objectives defined for our work are given at section 3. The work done in the nutshell on speech codec, statistical feature selection, DRSS feature selection using MSDI criteria and the experimental results of speaker & language identification are presented in section 4. In section 5, we have presented few issues which can be addressed in future along with the conclusion.

2. Literature Survey

Automatic Speaker Identification has been a major research topic in the area of audio processing since the 1970s. We studied in [1][10] that the modern ASR systems are build based on series of digital signal processing operations and finally applying the yield of these operations to nonlinear classifier. The signal processing operations includes the pre-emphasis of the signal to boost the amount of energy in the high frequencies, framing and windowing(hamming) of the signal in order to remove the spectral discontinued at the ends due to the quasi-stationary of the speech signal. The signal is shifted to frequency domain by \textit{dft} to find that how much energy it contains at different frequency bands. Thereafter the human hearing response is simulated using a Mel-Filterbank and finally we take \textit{log} and \textit{idft} of the signal to get the final \textit{cepstrum} coefficients which are used to train and test a classifier.

A typical ASR system, which needs to decompress the compressed voice packets into a voice signal waveform, does not scale in terms of CPU, disk access, and memory for live VoIP applications where speech is transmitted in the coded format. To make the real-time speaker identification feasible in the compressed domain in the recent past a few compressed domain ASR techniques have been proposed, which are ASR using direct codec parameters[3](2000), micro-clustering based[4](2005) and low complexity ASR using statistical features[5](2005). We implemented [3] and [5] and further established that there is room for further improvements if we use DRSS features instead using direct codec parameters as employed in the referred techniques.

In our work we kept the focus on most the most widely used speech codec for VoIP applications, which is the Conjugate Structure - Algebraic Code Excited Linear Prediction (CS-ACELP)[7] algorithm based ITU G.729[8] codec. The basic model used for G.729 codec is Code Excited Linear Prediction (CELP)[6]. We studied and used G.729 Annx A[8] & B[9] for all our experiments on compressed speech presented in this report. The G.729 coder operates on speech frames of 10ms corresponding to 80 samples at a sampling rate of 8000 samples per second. For every 10ms frame, the speech signal is analyzed to extract the parameters of the CELP model (linear-prediction filter coefficients, adaptive and fixed-codebook indices and gains), these parameters are encoded and transmitted on the wire.

We started with all the fifteen codec parameters (shown in Table I) and selected top eleven parameters after

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter description</th>
<th>No. of Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>L0</td>
<td>Predictive index of LSP quantizer</td>
<td>1</td>
</tr>
<tr>
<td>L1</td>
<td>First stage vector of LSP quantizer</td>
<td>7</td>
</tr>
<tr>
<td>L2</td>
<td>Second stage, lower vector of quantizer</td>
<td>5</td>
</tr>
<tr>
<td>L3</td>
<td>Second stage, higher vector of quantizer</td>
<td>5</td>
</tr>
<tr>
<td>PI</td>
<td>Pitch delay</td>
<td>3</td>
</tr>
<tr>
<td>P9</td>
<td>Parity # for pitch delay</td>
<td>1</td>
</tr>
<tr>
<td>X1</td>
<td>Signs of fixed-codebook pulses</td>
<td>4</td>
</tr>
<tr>
<td>C1</td>
<td>Pulse positions of fixed codebook</td>
<td>33</td>
</tr>
<tr>
<td>GA1</td>
<td>Gain codebook (Stage I)</td>
<td>5</td>
</tr>
<tr>
<td>GA2</td>
<td>Gain codebook (Stage II)</td>
<td>4</td>
</tr>
<tr>
<td>PI</td>
<td>Pitch delay</td>
<td>5</td>
</tr>
<tr>
<td>S2</td>
<td>Signs of fixed-codebook pulses</td>
<td>4</td>
</tr>
<tr>
<td>C2</td>
<td>Pulse positions of fixed codebook</td>
<td>15</td>
</tr>
<tr>
<td>GA3</td>
<td>Gain codebook (Stage II)</td>
<td>3</td>
</tr>
<tr>
<td>OR2</td>
<td>Gain codebook (Stage II)</td>
<td>4</td>
</tr>
</tbody>
</table>

2. Literature Survey

Automatic Speaker Identification has been a major research topic in the area of audio processing since the 1970s. We studied in [1][10] that the modern ASR systems are build based on series of digital signal processing operations and finally applying the yield of these operations to nonlinear classifier. The signal processing operations includes the pre-emphasis of the signal to boost the amount of energy in the high frequencies, framing and windowing(hamming) of the signal in order to remove the spectral discontinued at the ends due to the quasi-stationary of the speech signal. The signal is shifted to frequency domain by \textit{dft} to find that how much energy it contains at different frequency bands. Thereafter the human hearing response is simulated using a Mel-Filterbank and finally we take \textit{log} and \textit{idft} of the signal to get the final \textit{cepstrum} coefficients which are used to train and test a classifier.

