行政院及所屬各機關出國報告

（出國類別：實習）

實習對話式語音辨識技術

出國報告

服務機關：中華電信研究所
出國人　職　稱：助理研究員
　　姓　名：廖宜斌

出國地點：新加坡

出國期間：91 年 12 月 16 日至 92 年 6 月 6 日

報告時間：92 年 7 月 27 日

系統識別號:C09202789

公 務 出 國 報 告 提 要

頁數: 23　含附件: 否

報告名稱:

實習對話式語音辨識技術

主辦機關:

中華電信研究所

聯絡人／電話:

楊學文／03-4244218

出國人員:

廖宜斌　中華電信研究所　網路及多媒體應用技術研究室　助理研究員

出國類別: 實習

出國地區: 新加坡

出國期間: 民國 91 年 12 月 16 日 -民國 92 年 06 月 06 日

報告日期: 民國 92 年 07 月 27 日

分類號/目: H6／電信　／

關鍵詞:　對話式,語音,辨識技術

內容摘要: 本公司對於語音加值服務的開發一直不遺餘力，語音是使用者與電腦溝通最自然的介面。語音技術發展到現今已經可以提供相當程度的服務，未來會提供用戶更多個人化的服務，而在與使用者的對話過程中，使用者身分確認更是重要的一環，因此本次實習重點在聲紋識別的研究，了解最新的聲紋識別技術與方法，期待能爲使用者提供更人性化的服務

本文電子檔已上傳至出國報告資訊網

摘要

本公司對於語音加值服務的開發一直不遺餘力，語音是使用者與電腦溝通最自然的介面。語音技術發展到現今已經可以提供相當程度的服務，未來會提供用戶更多個人化的服務，而在與使用者的對話過程中，使用者身分確認更是重要的一環，因此本次實習重點在聲紋識別的研究，了解最新的聲紋識別技術與方法，期待能為使用者提供更人性化的服務。

目　　錄

1. 前言

本公司對於語音加值服務的開發一直不遺餘力，語音是使用者與電腦溝通最自然的介面。語音技術發展到現今已經可以提供相當程度的服務，未來會提供用戶更多個人化的服務，而從語音技術的觀點，語者的驗證與特徵求取將是關鍵的技術。Yankee group 在 2002 年行動用戶調查中發現 Voice Activated Dialing (VAD)語音撥號已經是前三個客戶最想獲得進一步的語音增強服務。這說明了語音技術在語音加值服務是不可缺少的角色，而個人化的服務是最讓人嚮往的。

為因應未來語音加值服務所需，職奉派前往新加坡大學(National University of Signapore)計算機學院(School of Computing)，多媒體實驗室(Multimedia Processing Lab)，作語音多媒體技術之研習。以下就研習的過程與內容做進一步的描述，研習期間的研究報告請參考附錄。

2. 實習過程

行程概要

整個行程從91年12月16日出發，至92年6月6日返國，共計173天。其受訓過程如下表：

日期	主題
12/16	起程
12/17~6/5	語者識別技術研習
6/6	回程

• 實習內容

聲紋識別(Voiceprint Recognition, VPR)，也稱為說話人識別(Speaker Recognition)，有兩類，即說話人辨認(Speaker Identification)和說話人確認(Speaker Verification)。前者用以判斷某段語音是若干人中的哪一個所說的，是“多選一”問題；而後者用以確認某段語音是否是指定的某個人所說的，是“一對一判別”問題。不

同的任務和應用會使用不同的聲紋識別技術，如縮小刑事偵查範圍時可能需要辨認技術，而銀行交易時則需要確認技術。不管是辨認還是確認，都需要先對說話人的聲紋建立模型，這就是所謂的“訓練”或“學習”過程。

從另一方面，聲紋識別有文本相關的(Text-Dependent)和文本無關的(Text-Independent)兩種。與文本有關的聲紋識別系統要求用戶按照規定的內容發音，每個人的聲紋模型逐個被精確地建立，而識別時也必須按規定的內容發音，因此可以達到較好的識別效果，但系統需要用戶配合，如果用戶的發音與規定的內容不符合，則無法正確識別該用戶。而與文本無關的識別系統則不規定說話人的發音內容，模型建立相對困難，但用戶使用方便，可應用範圍較寬。根據特定的任務和應用，兩種是有不同應用範圍的。比如，在銀行交易時可以使用文本相關的聲紋識別，因為用戶自己進行交易時是願意配合的；而在刑事偵查應用中則無法使用文本相關的聲紋識別，因為你無法要求犯罪嫌疑人配合。

