2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.81-88.

Effective Preprocessing of Speech and Acoustic Features Extraction for Spoken Language Identification

Abhijeet Kumar, H. Hemani, N. Sakthivel and **S. Chaturvedi**

Computational Analysis Division, Bhabha Atomic Research Center, Visakhapatnam, India

***Abstract*— Language identification (LID) systems have become very popular and indispensible in multilingual speech processing applications where there is need of preprocessing of machine systems and preprocessing of human interface. The system predicts the best identified language given the speech utterance. The proposed LID system uses a gaussian mixture model (GMM) based LID which uses generatively trained language models on acoustic features of a particular language. Acoustic approach requires only the digitized speech utterance and their language labels which are less expensive computationally than the alternative approaches which also require phonetic transcription of speech. This paper investigates the different preprocessing techniques for noise removal, speech activity detection (SAD), speaker normalization and channel normalization. Also, the extraction procedure of cepstral features that captures the phonetic characteristics of signal is illustrated. We also give a comprehensive review of the current trends in feature extraction and compare the results of the same. Notably, Shifted delta cepstral (SDC), a quintessential feature for LID systems derived from Mel frequency cepstral features (MFCC) have been successfully tested with GMM based classifier. A comparative study between use of MFCC and SDC features in LID has been conducted and presented.**

***Keywords*— Speech activity detection; cepstral mean subtraction, Feature extraction, Vocal tract length normalization, Mel Frequency Cepstral Coefficient, Shifted delta cepstral, Gaussian mixture modeling, Language identification**

I. INTRODUCTION

The aim of LID system is to identify the spoken language in a given segment of speech. Research in designing a language recognition system is not new. There are wide ranges of multilingual speech processing applications such as multilingual speech recognition by routing the speech to speaker of corresponding language, spoken document retrieval, spoken language translation (speech to text) and multilingual voice-controlled travel-information retrieval system [1] [2]. The LID system is implemented in two phase. In first phase, a language model is trained based on some language cues. In second phase, the testing of speech segment is performed against the trained language models. In general two main approaches are being followed in mainstream research for building a LID system. The first approach known as acoustic phonetic method uses the acoustic features and unlabeled training data. The acoustic features are extracted and then a language model is trained typically using a classifier like Gaussian mixture model (GMM) [1][3][4]. People have investigated other classifier like SVM also but GMMs are used because of their high performance and efficient computations [5][6]. The second approach is phonotactic approach which uses unique set of lexical phonological rules that governs the combinations of different phonemes. This method uses phone n-gram models [1]. In acoustic phonetic approach, earlier the acoustic features extracted from speech for language identification were Perceptual linear predictive (PLP), Linear Predictive Cepstral Coefficients (LPCC), MFCC and their derivatives [4]. The accuracy achieved with these features in acoustic approach was less than accuracy achieved from phonotactic approach [2]. But with the advancements in preprocessing techniques and invent of shifted delta cepstral (SDC) features [7], Acoustic method achieves accuracy as compared to phonotactic method [1].

This paper presents an implementation of an LID system that uses acoustic phonetic approach and investigates the key issues related to preprocessing and extraction of acoustic features from a speech signal. To be specific, speech utterance carries information from various sources. The actual uttered speech signal is distorted due to convolution of signal with constant channel dependent factor and speaker dependent factor. It is necessary for LID to be robust to channel distortion. A normalized acoustic model which is independent of such distortions is required in order to identify a language. Normalization methods affect the signal such that it cancels out the unwanted distortions. Due to the distortions introduced by different channels like microphone, telephone, broadcast channel, radio etc; variety of adaptation algorithm are used like RASTA filtering [8] [9], Gaussian dynamic cepstral

978-1-4799-9855-5/15/$31.00 ©2015 IEEE

representation (GDCR) [8] for normalizing channel. The state-of-the-art recognizers use a simple but very effective algorithm: Cepstral mean subtraction (CMS) for channel normalization [8] [10]. As a part of this paper, we investigate the basic algorithm of CMS and reasons of being a successful channel normalization technique. Although CMS also eliminates the speaker-dependent part when a speaker based CMS is done, there are issues when it is used for conversational speech. Therefore, in order to make the signal speaker-independent, we explored a well-known technique called Vocal Tract Length Normalization (VTLN) [11] [12]. This involves frequency warping for each speaker and normalizes the effect of vocal tract excitation which is specific to each speaker. In section 2, the flow of feature extraction process from a speech signal is depicted. The role of CMS in LID experiments is presented in subsection of section 2. In section 3, algorithms of VTLN techniques have been investigated. MFCC and SDC feature extraction techniques are illustrated in section 4. A brief description about language model training based on Gaussian Mixture Models (GMM) is given in section 5. Finally experiments and results comparing the effects of these techniques are presented in section 6.