A typical ASR system, which needs to decompress the compressed voice packets into a voice signal waveform, does not scale in terms of CPU, disk access, and memory for live VoIP applications where speech is transmitted in the coded format. To make the real-time speaker identification feasible in the compressed domain in the recent past a few compressed domain ASR techniques have been proposed, which are ASR using direct codec parameters[3](2000), micro-clustering based[4](2005) and low complexity ASR using statistical features[5](2005). We implemented [3] and [5] and further established that there is room for further improvements if we use DRSS features instead using direct codec parameters as employed in the referred techniques.

In our work we kept the focus on most the most widely used speech codec for VoIP applications, which is the Conjugate Structure - Algebraic Code Excited Linear Prediction (CS-ACELP)[7] algorithm based ITU G.729[8] codec. The basic model used for G.729 codec is Code Excited Linear Prediction (CELP)[6]. We studied and used G.729 Annx A[8] & B[9] for all our experiments on compressed speech presented in this report. The G.729 coder operates on speech frames of 10ms corresponding to 80 samples at a sampling rate of 8000 samples per second. For every 10ms frame, the speech signal is analyzed to extract the parameters of the CELP model (linear-prediction filter coefficients, adaptive and fixed-codebook indices and gains), these parameters are encoded and transmitted on the wire.

We started with all the fifteen codec parameters (shown in Table I) and selected top eleven parameters after
evaluating their discriminatory power ranking. Thereafter to reduce the dimension of training and the test data few statistic (Coefficient of Covariance, Skewness, Kurtosis and Lag-3 Autocorrelation) were extracted from the data. To further reduce the dimension of the data optimal statistical features (DRSS) are selected using ‘new correlation’ based criteria of MSDI. The significant difference is based on $F$-statistic of ANOVA and for measuring independence we evaluated simple correlation[18], mutual information[21] and Bergsma’s ‘new correlation’[19]. We established that for our system computationally, Bergsma’s ‘new correlation’[19] is the better choice than mutual information, whereas simple correlation has it’s own inherent disadvantages[15][16] as a measure of independence.

We also studied that similar to speaker identification, the Automatic Language Identification (LID) systems also involves considerable signal processing overheads[2] and accepts the coded speech only in the un-compressed format. Hence they are unsuitable for real-time VoIP applications. In the literature, identification of the languages using coded parameters was first mentioned in [3](2000). In our work, an approach similar to speaker identification was adopted for language identification. However we found that a different set of codec parameter with a different discriminatory ranking than speaker identification is obtained for language identification dataset.

3. Motivation and Objectives

We had seen during the literature survey that in VoIP networks, generally the speech is transmitted in the compressed format using some speech compression algorithm, whereas a typical ASR/LID System is not capable of handling compressed speech. The computational overhead incurred in re-synthesizing the speech to original waveform and subsequent signal processing operations makes it impractical for live VoIP applications. It is established that direct usage of codec parameters can overcome the need of resynthesizing of the coded speech to the original waveform. However there is a need for much faster and efficient speaker & language identification approach, so that the speaker & language identification be made practically achievable in the live network environment. With this motivation the following objectives were formulated for our work:-

- Study the different aspects of the speaker identification problem in compressed domain.
- Evolve an efficient speaker identification approach for VoIP networks.
- Explore the language identification problem in the compressed domain to argument the speaker identification problem.

We were able to meet the above objectives by evolving DRSS based feature selection approach presented in this report. The accuracy(>99%) and speed (in µsec) in speaker and language identification is yielded during our experiments.
4. Work done

In our work a major thrust has been upon the speaker identification problem. The language identification can be helpful in the categorization of speakers into various linguistic domains, which may be helpful in reducing the classification overheads of speaker identification problem for a large speaker set or may be employed as a separate tool in various futuristic VoIP based applications.

