IDIAP COMMUNICATION

IDIAP

Martigny - Valais - Suisse

1997 NIST Evaluation: Text independent speaker detection (verification)

Dominique Genoud * Gilles Caloz *

IDIAP–Com 97-03

July 1997

Dalle Molle Institute for Perceptive Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

* IDIAP, CP592 CH-1920 Martigny

1 Introduction

Every year the US government institute NIST (National Institute for Standardization and Technologies) [3] organize a speaker verification/identification evaluation. Any research institute can participate. In 1997 IDIAP choosed to participate in collaboration with ENST (Paris/France). The common part of the IDIAP and ENST system was the threshold calculator, and the first tests on GMM systems.

2 Classification task

The 1997 classification evaluation is a text independent speaker detection (verification) task. the **training set** is composed of "One session", "One handset" and "Two handset" data. the speech duration for each speaker on each training condition is 1 minute.

The **test set** has 3 different speech duration 3 seconds, 10 seconds and 30 seconds. these tests segments have to be used on the three different training conditions.

Two types of results have to be given:

- A true/false decision for each test speech segment. A COST function C_{det} will be calculated.

  $C_{det} = C_{fr} * P_{fr|target} * P_{target} + C_{fa} * P_{fa|nonTarget} * P_{nonTarget}$

  were

  $C_{fr} = 10; C_{fa} = 1; P_{target} = 0.01; P_{nonTarget} = 1 - P_{target} = 0.99$

- A score for each test speech segment, this allow to generate a COR curve.

Details of the evaluation protocol are available on [4].

The 1997 evaluation focused on the different handset conditions.

3 The IDIAP system

3.1 Parametrization

The basis vector parameters are 16 LPC coefficient, 16 δLPC coefficient, 16 $\delta\delta$LPC coefficient, δenergy and $\delta\delta$energy.

We used only the last 8 (c9-c16) LPC coefficient, the 16 δLPC, the 16 $\delta\delta$LPC and energy coefficients.

The speech signal is windowed at 32[ms] shifted each 10[ms], pre-emphasis 0.95, liftering 16. Cepstral mean subtraction (CMS) is used for channel compensation.

We didn't used any other normalization (i.e handset normalization).

3.2 Classifier

As classifier an MLP system is used [10]. The size of the MLP is 462 input neurons, 100 neurons on the hidden layer and 2 neurons on the output layer. The 462 input neurons correspond to 11 consecutive input vectors, in order to capture more long term speech events. the 2 neurons of the output layer are the local log likelihood score (LLS) of the target speaker (LLS_{sp}) and the non-target speaker (LLS_{ns})(also named *world* or *cohort*). These LLS are summed along the speech segment (using N frames) to obtain a total log likelihood TLL_{sp} for the target speaker and TLL_{ns} for the non-target speaker.

$TLL_{sp} = \sum_{1}^{N} LLS_{sp}(N)$

$TLL_{ns} = \sum_{1}^{N} LLS_{ns}(N)$

$TLLR = TLLsp - TLLns$

The final score used for each speech segment is $TLLR$ which correspond to a log likelihood ratio [9].

One MLP system is built for each target speaker. The *cohort* speaker data were created from around 40 male and 40 female speakers speech extracted from Switchboard database. The total amount of speech for each training condition was balanced with the amount of data for each target speaker (i.e. 1 minute).

3.3 Threshold settings

In order to decide if a test speech segment was said by the target speaker, an *a priori* decision threshold has to be set. The threshold th_{tsp} chosen here is derived from the Furui threshold setting method [1, 2].

$th_{tsp} = C1 * (\mu_{ntsp} - \sigma_{ntsp}) + C2$

tsp=target speaker, $ntsp$=non-target speaker

An extended threshold determination is used here:

$th_{tsp} = a * \mu_{ntsp} * \sigma_{ntsp} + b * \mu_{ntsp} + c * \sigma_{ntsp}$

in this case, the followed transformation is applied:

$Th'_{tsp} = TLLR - (A * \mu_{ntsp} * \sigma_{ntsp} + B * \mu_{ntsp} + C * \sigma_{ntsp})$

so the threshold Th'_{tsp} becomes speaker independent, and it becomes possible to adjust the threshold to improve the cost function (see 2). The data used as non-target speaker data (for threshold setting) came from the training set of the 1996 NIST evaluation data. In order to determine μ_{ntsp} and σ_{ntsp} the non-target speaker data were "passed through" each target speaker model to obtain μ_{ntsp}, σ_{ntsp} and the three constants A,B,C.