II. Feature Extraction

Feature extraction comprises of two parts: Preprocessing of signal and extraction of cepstral feature vectors from preprocessed digitized segmented speech. A complete flow of steps is show in the fig 1.

In amplitude scaling, the training or testing speech signal is scaled between -1 and +1as loudness of speech is not relevant for LID.

A. Spectral Noise Removal

The background noise is eliminated in spectral domain using a two pass effect. The first pass is done over noise profile which is taken from the speech signal. For each windowed sample of the sound, FFT is taken and then statistics are calculated for each frequency band. The maximum level achieved by at least (n) sampling windows in a sequence is taken; the value of (n) varies. During the noise removal phase, a gain control for each frequency band is set such that if the sound has exceeded the previously-determined threshold, the gain is set to 0; otherwise the gain is lowered to suppress the noise. Then frequency smoothing is applied so that a single frequency is never suppressed or boosted in isolation, and then time-smoothing is applied so that the gain for each frequency band moves slowly. The gain controls are applied to the complex FFT of the signal, and then the inverse FFT is applied, followed by a hanning window. This well known noise removal technique has been applied to open source tools like sox and audacity [13].

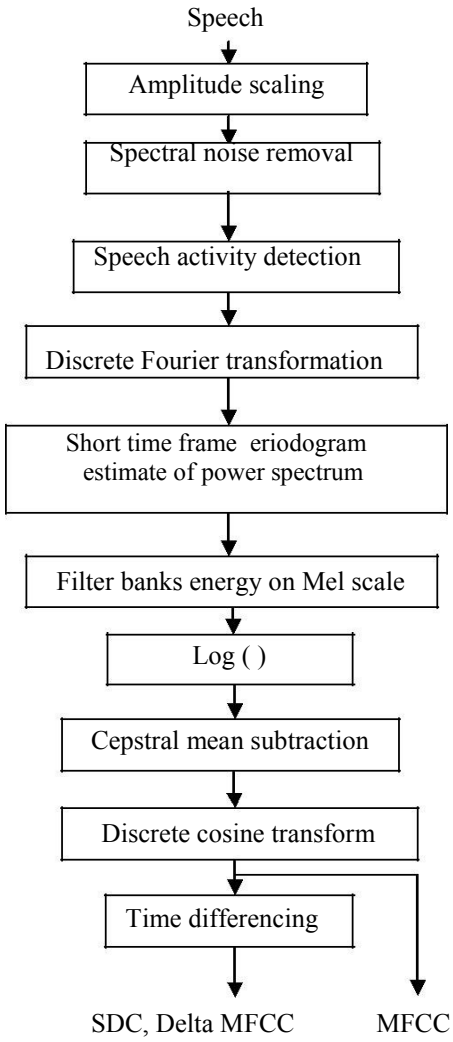

Fig. 1. A Complete Flow of Steps in Acoustic MFCC Feature Extraction From Digitized Speech Signal.

B. Energy Based Speech Activity Detection

SAD is done to distinguish silence, non-speech frames from speech signals. It has been found that the presence of non-speech frames considerably affects performance of the system. Energy based SAD are widely used in speech and speaker recognition application [14]. The state-of-the-art phoneme recognition technique [15] has also been used for SAD but due to its high computational cost, the method has been avoided here.

In our system, we have used energy based SAD which is straight forward. First, energy of all the speech frames is computed for a given speech utterance. An empirical

threshold can be selected from the frame energies. For instance, the threshold in reference [16] is selected as $0.06 \times E_{avg}$, where E_{avg} is the average energy of the frames in a speech utterance. In reference [17], the threshold is determined from the maximum energy of the speech frames. Both the methods have been investigated and applied.