4.1 Speaker Identification

The overall speaker identification approach involves the extraction of DRSS features from G.729 coded speech and subsequent training and testing of the classifier. The work done in brief has been presented in the following sub-sections:-

4.1.1 Coding of Speech and Discriminative Properties of Codec Parameters

The speaker identification process is realized by capturing the G.729B coded speech of the selected speaker set of fourteen arbitrary speakers. We analyzed all fifteen parameters(Table I), which are sufficient to describe the characteristics of the speaker and speech. To analyze the nature of the G.729 coder parameters, we plotted the probability density plots considering a particular parameter for different users. The shapes of these plots confirm that a particular codec parameter (say P1) has discriminative property for the different users, because the density plots of two different speakers are statistically different for the same parameter (Fig 1). Whereas, when the density plots of a parameter captured at different time instances for the same speaker were plotted, statistically similar plots were obtained (Fig 2). We evaluated the density plots of all fifteen parameters and selected nine coded parameters ($L_1, L_2, L_3, P_1, P_2, GA_1, GA_2, GB_1, GB_2$) which are sufficient to discriminate between various spoken languages. To conclude this we used a heuristic that “accuracy increases for the addition or deletion of a parameter to a selected parameter set”.

Fig. 1. Density Plots of parameter P1 for two different speakers

Fig. 2. Density Plots of parameter P1 for two different speakers with multiple speech samples taken at different time instants
4.1.2. Speaker Identification based on Direct Codec Parameters and Extraction of Statistical Features

In the previous section it has been seen that the coded parameters have discriminative properties, which can be used to distinguish between various speakers. Intuitively, first of all we considered the direct use of the codec parameters for the speaker identification task. We selected a target set of 10 male and 4 female speakers and captured their G.729B coded speech packets equivalent to 90sec of the speech for each speaker. On an average G.729A coder produces about 9000 packets for 90sec of speech. Where as G.729B coder removes the silence periods within the speech hence reduces the number of packets by percentage of silence presence in the speech. For each user a set of 9 dimensional vectors \(<L1, L2, L3, P1, P2, GA1, GA2, GB1, GB2>\) is generated for training (say, a matrix of dimension 9x9000). These training vectors were directly used to train ANN-MLP. Similarly for speaker identification, for a speaker coded packet corresponding to 90sec of individual speech is nearly 9x126000 which is too large, whereas training with speech sample lesser than 90sec may not be represents the speaker’s acoustic characteristics in totality. Hence instead if using the coded speech data directly we decided to extract few statistic from the captured speech data, thereby reducing the size of the training data horizontally. A window of 5sec of the speech was taken for a particular feature and the corresponding statistic were calculated. We considered statistic \(COV\) (Coefficient of Variance), \(Skewness\) and \(Kurtosis\) [18] for our analysis. A 27x1 matrix was computed from a 9x500 matrix by calculating the \(COV\), Skewness and Kurtosis for each of the parameter considering a window of 500 (5 Sec of Speech). Hence for 90sec of training speech of a particular speaker a training matrix of dimension 27x18 (for G.729A) is obtained. For a set of 14 speakers there will be 14 such training matrix hence overall dimension for the training data will be 27x252 in case of G.729A codec. It is evident that the feature set has increased from initial nine to twenty seven whereas the training data has been reduced considerably in comparison with the earlier described technique where the parameters were used directly for training; in that case the size of similar training matrix would have been of the dimension 9x126000. The process of extraction of statistical features data is depicted in Fig. 3.
Using these statistical features a considerable improvement in speaker identification accuracy and reduction in training and testing time of ANN-MLP were observed. For a set of 14 speakers a confusion matrix shown as Table II was obtained by capturing 90 sec compressed speech for training and 30 sec of compressed speech for testing of ANN. The ANN-MLP was trained using training matrix prepared by extracting data corresponding to statistical features from 90 sec of captured compressed speech for each of the fourteen speakers. Here five test vectors were prepared by extracting statistical features data from the sample test speech (30 sec) of the speaker to be classified. These vectors are applied to already trained ANN-MLP classifier. We used five test vectors to minimize the possibility of any wrong decision as all these vectors will be tested on already trained classifier and the speaker with maximum score will be declared as identified.

The confusion matrix shows that there was only one misjudgment for a set of 14 speakers, which was for speaker S9. Hence, the accuracy in speaker identification using statistical features is nearly 92%. The average times for training and testing on Pentium-IV PC simulating MLP with 54 hidden neurons and 1000 epochs were 16.59sec and 648µsec respectively.

### 4.1.3. Significant Features Selection

The significant features were selected using a ‘new correlation’ based criteria of ‘maximization of significant difference and independence’ (MSDI). We have used features selection procedure based on STSF[23] and a variant of similar approach presented in mRMR[24]. In our work, we introduced ‘new correlation’ as a measure of independence for feature selection. The new correlation is computationally efficient and capable of measuring arbitrary (linear/non-linear) dependencies between the random variables. However, before conducting any further statistical analysis we wanted to ascertain the nature of the feature data. The Bera-Jarque[13] & Shapiro-Wilk[14] parametric hypothesis tests of composite normality confirmed that the features has normal distribution with unspecified mean and variance for an acceptable significance level.