在說話人辨認方面，根據待識別的說話人是否在註冊的說話人集合內，說話人辨認可以分為開集(open-set)

辨認和閉集(close-set)辨認。前者假定待識別說話人可以在集合外，而後者假定待識別說話人在集合內。顯然，開集辨認需要有一個對集外說話人的“拒絕問題”，而且閉集辨認的結果要好過開集辨認結果。本質上講，說話人確認和開集說話人辨認都需要用到拒絕技術，為了達到很好的拒絕效果，通常需要訓練一個假冒者模型或背景模型，以便拒絕時有可比較的對象，臨界值容易選定。而建立背景模型的好壞直接影響到拒絕甚至聲紋識別的性能。一個好的背景模型，往往需要通過預先採集好的若干說話人的語音，通過某種算法去建立。

聲紋識別可以說有兩個關鍵問題，一是特徵提取，二是模型匹配(模型識別)。

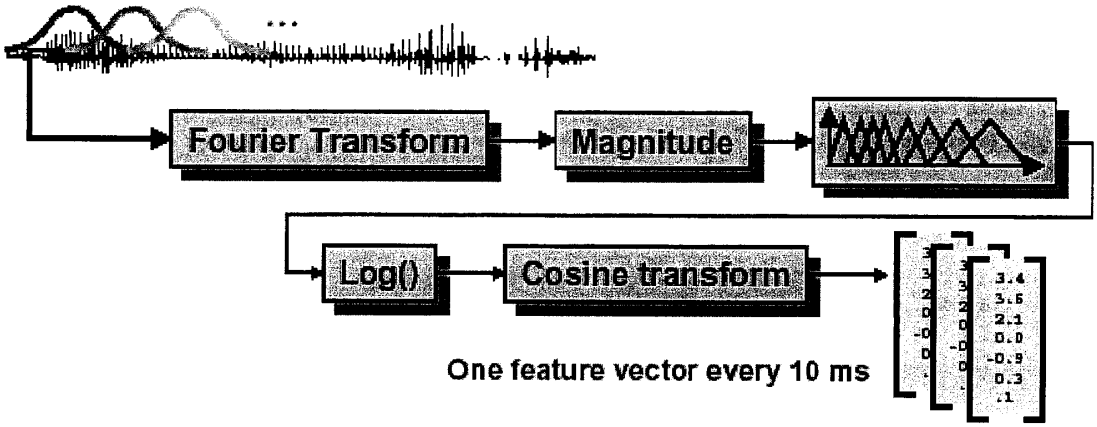
特徵提取的任務是提取並選擇對說話人的聲紋具有可分性強、穩定性高等特性的聲學或語言特徵。與語音識別不同，聲紋識別的特徵必須是“個性化”特徵。雖然目前大部分聲紋識別系統用的都是聲學層面的特徵，但是描述一個人的特徵應該是多層面的，從利用數學方法可以建立模型的角度出發，聲紋自動識別模型目前可以使用的特徵包括：(1)聲學特徵(倒頻譜)；(2)詞法特徵

(說話人相關的詞n-gram，音素n-gram)；(3)韻律特徵(利用 n-gram 描述的基音和能量“姿勢”)；(4)語種、方言和口音信息；(5)通道信息(使用何種通道)；等等。

相關技術:

特徵抽取

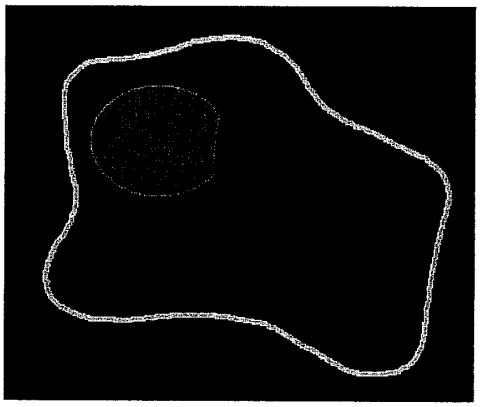
聲學的特徵(Acoustic Feature)是每個音框求取梅爾刻度式倒頻譜參數及對應差量參數，其求法如圖一。