4 Results

There were 9 participants to the 1997 NIST Evaluation, to see the IDIAP results, please consult [5]. To see the other labs results please consult [5] (IDIAP internal only). As there were 9 different tests IDIAP is third for the best and 6th for the worst place. This variability in the results can be explained because IDIAP didn't use any handset normalization, but the MIT [7] and Dragon [6] used one.

4.1 MIT handset detector

The MIT used a carbon/electret microphone detector based on a GMM (Gaussian Mixture Model) of 1024 Gaussian. They used 5 hours of speech coming from LLHDB database to train their detector.

4.2 Dragon handset detector

Dragon used a 512 mixtures GMM detector trained on NTIMIT database.

5 Formats/Software used

5.1 NIST CD distribution

The 1997 NIST evaluation is divided in 6 CDs:

CDNo	Name	Contents
CD1	training set female	sid97_fe
CD2	training set male	sid97_ma
CD3-4	test set female	sid97e1f-sid97e2f
CD5-6	test set male	sid97e1m-sid97e2m

5.2 STRUT software

To have more details about the STRUT toolkit see [8].

5.3 File output format of STRUT/MLP

The output format for the files coming from a enhanced version of STRUT (STRUT + shell scripts) is:

- one test per line:
FileName IDprocl NbofFrame LLKspeaker LLKcohort IDvrai

FileName	Name of the speech file.
IDprocl	Name of the speaker which has to be verified.
NbofFrame	Number of speech frames.
LLKspeaker	Log Likelihood of the speech data on the true speaker output of the model.
LLKcohort	Log Likelihood of the speech data on the cohort output of the model.
IDvrai	Name of the current speaker from which the speech is taken.

for example: 0005.wav 1103 22 -11.462109 -2.864248 1010

5.4 Threshold setting, decision programs

In order to set the *a priori* thresholds :

- Generate speaker μ_{ntsp} and σ_{ntsp} (program impodist) using impostor access.
- Calculate extended Furui's method constants A,B and C (Program Indiveval).
- Decide if it is or not the target speaker using μ_{ntsp},σ_{ntsp},A,B,C (Program Scoreval2).

5.5 Programs and scripts available

The Programs ar available at IDIAP in the
/home/polyphone6/NIST/Progs directory
the STRUT scripts are available at IDIAP in the
/home/polyphone6/NIST/STRUT directory.

5.6 Evaluation file format

the final output format for NIST evaluation is (one test per line):
Sex TrainCond TargetID Duration FileName Decision Score

Sex	male or female
TrainCond	1 session(1s), 1 handset(1h), 2 handset(2h).
TargetID	Name of the target speaker.
Duration	3,10 or 30 seconds.
FileName	Name of the speech file.
Decision	True of False.
Score	the current score ($TLLK'$ in our case).

References

[1] Sadaoki Furui, *Cepstral Analysis Technique for Automatic Speaker Verification*, in IEEE transactions on acoustics, speech and signal processing, VOL ASSP-29, No2,p 258 April 1981.

[2] Dominique Genoud, Frédéric Bimbot, Guillaume Gravier, and Gérard Chollet, *Combining methods to improve speaker verification decision*, in Proc. of the 4th International Conference on Spoken Language Processing, ICSLP 96, Philadelphia, USA, Oct. 3-6, 1996

[3] NIST *http://www.itl.nist.gov/div894/894.01/*

[4] *Nist Site Evaluation plan* http://www.itl.nist.gov/div894/894.01/sp_v1p1.htm

[5] *Nist Speaker detection results* http://www.idiap.ch/ genoud/public/Nist/index.htm

[6] *Dragon Dictate system* http://www.idiap.ch/ genoud/Nist/desc/dragon_sys1.desc

[7] *MIT system* http://www.idiap.ch/ genoud/Nist/desc/mitll.GMM-UBM.sys1.desc

[8] *The Strut manual* file://localhost/home/speech01/STRUT/src/strut-1.03/doc/html/users-guide/users-guide.html

[9] G. Saporta, *Probabilités analyse des données et statistique*, p301, Ed Technip, Paris, 1990.

[10] Nelson Morgan and Hervé Bourlard *An introduction to the hybrid HMM/Connectionist approach*, IEEE signal processing magazine, p 24–42 may 1995