After arriving at a conclusion about the distribution of the features; which are normally distributed, we formulated a feature selection procedure based on MSDI. The ‘significant difference’ among classes is measured using ANOVA (Analysis of Variance)[17] F-test. It measures the speaker recognition ability of individual features. For measuring ‘independence’ (i.e. the non-similarity among the features) we considered new correlation coefficient[19]. The basic idea for measuring independence is to eliminate those features which can be constructed by arbitrary combinations of other features and significant

### TABLE II

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
<th>S13</th>
<th>S14</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S11</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
difference is measured to eliminate those features which have lower classification capability. Overall those feature will be selected which maximize the significance difference and independence.

The most commonly used measure for dependence between two variables is the mutual information [20]. As we know that true estimation of the mutual information between random variables practically is a hard problem due to approximation of joint and marginal probabilities used in the calculation of mutual information. Whereas, The simple correlation[18] between two random variables does not account for the nonlinear dependencies[15][22]. Moreover dependency between two random variables implies correlation whereas converse is not true[16]. To overcome these fallacies a new correlation coefficient[21] is proposed. It caters for the non-linear dependencies between random variables and has computational complexity of the order $O(n^2)$, where n is number of values for random variables.

The feature significance of a candidate feature is the contribution towards the improvement of the speaker identification when the candidate feature is added in a feature set. The measurement of significance of a feature $sf$ can be estimated by the product of the significant difference $sd$ among classes and the independence $L$ between the candidate feature and the already selected features.

$$sf_v = sd_v \times L_{u1u2...uv}$$

here $v$ represent the candidate feature; $u1, u2,..., up$ represent the already selected features.

The procedure for feature selection following the criteria of MSDI is explained using the flowchart shown in Fig. 4.

![Flowchart for Feature Selection](image)

The feature vectors shown in Fig.5 were obtained after running the feature selection algorithm employing new correlation as the measure for independence for a selected speaker set. The ordering of features is dependent on the acoustic behavior of the speakers and their speech’s inter-relationship hence the ordering will be different if we add or remove a speaker to a speaker set. For a simulation on a Pentium-IV PC with 14 speaker data set the average time taken for online selection was approx. 4.86 sec.
4.1.4. Experimental Results

We selected first eighteen DRSS features by using the simple heuristic that “accuracy in identification increases by adding or deleting a particular DRSS feature in a selected feature set” out of initial 27 statistical features and formed a training matrix corresponding to 90 sec of compressed speech of all the speakers and thereafter the classifier (ANN-MLP) is trained. We captured the 30 sec of coded speech of an anonymous speaker on the network and extracted already known DRSS features. As explained earlier, five test vectors were generated to minimize the possibility of any wrong decision as these vectors are tested on already trained classifier and the speaker with maximum score is declared as identified. These training matrix and test vectors are used for training and testing of MLP-ANN. The confusion matrix shown as Table III was obtained, which inferences accuracy in speaker identification nearly 99% for 14 speaker dataset. For a simulation on a Pentium-IV PC, the average times for the training and identification are 15.83 sec and 593 μsec respectively. We summarized the simulation results obtained for various experiments described earlier in this paper for comparative analysis in Table IV.

<table>
<thead>
<tr>
<th>Table IV</th>
<th>Speaker Identification Simulation Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Direct Parameters</td>
<td>70%</td>
</tr>
<tr>
<td>Statistical Features</td>
<td>92%</td>
</tr>
<tr>
<td>DRSS Features</td>
<td>&gt;99%</td>
</tr>
</tbody>
</table>

4.2 Language Identification

We identified that nine codec parameters (L1, L2, L3, GA1, GA2, GB1, GB2, C1, C2) are sufficient to discriminate between various spoken languages based on the shape of density plots and the heuristics that “accuracy increases for the addition or deletion of a parameter to a selected parameter set” out of initial 15 codec parameters. Here we captured 180 sec of compressed speech for each speaker (8 speakers per language and total of 4 Indian languages) and formed a training matrix corresponding to 180 sec of compressed speech of all the speakers, we extracted statistic (COV, Skewness and lag-3 Autocorrelation) and we trained the ANN-MLP using statistical features. Further we tested the ANN-MLP using statistical features equivalent to 30 sec speech of an arbitrary speaker and obtained the confusion matrix as shown in Table V. As explained earlier, five test...
vectors were generated to minimize the possibility of any wrong decision as all these vectors are tested on already trained classifier and the language with maximum score is declared as identified. Thereafter same experiment was repeated using DRSS features. Here we captured the 30sec of coded speech of an anonymous speaker on the network and extracted already known DRSS features. The selected ranked DRSS features are as shown in Fig 6. We considered first 22 DRSS features for evaluation of the language identification system. The confusion matrix shown as Table VI was obtained, which inferences accuracy in language identification nearly 99% for 4 language dataset. For a simulation on a Pentium-IV PC, the average times for the training and identification are 45.80sec and 652µsec respectively.