圖一 **Mel-Cepstrum**

模型建立 Gaussian Mixture Model

語者的模型(圖二)是以數個獨立的高斯分布加權在一起的混合模型來模擬語者的觀測機率。



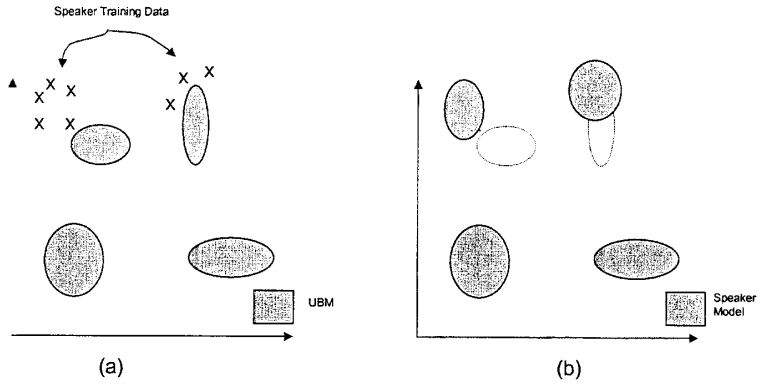
$$P(\vec{O}) = \sum_{m=1}^{M} w_m \mathcal{N}(\vec{O} \mid \mu_m, \Sigma_m)$$

圖二 **Gaussian Mixture Model**

模型調適

背景模型的好壞決定語者識別的效能因此通常用

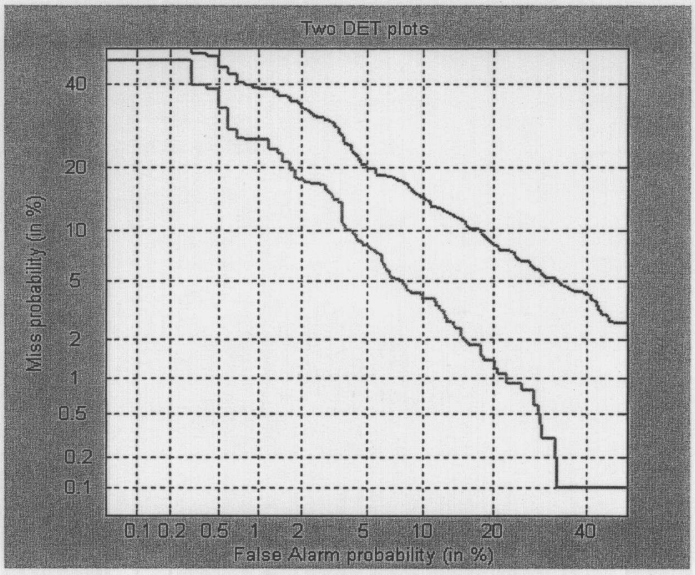
很多高斯分布來描述這一個模型而語者模型利用 MAP 的調適(圖三)只調整模型均值而得到的，實驗結果也證實會有不錯的效果。



圖三 MAP Adaptation

性能評估

錯誤拒絕(Miss)與錯誤接受(False Alarm)是語者識別會發生的兩種錯誤，而一個語者識別系統的好壞是希望兩種錯誤盡量降低，DET Plot(圖四)是評估系統效能，可以看出不同臨界值的選取對這兩種錯誤產生的影響，也可以比較兩個系統的好壞。



圖四 Detection Error Trade-off

3. 未來方向

3.1 強健性

要增加系統的強健性，我們必須減少訓練與測試之間的不匹配性。儘管 Cepstral mean subtraction (CMS) 與 RASTA 是兩種在特徵空間減少通道效應的方法，但是通道不匹配造成的問題依然嚴重。最近一種在頻譜空間上的轉換，使得訓練與測試語料間的機率分布盡量一致，用來克服不匹配的問題得到不錯的效果。聲學的特徵已經包含大部分語音的特徵，只要克服不匹配的問題系統通常有不錯的效果。

3.2 韻律

有別於聲學訊息 韻律也可以當作一種特徵它包含.

