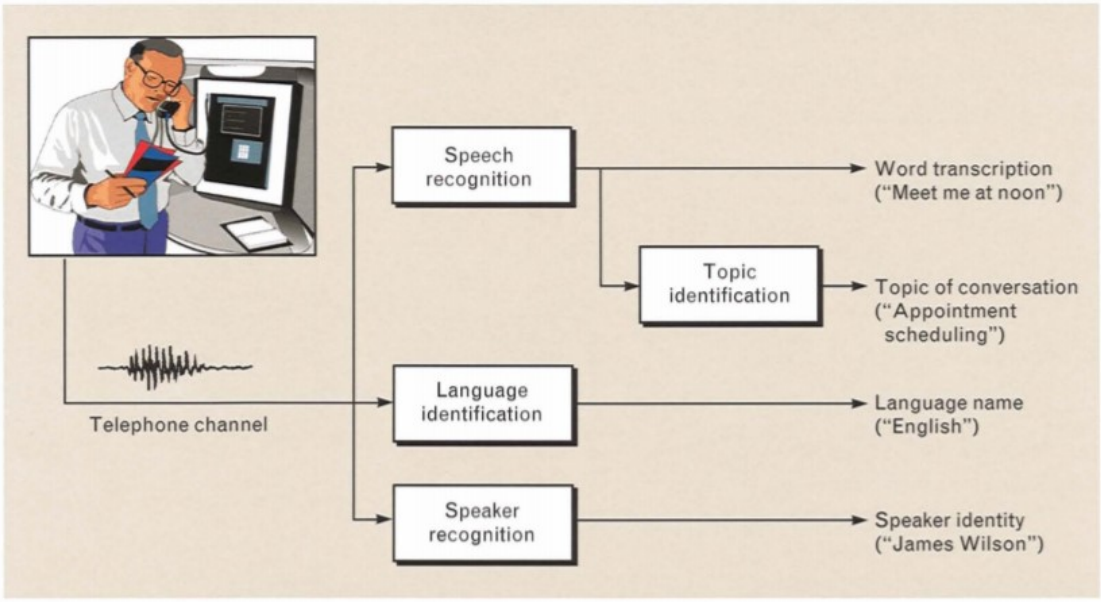
Chapter 2- Literature Survey

2.1 Marc A. Zissman's summary on Phonotactic Language Identification Systems

Lincoln Laboratory has investigated the development of a system that can automatically identify the language of a speech utterance. To perform the task of automatic language identification, they have experimented with four approaches:

1. Gaussian mixture model classification;
2. Single-language phone recognition followed by language modelling (PRLM);
3. Parallel PRLM, which uses multiple single-language phone recognizers, each trained in a different language;
4. Language-dependent parallel phone recognition.



https://www.ll.mit.edu/publications/journal/pdf/vol08_no2/8.2.1.languageidentification.pdf

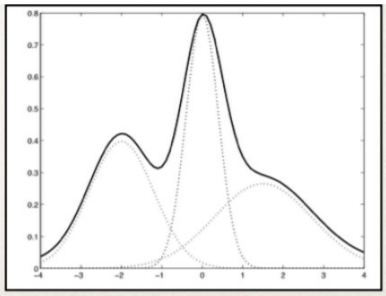
2.1.1 Approach1: GMM Based

A GMM language-ID system served as the simplest algorithm for this study. As shown below, GMM language ID is motivated by the observation that different languages have different sounds and sound frequencies. Under the GMM assumption, each feature vector V t at frame time t is assumed to be drawn randomly according to a probability density that is a weighted sum of unimodal multivariate Gaussian densities. Following is a brief introduction to GMM.

2.1.1.1 Gaussian Mixture Model

- GMMs are used to represent frame-based speech features
- Used for estimating acoustic likelihoods
- GMMs are a weighted sum of multivariate Gaussians

$$f(x|\mu,\Sigma) = \sum_{k=1}^{M} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} exp[(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)]$$



The basic Aim of Using Gaussian Mixture Model is to minimize the log likelihood function

❑ Consider log likelihood

$$\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathbf{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \right\}$$

ML does not work here as there is no closed form solution

Parameters can be calculated using Expectation Maximization (EM) technique

For which it uses the Expectation Minimisation (EM) Algorithm

❑ From Bayes rule

$$\gamma_k(\mathbf{x}) = p(k \mid \mathbf{x}) = \frac{p(k)p(\mathbf{x} \mid k)}{p(\mathbf{x})} = \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma_j)} \quad \text{where, } \pi_k = \frac{N_k}{N}$$