C. Cepstral Mean Substraction

To perform CMS, the mean 'M' of the feature vectors is calculated and then 'M' is subtracted from each feature vector. Despite its simplicity it is very effective technique. CMS can be performed either in log-spectrum domain or in cepstrum domain because the cepstrum is a linear transformation of log spectrum. In time domain, signal 'y' is convolution of uttered speech signal 'x(n,k)', characteristic of a speaker dependent part v(n,k) and a constant channel noise c(n,k). So in DFT domain a channel model looks like [18]:

$$Y(n,k) = C(n,k).V(n,k).X(n,k) + N(n,k)$$

In the model, n is the frame index and k is the frequency bin. Also we consider an additive noise N (n, k). In log spectral or cepstral domain the channel model can be seen for speech and non speech frames.

Speech frames:

$$\log Y(n, k) = \log (C(n, k).V(n, k). X(n, k) + N(n, k))$$

$$\log Y = \log (C.V.X+N) = \log (C.V.X) + \log (1+N/C.V.X)$$

This can be expressed by notations in log domain:

$$y = c + v + x + r \text{ where } r = \log (1+N/V.C.X)$$

In case of dominant speech, $N << X.V.C$

$$y = c + v + x$$

Non speech frames (pause): x, v = 0 so, only additive noise 'n' remains

$$\log Y = \log N(n,k) \Rightarrow y = n$$

Now the mean of the received signal 'y' of the same model can be expressed in two terms weighted by proportion [18] of speech and non speech (pause) frames

$$m = \alpha * y_{speech} + \beta * y_{pause}$$

$$m = \alpha * (x_{avg} + v + c + r_{avg}) + \beta * n$$

for long speech utterances with high signal to noise ratio. we can neglect x_{avg} and r_{avg} and get the following approximation :

$$m = \alpha (v + c) + \beta * n$$

Now, after doing the mean subtraction (CMS)

$$z = y - m$$

speech: $y = x + v + c - \alpha * (v + c) - \beta * n$

$$= x + v + c - v - c + \beta v + \beta c - \beta n$$

$$= x + \beta v - \beta c$$

non speech: $y = n - \alpha v - \alpha c - n + \alpha n$

$$= \alpha n - \alpha v - \alpha c$$

As most of the non-speech frames are eliminated at SAD step, the value of $\beta << \alpha$ and in speech frame only 'x' remains as the significant quantity as expected. It can be also seen that in speech frames constant channel component is cancelled out after CMS. Although in non-speech frames there is a shift related to channel. In conversational speech there is greater variance in proportion of β. It has been found on switchboard corpus that relative word error rate is 4% less in speaker based CMS than conversational speech based CMS [18].

Earlier a channel normalization technique RASTA was used but it has been seen that CMS significantly outperforms RASTA [8]. In time domain, shape of speech signal is distorted after RASTA filtering whereas the same is preserved in CMS. A phase distortion is introduced by the RASTA filter. A detailed review of such aspects is investigated in [8] [9].

III. VOCAL TRACT LENGTH NORMALIZATION

The MFCC feature vectors which come as a result of feature extraction process represent the envelope of short time power spectrum which in turn is a manifestation of shape of vocal tract of a person. As the vocal tract of each person is different, MFCC feature vectors are speaker dependent and can be a good discriminator for speech recognition. But as far as LID systems are concerned, the feature vector must be speaker independent. It is known that the length of the human vocal tract has an inverse relationship to each formant frequency resulting in formant center frequency variation as much as 25% among speakers [11] [12].

Therefore, speaker normalization may be performed by rescaling the frequency axis according to the length of the vocal tract or a corresponding normalization factor in order to build a single gender independent model. In the present work, we trained a generic voiced speech model to select warp scale (normalization) factor for each speaker. For determining a generic voiced speech model, a set of acoustic data was used. This data was balanced with respect to speakers and languages. The acoustic data used is trained in an iterative process to get a final generic speech voiced model. The VTLN process of training and testing is explained as follows:

A. Training

1. Train UBM which is single probability distribution consisting of an appropriate number of multivariate Gaussians with balanced acoustic data from all languages. In our case we take several hours of training speech of a large number of male speakers (21 speakers from each of four languages).
2. Select α for each speaker, over the range 0.88-1.12 such that it maximizes the likelihood score against UBM when corresponding speaker data is warped with α. It guarantees that the warped version will give better likelihood scoring against the trained models.
3. Re-train the UBM with the warped acoustic data.
4. Repeat step 2 and 3 until there are no further significant changes in α of speakers.

B. Testing

1. Select the warping factor α for the input test speech by testing it against the generic converged UBM model trained in iterative process
2. Warp the frequency scale for the test speech with α and then test sample is tested against the trained independent language models.
3. Iterate until convergence.