聲調(Intonation): 基週軌跡的起伏

重音(Stress): 音節加重語氣的地方

節奏(Rhythm): 說話速度與節拍

3.3 慣用語

為了要了解更高層次語者的語音特性 Dr. Doddington 統計 SwitchBoard 資料庫中每一位語者說話

內容的 Bigram 統計資料，了解每個人說話的特性，用此 bigram model 在語者識別上，效果令人滿意。因此他們建議這種更高層次的語音訊息值得進一步研究。

4. 建議

National Institute of Standards and Technology 機構位於美國 Gaithersburg, MA 這幾年主要工作在發展應用技術量測與標準的制定，在工業界扮演很重要的角色。

NIST 這幾年舉辦了一連串語者識別的相關競賽，提供了在這方面的領域很多重要的研究成果與研究方向，每年來自世界各地重要的研究室(MIT Lincoin Lab , Dragon inc , Microsoft China , IBM ,...)都有參與年度評比。評比的內容集中在電話語音上面的語者識別與本公司的應用方向符合，評比的目的是探索語者識別的新技術與量測這些技術的方法。

我建議可以每年參與這個評比取得語者識別技術最新的訊息，了解未來趨勢，相信會有助於本公司開發語者識別相關應用服務的競爭力。

.5. 附錄

GMM-Based Speaker Verification System

I Bin Liao

Email : snet@cht.com.tw , liaoib@comp.nus.edu.sg

Abstract

This report summarizes the GMM-based speaker verification baseline system. Firstly, we normalize acoustic vectors at the filter-bank level such that the test data distribution matches the training data distribution. And then, the system is built around the likelihood ratio test for verification, using GMMs for likelihood functions, a universal background model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from UBM. Experiments with this baseline system on the development data of the NIST 2001 speaker recognition evaluation corpus are reported.

Introduction

Background noise or distortions caused by the transmission usually lead to mismatch between the test conditions and the training data. The mismatch can severely deteriorate the system performance. To improve the performance, the mismatch should be reduced between training and testing. Cepstral mead subtraction (CMS) and RASTA are two of the standard feature-based approaches. But channel and handset mismatch can still cause lots of errors after CMS or RASTA. Recently, a new approach is feature warping which transform the distribution of filter-bank coefficients such that the test data

distribution matches the training data distribution. This technique brought significant improvements for speaker verification compared to standard techniques.

In recent years, GMM-based systems have been applied to the annual NIST Speaker Recognition Evaluation (SRE). These systems, fielded by different sites, have consistently produced state-of-the-art performance. In particular, a GMM-based system developed by MIT Lincoln Laboratory [1], employ Bayesian adaptation of speaker models from a universal background model and handset-based score normalization, has been the basis of the top performance in the NIST SREs since 1996.

The aim of the baseline system is to setup evaluation environment and compare to the system which gets the result on NIST 2001 speaker recognition evaluation corpus. In this report, we describe in Section 2 the framework of the baseline system. In Section 3, we report on some experiments and some comparisons.

The Framework of baseline system

Platform Architecture

The baseline is composed of the following main modules: speech feature extraction, modeling, score normalization, and decision. Feature warping is applied during MFCC feature extraction. The modeling module is based on Gaussian mixture models (GMMs) with maximun a posteriori (MAP) adaptation of speaker independent model. Score normalization, is also applied. The decision module

makes the decision by comparing a normalized likelihood ratio to a threshold and plots the DET curves [4].

Feature Warping

The warping can be viewed as a nonlinear transform.

$$Y_k^{eq}[t] = T_k(Y_k[t]) \qquad (1)$$

$Y_k[t]$ denote output of the kth Mel scaled filter after applying a 10^{th} root compression at time frame t. The transform T_k use here is a power function.

Before actually applying the power function transformation the filter output values $Y_k[t]$ are scaled to the interval [0,1] by dividing them through the maximal value Q_{k,N_Q}. Then the transformation is applied and the resulting values are scaled back to the original range. The symbols used in the following equations are : N_Q the number of quantiles. Q_i^{train} the i th quantile on the training data, these are estimated globally not dependent on the filter channel k. $Q_{k,i}$ the i th quantile estimated on the test utterance for filter channel k.

$$T_k(Y_k[t]) = Q_{k,N_Q}(\alpha_k(\frac{Y_k[t]}{Q_{k,N_Q}})^{\gamma_k} + (1-\alpha_k)\frac{Y_k[t]}{Q_{k,N_Q}}) \qquad (2)$$

The transformation parameters α_k and γ_k are chosen to minimize the squared distance between the current quantiles $Q_{k,i}$, and the training quantiles Q_i^{train}:

$$\{\gamma_k, \alpha_k\} = \arg\min_{\{\gamma_k, \alpha_k\}} \left(\sum_{i=1}^{N_Q - 1} (T_k(Q_{k,i}) - Q_i^{train})^2 \right) \quad (3)$$

It is useful to apply the normalization both in training and test. That is, the overall distribution of all training data is used as reference (target histogram), and the data of each test and training speaker is transformed to match the target histogram.