Latent Variable

2. E step. Evaluate the responsibilities using the current parameter values

$$\gamma_j(\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma_j)}$$

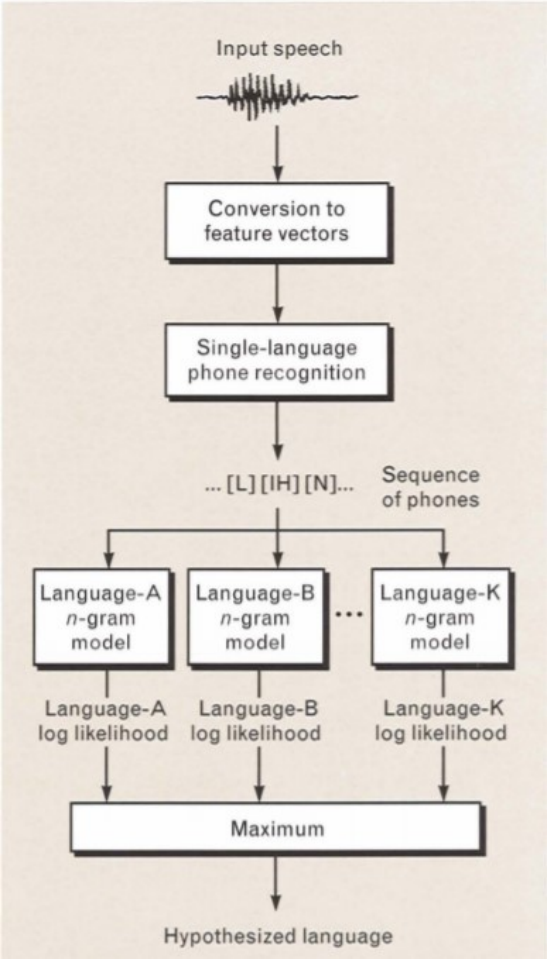
3. M step. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)} \qquad \Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_j)(\mathbf{x}_n - \boldsymbol{\mu}_j)^T}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)} \qquad \pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)$$

4. Evaluate log likelihood

$$\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathbf{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \right\}$$

If there is no convergence, return to step 2.

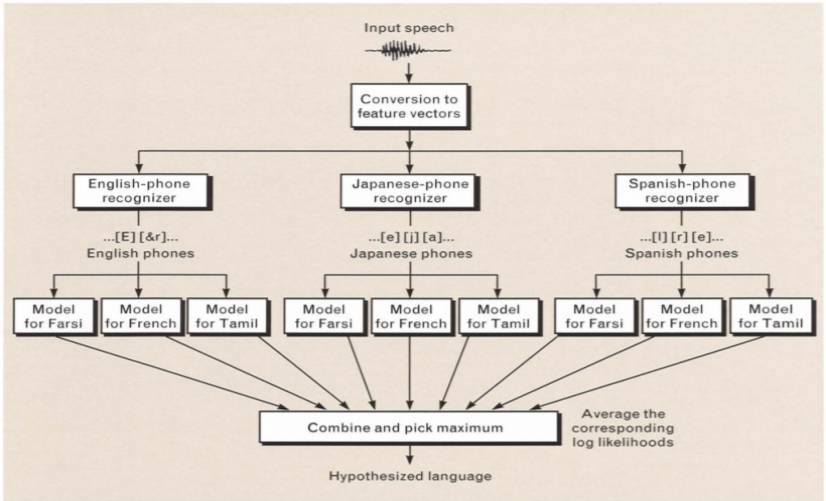


2.1.2 Approach 2: Phone Recognition Followed by Language Modelling (PRLM)

The second language-ID approach they tested comprises a single-language phone recognizer followed by language modelling with an n-gram analyzer, as shown in **FigureXXXXXXXXX** [Inclusion Needed] In the PRLM system, training messages in each language l are tokenized by a single-language phone recognizer, the resulting symbol sequence associated with each of the training messages is analyzed, and an n-gram probability-distribution language model is estimated for each language l. During recognition, a test message is tokenized and the likelihood that its symbol sequence was produced in each of the languages is calculated