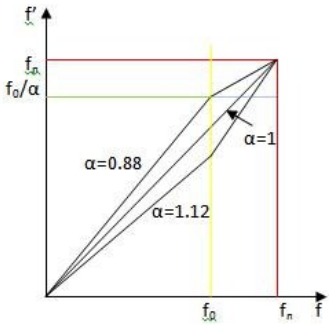

Fig. 2, Piecewise Linear Transforms of the Frequency Axis to Achieve Speaker Normalization.

This frequency warping/normalization/scaling process can be achieved using a piecewise linear transformation [11] of the frequency axis as depicted at Figure 2.

The warping function is as:

$$f' = \begin{cases} \alpha^{-1} f, & \text{if } f < f_0 \\ b * f + c, & \text{if } f > f_0 \end{cases}$$

Where f ‟ is the normalized/warped frequency, α is the speaker-specific normalization factor, f_0 is a fixed frequency (0.875 * Nyquist frequency) [19] set to handle the bandwidth mismatching problem. From the equation of line b and c can be calculated as „b‟ is slope of line and c is constant when $f > f_0$

$b = (y_2-y_1/x_2-x_1)$

$= (f_n‟ – f_0‟) / (f_n-f_o)$

$= (f_n-f_0/ \alpha) / (f_n-f_o)$

$= (\alpha f_n-f_o) / \alpha (f_n-f_o)$

Here, $f_n‟ = f_n$ (needs to be same to avoid bandwidth mismatching problem) and $f_0‟ = f_o/\alpha$ (according to warping/scaling function)

$c = f‟ – bf$

$= f_0/\alpha – bf_0$

$= (f_0 - \alpha b f_0) / \alpha$

Thus speaker normalization can be achieved by following the above process. The corpus taken for experiment had the entire male speaker thus the model trained was gender dependent. The frequency distribution of warping factors for all the speakers in training dataset is depicted in the histogram as:

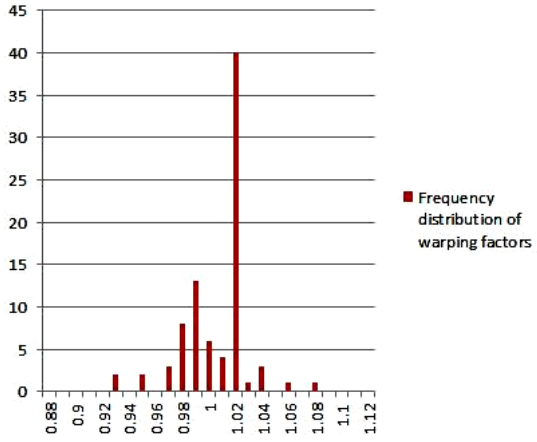

Fig. 3. Frequency Distribution of Warping Factors

The accuracy in VTLN based language identification is not improved much in comparison to gender dependent model as investigated. The reason behind this is less variation in speech frequency within a gender. Although results have shown that there is significant increase in accuracy when compared to a gender independent model as there are more variations in frequency of male and female [12].

IV. MFCC AND SHIFTED DELTA CEPSTRAL FEATURE EXTRACTION

The MFCC features are very well known and are calculated from speech signal in as shown in Fig. 1. After taking DCT, only few lower coefficients of the total DCT coefficients are kept. This is because the higher DCT coefficients represent fast changes in the filter bank energies. The initial coefficients are representation of vocal tract. Before the discovery of SDCs, MFCCs and their first and second derivative were used for language identification. MFCC features simulate the vocal tract of an individual so they were not very useful in this case. The SDCs which are computed from the deltas of MFCCs can capture a wide range of dynamics of an utterance. So, they showed significant performance improvement in language identification. They have four parameters N-d-P-k, "where N is the number of cepstral coefficients computed at each frame, d represents the time advance and delay for the delta computation, k is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and P is the time shift between consecutive blocks" [7].