Adaptation of Speaker Model

The UBM is a large GMM trained to represent the speaker-independent distribution of features. In our system, we derive the hypothesized speaker model by adapting the parameters of the UBM using the speaker's training speech and a form of Bayesian adaptation (Fig. 1). In previous result [1], the best overall performance is from adapting only the mean vectors. In our system, we also adapt mean only.

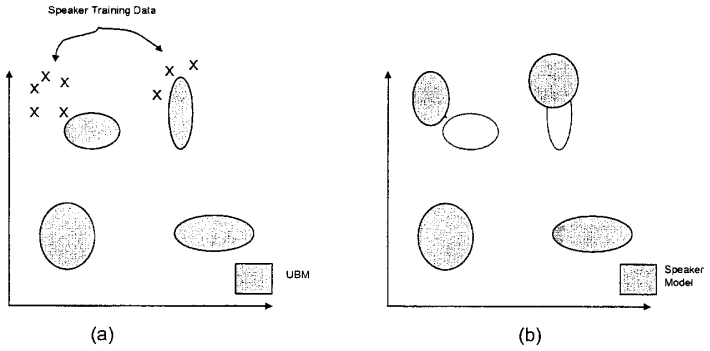


Fig. 1 Pictorial example of two steps in adapting a hypothesized speaker model. (a) The training vectors (x's) are probabilistically mapped into the UBM mixtures. (b) The adapted mixture parameters are derived using the statistics of the new data and the UBM mixture parameters. The adaptation is data dependent, so UBM mixture parameters are adapted by different amounts.

•

Score Normalization

The log-likelihood ratio for a test sequence of feature vector X is compute as $\Lambda(x) = \log P_{avg}(X \mid \lambda_{hyp}) - \log P_{avg}(X \mid \lambda_{UBM})$ (4)

where $\log P_{avg}(X \mid \lambda_{hyp})$ is average log-likelihood of frame of the test utterance for hypothesis model and $\log P_{avg}(X \mid \lambda_{UBM})$ is average log-likelihood of frame of the test utterance for UBM model.

This equation gives a relative log-likelihood score between a speaker and a background model for the observation X. The effect of Eq. (4) on the speaker verification task is that quality mismatches which occur between the test observation X and the speaker model λ_{hyp} will have a corresponding effect on the background model λ_{UBM}.

Therefore, effects which lead to a bias in $P_{avg}(X|\lambda_{hyp})$ are eliminated in $\Lambda(x)$ due to the relative log-likelihood scoring.

The most frequently score-normalization techniques used are T-Norm and Z-Norm. These two score-normalization methods lead to better system performance but they need additional speech data or external speakers to be computed. In our baseline system, we use a logistic-regression model to map scores into probability estimates. The logistic regression model is

$$p(y=\pm 1 \mid x,w,b)=\frac{1}{1+\exp(-(b+yw^{T}x))} \qquad (5)$$

Given the training data set $\{(x_1,y_1),.....,(x_N,y_N)\}$, we wish to maximize the likelihood of the observed data. To do this, we make use of gradient information of the likelihood, and then ascend the likelihood and we can obtain estimated parameters w and b.

Experiments

Database

The baseline system is evaluated on the cellular telephone speech, used in the NIST speaker recognition evaluation for 2001. 2 hours of speech from 38 male and 22 female speakers, with 2 minutes each speaker, are used for training the background model. There are 74 male and 100 female target speakers. Each speaker has 2 minutes of speech for training. 20380 gender-matched verification

trials from the test set. The duration of each test segment varies from a few seconds to one minute, with the majority of tests falling into a range between 15 to 45 seconds. The ratio between target and imposter is roughly 1:10.

Evaluation measure

The evaluation of the speaker verification system is based on Detection Error Tradeoff curves, which show the tradeoff between false alarm (FA) and false rejection (FR) errors. Besides, the equal error rate (EER), there is also a detection cost function (DCF) defined for the NIST evaluation:

$$DCF = C_{FA} \Pr(FA \mid N) \Pr(N) + C_{FR} \Pr(FR \mid T) \Pr(T) \quad (4)$$

Where Pr(N) and Pr(T) are the a prior probability of non-target and target tests with Pr(N)=0.99 and Pr(T)=0.01. And the specific cost factors $C_{FA} = 1$ and $C_{FR} = 10$. So the point of interest is shift towards low FA rates.

