Gist of the project "Automatic Spoken Language Identification among five regional languages of Tripura State"

The project is about automatic spoken language identification . In this project language identification model is made for 5 languages -Kok-borok, Reang,Chakma, Bengali and Manipuri of Tripura state, where 5 male and 5 female speech sample are recorded in a noise free environment and collected for each language with a duration of 10 miniute by the students of NIELIT Agartala Centre.

The challenging issues faced during language identification: Variation in speaker characteristics, Variation in accents, Variation in environment and channel characteristics, variation of dialects, Similarities of languages, Extraction and representation of language-specific prosody.

Two types of language identification system is there. Implicit and Explicit . Explicit system language performance is better than implicit system but at the cost of additional complexity of using a subword unit recognizer which is not appropriate in Indian languages .The LID system that operates on features derived directly from the speech signal is appropriate in Indian context.

Language identification task involve 3 stages- feature extraction, modeling and evaluation.

language identification is carried out using spectral features such as mel-frequency cepstral coefficients (MFCC). MFCCs are determined from speech using the following steps-

1. Pre-emphasize the speech signal- Pre-emphasis refers to filtering that emphasizes the higher frequencies to balance the spectrum of voiced sounds that have a steep roll-off in the frequency region.

2. Divide the speech signal into sequence of frames with a frame size of 20 ms and a shift of 10 ms . Apply the Hamming window over each of the frames. This is done to enhance the harmonics, smooth the edges and to reduce the edge effect while taking the DFT on the signal.

3. Compute magnitude spectrum for each windowed frame by applying DFT.

4. Mel spectrum is a set of band-pass filters .A mel is a unit of measure based on the human ears perceived frequency. Mel frequency is computed by passing the DFT signal through mel filter bank. Triangular filter is most the most commonly used filter shaper and in some cases the Hanning filter is used.

5. DCT is applied to the log mel frequency coefficients (log mel spectrum) to derive the desired MFCCs. Since the vocal tract is smooth, the energy levels inadjacent bands tend to be correlated. The DCT is applied to the transformed mel frequency coefficients produces a set of cepstral coefficients. Traditional MFCC systems use only 8 to 13 cepstral coefficients. Prior to computing DCT the mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with quefrequency peak corresponding to the pitch of the signal and a number of formants representing low quefrequency peaks.

Dynamic MFCC features: The cepstral coefficients are usually referred to as static features, since they only contain information from a given frame. The extra information about the temporal dynamics of the signal is obtained by computing first order derivative called delta coefficients and second derivatives called delta-delta coefficients of cepstral coefficients. Delta coefficients tell about the speech rate, and delta-delta coefficients provide information similar to acceleration of speech.

GMM: unimodel probability density function with only one mean and covariance are unsuitable to model all variations of a single event in speech signals. Therefore, a mixture of single densities is used to model the complex structure of the density probability. The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight from all component densities.

The language-specific information from the extracted features is captured using Gaussian mixture models (GMMs).Here 5 GMMs are needed for 5 languages used.

Training the GMMs: To determine the model parameters of GMM of the speaker, the GMM has to be trained. In the training process, the maximum likelihood (ML) procedure is adopted to estimate model parameters. Most popular procedure is the iterative expectation maximization (EM) algorithm which uses iterative procedure because high non linear value of ML . The EM algorithm begins with an initial model and tends to estimate a new model such that the likelihood of the model increasing with each iteration. Em algorithm is repeated until a certain convergence threshold is obtained or a certain predetermined number of iterations have been made which in this case is 10.

Em algorithm follows this steps : Initialization , likehood computation, parameter update(mixture weight update, mean vector update, covariance matrix update).

In estimation of model parameters diagonal covariance matrix is used mostly because it has same model capability with the full matrices and cepstral features are more compactable, discriminative, and most important, they are nearly uncorrelated, which allows diagonal covariance to be used by the GMMs .

Maximum a posteriori (MAP) Adaptation: A trained GMM model needs sufficient data to create a model of the speaker. By using MAP of a background model (which encompasses different channel conditions, composition of speakers, acoustic conditions etc) estimating a statistical model with training data of short duration can be done. For each mixture i from the background model, posterior probability is calculated as Using Pr(i|xt), by which the statistics of the weight, mean and variance are calculated.

$$\Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)}.$$

$$n_i = \sum_{t=1}^{T} \Pr(i|x_t)$$

$$E_i(x_t) = \frac{\sum_{t=1}^{T} \Pr(i|x_t) x_t}{n_i}$$

$$E_i(x_t^2) = \frac{\sum_{t=1}^{T} \Pr(i|x_t) x_t^2}{n_i}.$$

i is the adaptation coefficient .Low values for i i! 0), will result in new parameter estimates from the data to be de-emphasized, while higher values (i ! 1) will emphasize the use of the new training data-dependent parameters. Generally only mean values are adapted.

Testing: In identification phase, mixture densities are calculated for every feature vector for all speakers and speaker with maximum likelihood is selected as identified speaker.

The decision rule , by Bays rule for the most probable speaker can be redefined as-

$$\hat{s} = \max_{1 \le s \le S} \sum_{t=1}^{T} \log P(x_t | \Omega_s)$$

Where S is the no. of speakers with T the number of feature vectors of the speech data set under test . Decision in verification is obtained by comparing the score computed using the model for the claimed speaker S given by P (Ωs |X) to a predefined threshold theta . The claim is accepted if P(ΩS |X) > theta , and rejected otherwise .

Experimental results and discussions:

In LID system is evaluated with varying number of Gaussian mixture components (32, 64 and 128).

2 types of LID studied here-

1. Gender independent LID study 2 .Gender dependent LID study.

In iteration 1, 45 speakers (speakers per language) are involved in building language model; whereas 5 speakers (1 from each language) which are not involved in building language model, are involved in language test wherein, 10 seconds duration of each speaker is used.

10 iteration of Independent LID shows average of performance for GMM=128 is 94.8% which is highest.

Performance % of Speaker independent LID per language is also obtained for the above test.

Gender dependent LID study has 3 parts :

1. male speakers are used for language modelling & language test.
2. female speakers are used for language modelling & language test.
3. cross gender LID study- has 2 parts.

a) male speakers for language modelling and female speaker for language test of all languages

b) female speakers are used for language modelling only and male speakers for language test.

For all above gender dependent test performance percentages are obtained .