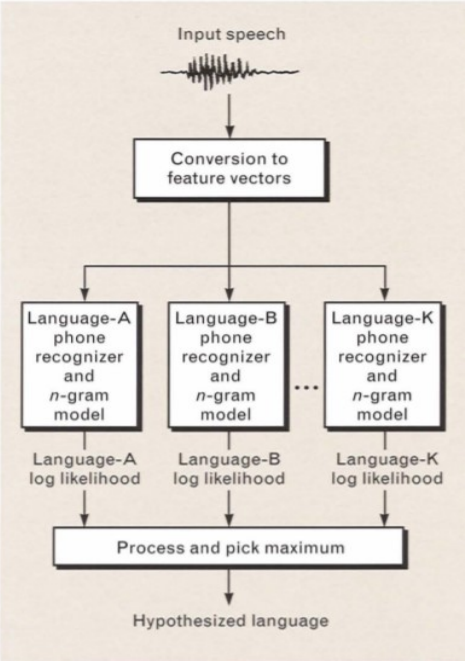
2.1.3 Approach3: Parallel PRLM

Although PRLM is an effective means of identifying the language of speech messages, we know that the sounds in the languages to be identified do not always occur in the one language that is used to train the front-end phone recognizer. Thus we look for a way to incorporate phones from more than one language into a PRLM-like system. Alternatively, the approach is simply to run multiple PRLM systems in parallel with the single-language front-end recognizers each trained in a different language. This approach requires that labelled training speech be available in more than one language, although the training speech does not need to be available for all, or even any, of the languages to be recognized. **FigureXXXX** shows an example of such a parallel PRLM system



2.1.4 Approach 4: Parallel Phone Recognition

The PRLM and parallel PRLM systems perform phonetic tokenization followed by phonotactic analysis. Though this approach is reasonable when labelled training speech is not available in each language to be identified, the availability of such labelled training speech broadens the scope of possible language-ID strategies; for example, it becomes easy to train and use integrated acousticl phonotactic models. If we allow the phone recognizer to use the language-specific phonotactic constraints during the Viterbi-decoding process rather than applying those constraints after phone recognition is complete (as is done in PRLM and parallel PRLM), the most likely phone sequence identified during recognition will be optimal with respect to some combination of both the acoustics and phonotactics. The joint acoustic-phonotactic likelihood of that phone sequence would seem to be well suited for language ID.



2.2 Language Identification based on GMM and Ceptral Features

The approaches discussed in the previous section were phonotactic in nature. These approaches require tokenization of speech in separate phones which is a computationally complex task and difficult to implement.

The ceptral features on the other hand are easy to compute and give a good representation of physical shape of vocal chords of the speaker. The variation of these coefficients can be used as a good measure for detecting the language.

2.2.1 Kshirod Sarmah and Utpal Bhattacharjee

Title- GMM based Language Identification using MFCC and SDC Features

In this paper, a baseline system for the LID system in multilingual environments has been developed using GMM as a classifier and MFCC combined with Shifted-Delta-Cepstral (SDC) as front end processing feature vectors. In this works, we used the Arunachali Language Speech Database (ALS-DB), a multilingual and multichannel speech corpus which was recently collected from the four local languages namely Adi, Apatani, Galo and Nyishi in Arunachal Pradesh including Hindi and English as secondary languages.The Gaussian mixture model with 1024 Gaussian components has been used for constructing language models. The individual language models were trained using the algorithm Expectation Maximization (EM) of 10 iterative steps. Training for the language model with equal number of male (50) and female (50) speaker's data with the same language. For the any one language suppose Adi language, the language model was created from 100 speakers utterance of Adi language. Similar approach is also applied for other five languages models Apatani, Galo, Nishi, Hindi and English.

2.2.2 Pedro A. Torres-Carrasquillo

Title- Language Identification using Gaussian Mixture Model Tokenization

Phone tokenization followed by n-gram language modeling has consistently provided good results for the task of language identification. In this paper, this technique is generalized by using Gaussian mixture models as the basis for tokenizing. Performance results are presented for a system employing a GMM tokenizer in conjunction with multiple language processing and score combination techniques. On the 1996 CallFriend LID evaluation set, a 12-way closed set error rate of 17% was obtained.

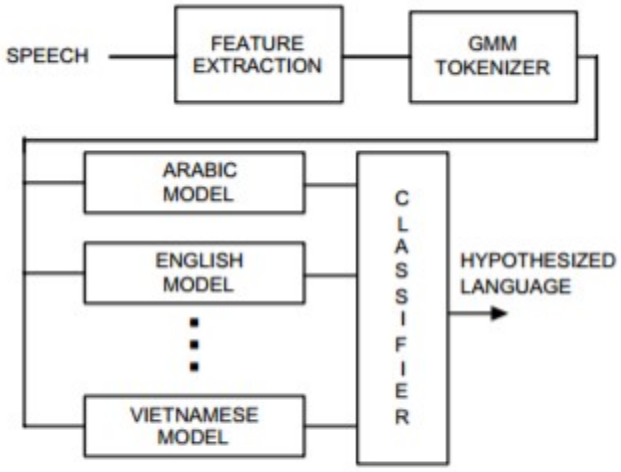


FIG 2. DIAGRAM FOR LID SYSTEM BASED ON GMM TOKENIZATION AND LANGUAGE MODELING.