Experiment results

As shown in Fig, 2, we extract 19 dimensional MFCC first. The frame is set to 10ms. Then delta coefficients are calculated to form 38 dimension feature vector. UBM is a 512mixture GMM and target model is obtained by adapting the mean of UBM.

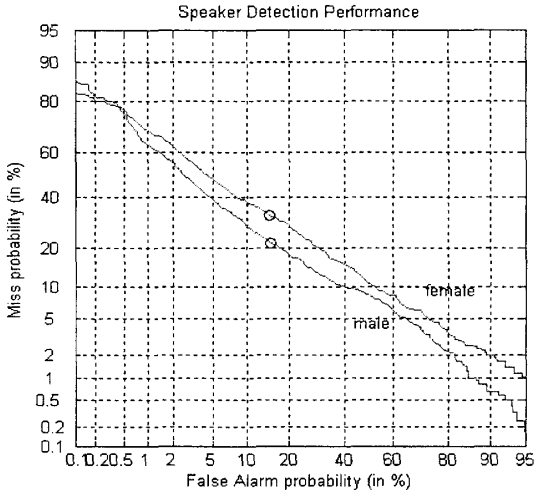


Fig. 2

As shown in Fig, 3 and Fig. 4, we extract 19 dimensional MFCC first. The frame is set to 10ms. Then delta coefficients are calculated and use normalized energy term to form 40 dimension feature vector. UBM is a 512mixture GMM and target model is obtained by adapting the mean of UBM. As we can see in the figure, energy is helpful to reduce equal error rate.

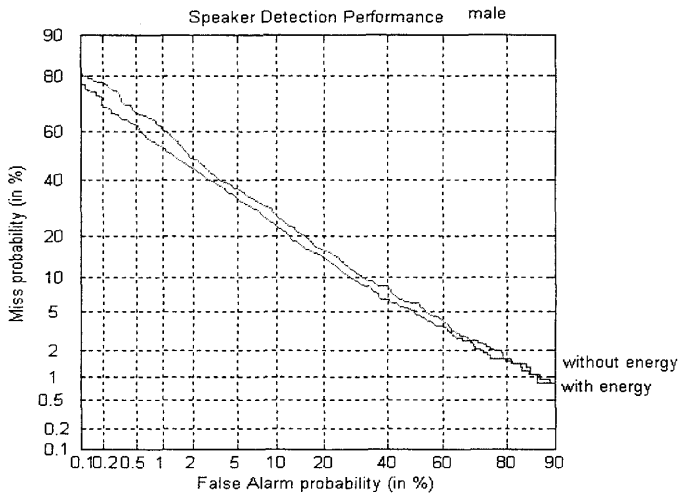
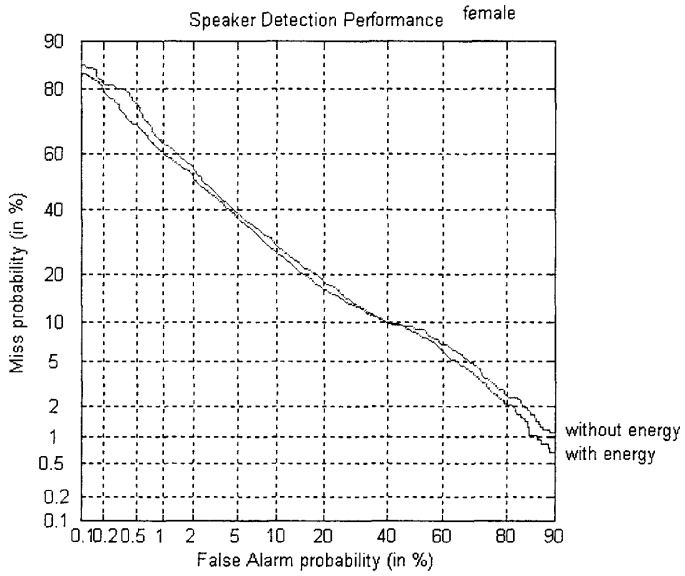

Speaker Detection Performance male
Miss probability (in %)
90
80
60
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40 60 80 90
False Alarm probability (in %)
without energy
with energy

Fig 3


Speaker Detection Performance female
Miss probability (in %)
90
80
60
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40 60 80 90
False Alarm probability (in %)
without energy
with energy

-
- Fig4
-
-

As shown in Fig, 5, we use 40 dimension feature vector. UBM is a 512mixture GMM and target model is obtained by adapting the mean of UBM. And we applied logistic regression method in score normalization. Since logistic regression is a discriminative score normalization technique, it can be combined with target speaker model training. In this report, I only use it to discriminate target scores and imposter scores.

