Investigation of Acoustic Features in Text-Independent Speaker Verification

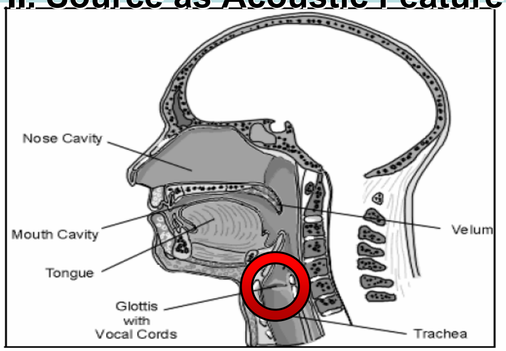
MERIT 2004

By: Thomas J. Plummer of University of Miami , Prof. Espy-Wilson & Gongjun Li. June 1- August 13, 2004

I. Introduction

- Objective of Speaker Verification (SV) is to verify the identity claim of a speaker from his or her speech
- Speech has strong biometric features like that of fingerprints and retinal pattern
- Text-Independent systems use long term statistics of speech signal to extract speaker specific data with 2 min. of speech for training & 3 sec. of speech for the verification process

II. Source as Acoustic Feature



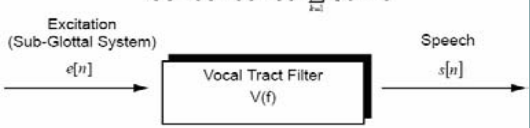
Path of Human Speech Production

Using Linear Prediction Coefficients (a_k) to define the predicted value:

$$\tilde{s}[n] = -\sum_{k=1}^{K} a_k s[n-k]$$

Then we can define the error or the *residual* as:

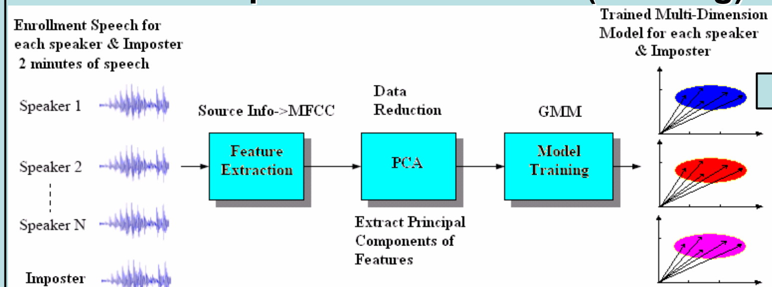
$$e[n] = s[n] - \tilde{s}[n] = s[n] + \sum_{k=1}^{K} a_k s[n-k]$$



The error risidual found from LPC method is the initial source of voiced speech as shown in the digital block diagram above

- Speech Source information is highly correlated unlike raw speech
- Does Source information exhibit desired Speaker dependent and text-independent characteristics?

III. Text-Independent Enrollment (Training)



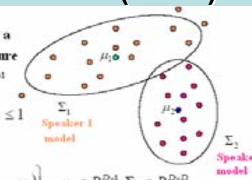
3.1. Feature Processing

- (MFCC) mel-frequency cepstral coefficients motivated by properties of human auditory system & the Ceptrum
- (PCA) Principal Component Analysis reduces data dimension by extracting top principal components containing most important Text- Independent Source information
- With Source based MFCC's represented with smaller dimension principal components, system accuracy increased by 1.21%

3.2 Gaussian Mixture Models (GMM)

GMM is a statistically adaptive model that consists of a weighted sum of M Gaussian densities used to measure the probability for a feature vector, say $x_0 \in R^{D\times 1}$

$$p(x_0 \mid \lambda) = \sum_{i=1}^{M} w_i g_i(x_0) \quad : \quad \sum_{i=1}^{M} w_i = 1 \;:\; 0 \le w_i \le 1$$



Gaussian density,

$$g_i(x_0) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x_0-\mu_i)'(\Sigma_i)^{-1}(x_0-\mu_i)\right\}, \quad \mu_i \in R^{D\times 1}, \Sigma_i \in R^{D\times D}$$

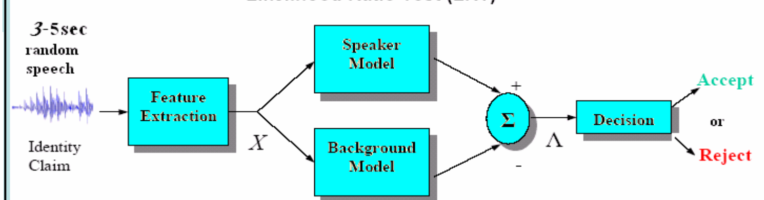
- A GMM is denoted as: $\lambda = \{w_i, \mu_i, \Sigma_i\}_1^M$
- The log-likelihood of a sequence of T feature vectors, $X = \{x_1, \ldots, x_T\}$

$$\log p(X \mid \lambda) = \sum_{t=1}^{T} \log p(x_t \mid \lambda)$$

- Statistical speaker representation with higher mixture degree for higher data diversity
- Source information is highly correlated, so lower mixture degree yields higher accuracy

IV. Text-Independent Verification (Testing)

Likelihood Ratio Test (LRT)



$p(\lambda_c \mid X)$: probability that X features belongs to the claimed speaker

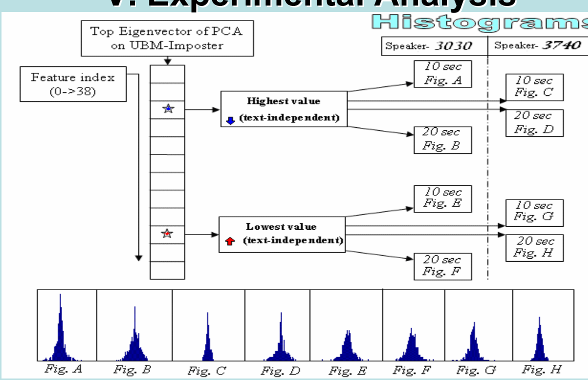
$p(\lambda_{\bar{c}} \mid X)$: probability that X features does not belong to the claimed speaker

Use Bayes, can measure (log) likelihood by:

$$\Lambda(X) = \log p(X \mid \lambda_c) - \log p(X \mid \lambda_{\bar{c}})$$

- The Imposter model a.k.a. Universal Background Model represents the similarities across all speakers in database
- Testing threshold based on the probability the speaker is the imposter
- Positive Λ would result in acceptance of identity claim

V. Experimental Analysis



- With Source information and PCA integrated into the present system, accuracy decreased 10.21% due to Speaker Independent Source properties seen in the similarities in Fig. (E-H) above for two different speakers
- Source information does show desired Text-Independent properties as histograms are similar from 10 to 20 seconds
- Source information benefits the system from fewer acoustical features and decreased mixtures