Computation of SDC feature vector from MFCC can be depicted in figure as:

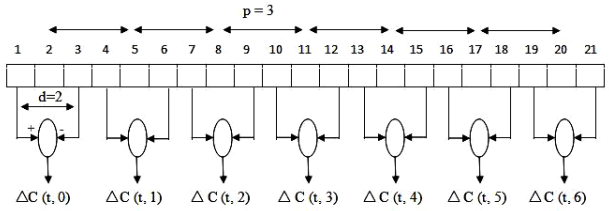

Fig. 4. Computation of SDC Feature Vector at Frame 'T' for Parameter 7-1-3-7 (N-D-P-K)

At a frame t, the SDC feature vector is:

$$SDC(t) = [\, c(t,0)^{T}, c(t,1)^{T}, c(t,2)^{T} \ldots\ldots c(t,6)^{T}\,]^{T}$$

Where $c(t,i) = c(t - i * p + d) - c(t + i * p - d)$

For instance, SDC feature vectors in a standard 7-1-3-7 configuration are 7(k) delta computations of 7 MFCC coefficients horizontally stacked over 21 frames (p*k) with difference of 3(p) frames. Also, delta computations are made between the preceding and succeeding frames from current frame (d =1). So, It is 49+7 dimensional vector with N *k delta MFFCs concatenated with 'N' MFCC coefficients. So, an algorithm which computes SDC takes mfcc_coeff (no_of_frames*7) matrix as input and give sdc (no_of_frames*56) matrix as output.

A SDC feature vector at a frame 't' uses k * P consecutive frames of cepstral coefficient which covers dynamics of a language. It could be the explanation of the accuracy achieved by SDC in discriminating language.

V. EXPERIMENTS

In this section, the training data set for the acoustic system and testing set is explained.

A. Training Data

We use an open speech corpus voxforge [20] for training the model. The closed set language identification consists of four languages taken from the corpus. For experiment, reduced training set of 1 hour is taken for each of the four languages: Dutch, Italy, Russian and French. The training set is well balanced as each language has 21 speakers with duration of utterances varying between 100 s - 300 s.

B. Test Data

The duration of test utterances is 30-120 seconds. Test set consists of 150 such utterances from the closed language set. In order to verify the credibility of language identification and annulling the argument that it is recognizing the speaker rather than language, all the test utterances are taken from different speakers and these 150 speakers of test set do not match with any of the speakers used in training the language models. The testing is done on 10 sec frame of a test file such that the language whose log-likelihood score is highest against the independent language models is the identified language of the particular frame. For each test file, the language identified in most number of 10 sec frames is declared as the language of the test file.

C. Model Training

For each of the four languages (Dutch, Italy, Russian and French), both MFCC and SDC feature vectors were extracted and multivariate GMMs consisting of 32, 64, and 128 mixture components were trained. The GMMs were trained using the parallel algorithm described in [21]. This brings down the time required for training the models considerably.

VI. RESULTS AND CONCLUSION

Several experiments on different models have been done on the data test against the GMM based language models by applying the explained preprocessing techniques. The depicted results in tables are in form of false rejection rate (FRR) and false acceptance rate (FAR). FRR is the number of test files of particular language having detected as another language divided by total number of files of the corresponding language. FAR is number of test files wrongly identified as particular language divided by total number of test files not in the corresponding language test set. The tables given below shows that GMM model on SDC features with 128 components perform the best with lowest FRR and FAR.

Table I. False rejection rate of test set showing effect of gmm components on reduced training set

Language	*False Rejection Rate (SDC features)*		
	32 (#GMM)	*64 (#GMM)*	*128 (#GMM)*
Dutch	29.73	27.03	32.44
Italy	21.06	21.06	18.43
Russian	27.28	30.31	15.16
French	21.43	14.29	11.91
Avg. FRR	24.88	23.17	19.49

Table II. False acceptance rate of test set showing the effect of gmm components on reduced training set

Language	*False Acceptance Rate (SDC features)*		
	32(#GMM)	*64(#GMM)*	*128(#GMM)*
Dutch			
Italy			
Russian	8.54	7.69	7.69
French	3.70	4.62	2.77
Avg. FAR	7.50	6.63	5.50

We conclude with the above results that as the number of gaussian mixture components are increased, the overall system performance increases. But the computation cost increases as the components are increased. More number of gaussians leads to better clustering of the large data points which are 56 dimensional feature vectors. Moreover, it can be seen that the performance of Dutch decreases with 128 gaussians. Further experiments were done to compare the results of MFCC and SDC features in LID systems. The histogram depicted below clearly shows that SDC features outperform the MFCC coefficients in identifying language.

The following results show the effect of CMS and normalization of SDC features on the accuracy and performance of the each of the four languages tested in closed set LID system.