As we can see in the figure, logistic regression is an effective discriminative score normalization technique. It has better system performance in terms of ERR and DCF.

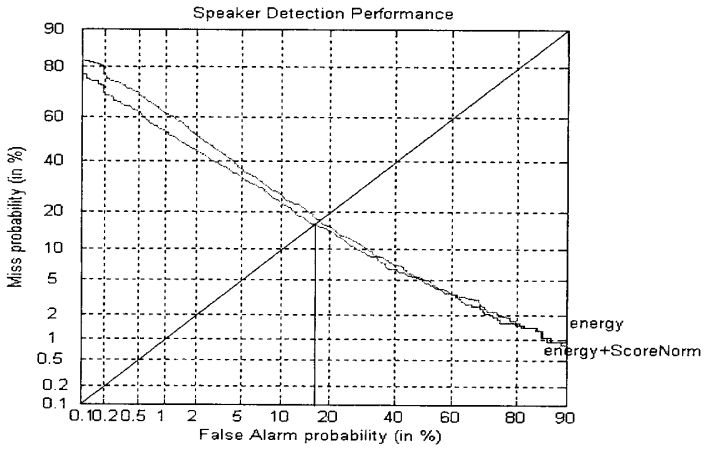


Fig 5

As shown in Fig, 6, we map acoustic vectors at the filter-bank level such that the test data distribution matches the training data distribution. But it can be seen in the figure that system performance decreases. I am still thinking about this result.

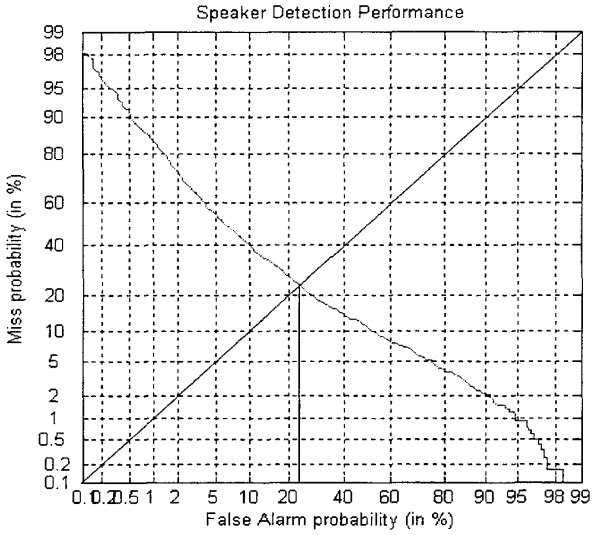


Fig 6

Reference

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn M.I.T. Lincoln Laboratory: "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing 10, 19-41, 2000

[2] Chengyuan Ma and Eric Chang, Microsoft Research Asia: "Comparison of Discriminative Training Methods for Speaker Verification"

[3] Florian Hilger, Sirko Molau, Hermann Ney University of Technology Ahornstr : "Quantile Based Histogram Equalization for Online Applications".

[4] Bing Xiang, Upendra V. Chaudhari, Jiri Navratil, Ganesh N. Ramaswamy, Ramesh A. Gopinath IBM T.J. Watson Research Center: "Short-Time Gaussianization for Robust Speaker Verification"

[5] "The NIST Year 2001 Speaker Recognition Evaluation Plan", http://www.nist.gov/speaker

[6] Kemal Somez Elizabeth Shriberg SRI International, Menlo Park, CA 94025 "Modeling Dynamic Prosodic Variation For Speaker Verification"

[7] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification", *ICSLP*, 1992.

[8] A. E. Rosenberg, O. Siohan & S. Parthasarathy, "Speaker verification using minimum verification error training", *ICASSP*, 1998.

[9] B.-H. Juang, W. Chou & C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. Speech and Audio Processing, Vol.5, No. 2,* pp.257-265, 1997.