2.3 Language Identification based on DNN and Ceptral Features

2.3.1 Ignacio Lopez-Moreno

Title: Automatic Language Identification Using Deep Neural Networks

This work studies the use of deep neural networks (DNNs) to address automatic language identification (LID). Motivated by their recent success in acoustic modelling, we adapt DNNs to the problem of identifying the language of a given spoken utterance from short-term acoustic features. The proposed approach is compared to state-of-the-art i-vector based acoustic systems on two different datasets: Google 5M LID corpus and NIST LRE 2009. Results show how LID can largely benefit from using DNNs, especially when a large amount of training data is available. We found relative improvements up to 70%,in Cavg, over the baseline system.

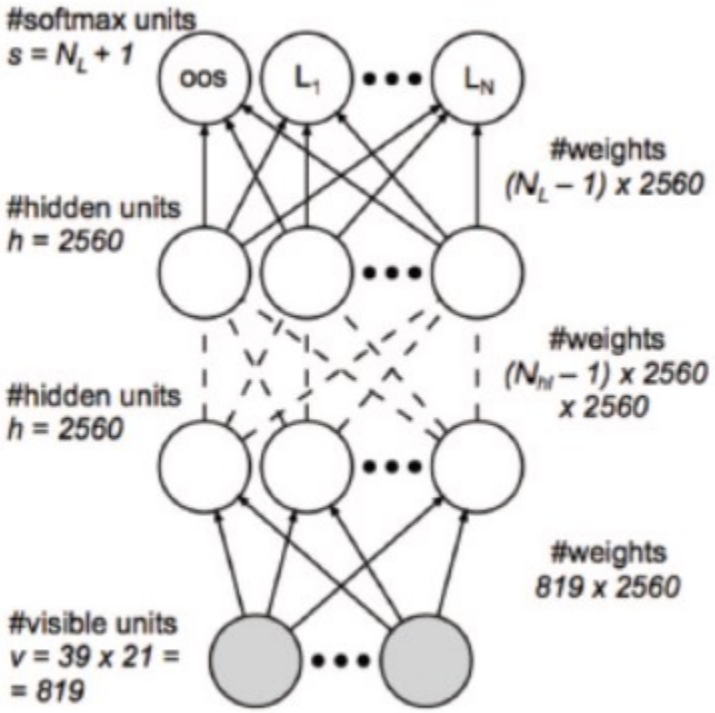


Fig. 1. DNN network topology

2.3.2 Julien De Mori

Title: Spoken Language Classification

This paper presented various approaches to language detection and they were compared. These methods were Support Vector Machine(SVM) based, Gaussian Mixture Model(GMM) based and Neural network.

However we were keen in its neural network implementation as it was quite simple and different from other implementations.

Instead of computing features for each 25ms frame and then making predictions on the frame level, they modelled each
clip as a multivariate Gaussian distribution by computing the mean and variance of each feature across each clip, changing the dimension of the feature vector to 78. This has the effect of averaging out any noise in the signal, as well as reducing the number of training examples, considerably decreasing computation time.

The minimum ERR rates for the features MFCC and SDC individually are 19.70% and 11.83% respectively.

2.3.3 Najim Dehak

Title: Language Recognition via Ivectors and Dimensionality Reduction

In this paper, a new language identification system is presented based on the total variability approach previously developed in the field of speaker identification. Various techniques are employed to extract the most salient features in the lower dimensional i-vector space and the system developed results in excellent performance on the 2009 LRE evaluation set without the need for any post-processing or backend techniques. Additional performance gains are observed when the system is combined with other acoustic systems.

The total variability space or i-vector approach concept was first introduced in the context of speaker verification. The basic idea of the total variability space consists of adapting the Universal Background Model (UBM) (which is trained on all the available language data for this paper) to a set of given speech frames based on the eigenvoice adaptation technique in order to estimate the utterance dependent GMM. The eigenvoice adaptation technique operates on the assumption that all the pertinent variability is captured by a low rank rectangular matrix T named the Total variability matrix. The GMM supervector (vector created by stacking all mean vectors from the GMM) for a given utterance can be modeled as follows

$$M = m + Tw + \epsilon$$

where m is the Universal Background Model supervector, the i-vector w is a random vector having a normal distribution N (0, I), and the residual noise term $\sim$ N (0, Σ) models the variability not captured by the matrix T . In our new modeling, we apply an SVM directly to the low dimensional i-vector (which is the coordinate of the speech segment in the total variability space) instead of applying the SVM in the GMM supervector space as done in. The process of training the total variability matrix T is a little bit different compared to learning the eigenvoice adaptation matrix . In eigenvoice training for speaker recognition, all the recordings of a given speaker are considered to belong to the same person; in the case of the total variability matrix however, we pretend that every utterance from a given speaker is produced by different speakers. If we follow the same total variability matrix training process for language identification, we assume that every utterance for a given language class is considered a different class.

It further uses dimensionality reduction by using a popular technique of Linear Discriminant Analysis and Neighbouring Component Analysis.

The maximum error obtained was 18.3% on 3 second uttarances.