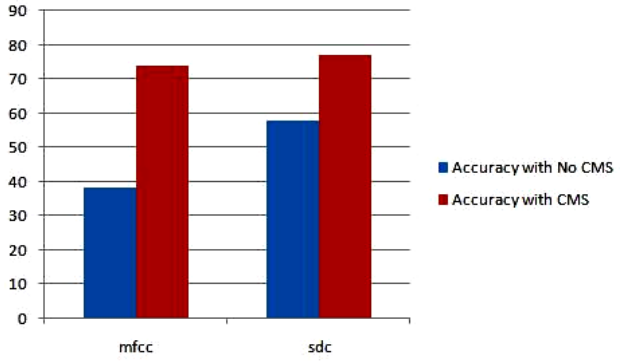

Fig. 5. Comparison between the Accuracy of Language Models Trained on MFCC and SDC (#64GMM)

Table III. FRR showing the effect of cms and normalization with 64component gaussians

Language	*FRR (#64 Gaussians Components)*		
	Raw SDC	*SDC with CMS*	*Norm SDC with CMS*
Dutch			
Italy			
Russian			
French	42.86	30.96	14.29
Avg.FRR	42.15	31.68	23.17

Table IV. Far showing the effect of cms and normalization with 64 components gaussians

Language	*FAR (#64 Gaussians Components)*		
	Raw SDC	*SDC with CMS*	*Norm SDC with CMS*
Dutch			
Italy			
Russian			
French			
Avg. FAR	13.55	10.03	6.63

GMM language models trained with raw sdc performs poor and shows correctness of 58%. With CMS there is a considerable rise in accuracy of system (correctness: 68%) as the channel distortion issue is addressed. As far as speaker normalization is concerned, VTLN technique does not show significant changes in the accuracy achieved as it is more or less same. Due to fewer variations within the same gender, the effect of VTLN cannot be observed. It is also important to normalize the SDC feature vector matrix before training and testing as well. The mean of the SDC feature vectors are kept 0 and variance 1. This improves the model training which in turn results in effective identification and better performance (correctness: 78%). False acceptance rate also depicts the same trend with 6.6%

avg. FAR as shown in table 4. The overall system performance can further be enhanced by increasing the training data as 1 hour training is not sufficient to achieve best results.

The overall system performance in terms of FRR and FAR is represented as histogram in fig.5 and Fig.6.

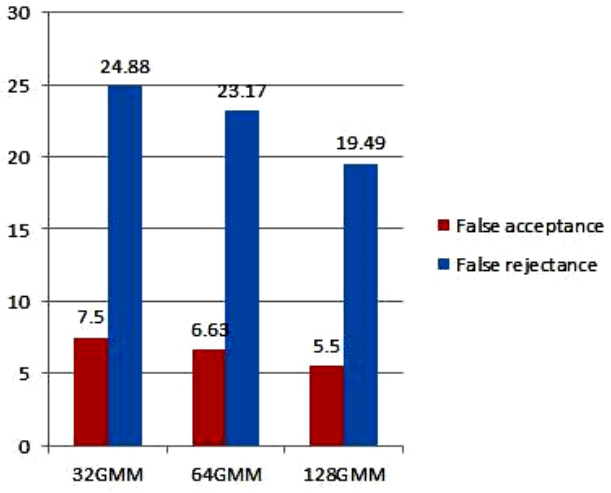

Fig. 5. Overall Systems FAR and FRR for Different GMM Components

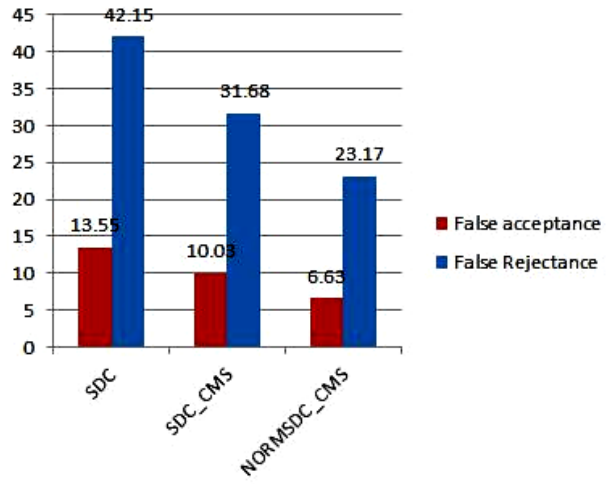

Fig. 6. Overall Systems FAR and FRR for Raw SDC, SDC with CMS and Normalized SDC with CMS

This paper has presented the complete feature extraction process for GMM based language identification. The primary aim of conducting the experiments was to identify the preprocessing techniques and study the effect of different acoustic features in building an LID system. The techniques hereby identified are important to develop a more flexible and adaptable system for language identification. Also, the results show significant improvement of the LID system using shifted-delta cepstral features over MFCC features and their delta derivatives.