We summarized the simulation results obtained for various experiments for language identification for comparative analysis in Table VII.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Accuracy</th>
<th>Training Time</th>
<th>Identification Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Features</td>
<td>&gt;99%</td>
<td>51.59sec</td>
<td>698µsec</td>
</tr>
<tr>
<td>DRSS Features</td>
<td>&gt;99%</td>
<td>45.83sec</td>
<td>652µsec</td>
</tr>
</tbody>
</table>

### 4.3 Proposed Architecture of Speaker & Language Identification Framework

We intend to propose, a formal speaker identification framework for VoIP networks based on experimental results and the feature selection based approach evolved, which is summarized in Fig. 7. To capture the live VoIP packets, a packet sniffer working in the promiscus mode at VoIP router/gateway can be employed. The compressed speech data (parameters) is extracted from the VoIP packets and the packet identity is established by reading the Source Address appended in the IP header of the packets. The subsequent process is divided in two phases, the background DRSS feature selection cum Training phase and on-line speaker identification phase. The training of the classifier is based on DRSS features and the class labels (i.e, identity tags for speakers). First of all the target speaker set is identified and for each speaker in the target set VoIP Packets corresponding to 90 sec speech are captured. Thereafter values corresponding to nine G.729 coder
parameter \((L_1, L_2, L_3, P_1, P_2, GA_1, GA_2, GB_1, GB_2)\) are extracted and stored in the vector format on a feature database. The vector set is dimensionally reduced (horizontally) by calculating the statistic \((COV, \text{Skewness and Kurtosis})\) for each parameter corresponding to 500 samples (5 sec of compressed speech) and statistical feature vectors are formed with the number of features grown from initial nine to twenty seven. The new statistical features give us the aggregate acoustic characteristics of a speaker. Thereafter from these statistical features, statistically significant \((DRSS)\) features are selected using new correlation based \(MSDI\) criteria as explained earlier. Now, the data corresponding to DRSS features is used for training of the classifier \((ANN)\). The trained classifier along with DRSS features will be stored on to a database for later references in re-training and testing phase.

Since the DRSS features are already known, for on-line speaker identification phase raw speech packets are sniffed corresponding to 30 sec of speech and the data is divided into five/six parts (depending on G.729 Annx A or B) of 5 sec each. The five test vectors are prepared by extracting DRSS features from the data. These vectors are applied to already trained classifier. We used five or six test vectors to minimize the possibility of any wrong decision as all these vectors are tested on already trained classifier and the speaker with maximum score is declared as identified. For language identification the background DRSS feature selection cum training phase is replaced by the off-line DRSS feature selection cum training phase and we extract the DRSS features for a fixed speaker set (language dataset). We further propose that language identification training using some standardized language dataset, which will surly improve the performance and accuracy of LID system.

5. Conclusion

We have shown in our work that the on-line speaker and language identification can be realized on VoIP Networks by extracting and statistically condensing the acoustic behavior of a speaker and the spoken language from the compressed speech. A significant improvement in speaker and language identification accuracy and reduction in time-space complexity were observed when the Dimensionally Reduced and Statistically Significant \((DRSS)\) features are used, instead of speech codec parameters obtained directly from the VoIP compressed speech packets for the training and testing of the classifier. We employed MLP as classifier for our experiments, the basic and much researched flavor of ANN. In future we further plan to conduct the similar experiments using more efficient classifiers such as other flavors of ANNs, Multiclass-Support Vector Machine etc. Also we plan to validate results presented here on a large speaker set and report the performance. In particular for language identification we have not used any standard language dataset instead we captured G.729 coded voice for an arbitrary duration, hence there is scope for improvements if the classifier is trained using DRSS features obtained using some standard language identification dataset. The simulation and experimental results presented here motivate us to work in future in the pursuit of a fast signal processor based implementation of on-line speaker and language identification framework for the live VoIP networks.
REFERENCES


[22] Qianren Xu et al., “A Significance Test for Feature Sub-set Selection on Image Recognition”, University of Waterloo, Waterloo, Ontario, Canada, 2004