The future work includes further analysis of the amount of data needed for training the language models. Further to increase accuracy more, implementation of discriminative modeling [22] is required instead of generative modeling. Also, future experiments will be conducted to assess the system performance using shorter speech utterances. Study of the overall system performance of LID system has to be done with full training set tested across the speech corpus.

REFERENCES

[1] Haizhou Li, Bin Ma and Kong Aik Lee," Spoken Language Recognition: From Fundamentals to Practice" Proceedings of the IEEE | Vol. 101, No. 5, May 2013

[2] Marc A. Zissman, Automatic Language Identification ofTelephone Speech, Lincoln Laboratory Journal, Volume 8, Number 2.

[3] E. Wong, J. Pelecanos, S. Myers and S. Sridharan, "Language Identification Using Efficient Gaussian Mixture Model Analysis" SST-2000: 8th Aust. Int. Conf. Speech Sci. & Tech.

[4] Marc A. Zissman, "Comparison of :Four Approaches to Automatic Language Identification of Telephone Speech", lEEE Transactions on Speech and audio processing, Vol. 4, no. 1, Jan 1996

[5] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, P. A. Torres-Carrasquillo, Support Vector Machines for Speaker and Language Recognition, Computer Speech & Language Volume 20, Issues 2–3, April–July 2006, Pages 210–229.

[6] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, D. A. Reynolds "Language Recognition with Support Vector Machines", Lincoln Laboratory, unpublished.

[7] Pedro A. Torres-Carrasquillo et all, "Approaches to language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features," Proc. Intl Conf.Spoken Language Processing, Denver, Sep. 2002, in press

[8] Johan de Veth, Louis Boves, "Comparison of Channel Normalization Techniques for Automatic Speech Recognition over the phone" In: Proc. ICSLP-96, pp. 2332-2335, 1996.

[9] Hynek Hermansky, and Nelson Morgan," RASTA Processing of Speech" IEEE transactions on speech and audio processing, vol. 2. no. 4,Oct I994

[10] Pavel Mateˇjka, Luka´sˇ Burget, Petr Schwarz and Jan Cˇernocky," Brno University of Technology System for NIST 2005 Language Recognition", 2006 IEEE

Odyssey - The Speaker and Language Recognition Workshop.

[11] Eddie Wong & Sridha Sridharan,"Utilise Vocal Tract Length Normalization for Robust Language Identification" Proceedings of the 9th Australian International Conference on Speech Science & Technology Melbourne, December 2 to 5, 2002.

[12] Steven Magwann, Don McAlaster, Jeremy Orelof, Barbara peskin "Speaker normalization on conversational telephone speech", Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on (Volume:1).

[13] Dominic Mazzoni, https://code.google.com/p/audacity/source/browse/audacity- src/trunk/src/effects/NoiseRemoval.cpp

[14] Md Sahidullah, Goutam Saha, "Comparison of Speech Activity Detection Techniques for Speaker Recognition" http://arxiv.org/abs/1210.0297

[15] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Brno University of Technology, 2009.

[16] S. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 8, pp. 2552– 2565, Nov. 2011.

[17] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, no. 1, pp. 12–40, 2010.

[18] Martin Westphal, "The use of cepstral means in coversational speech recognition", In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)

[19] Marcel Kockmann, Luka´s Burget, Jan Honza C ernocky "Application of speaker- and language identification state-of-the-art techniques for emotion recognition", Speech Communication 53 (2011) 1172–1185.

[20] http://www.repository.voxforge1.org/downloads/

[21] Ayush Kapoor, H. Hemani, N. Sakthivel, S. Chaturvedi, "MPI Implementation of Expectation Maximization Algorithm for Gaussian Mixture Models" Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2, Advances in Intelligent Systems and Computing Volume 338, 2015, pp 517-523

[22] Qu Dan and Wang Bingxi, Yan Honggang and Dai Guannan "Discriminative training of GMM based on Maximum Mutual Information for language identification", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China in press.