ABC System description for NIST SRE 2010

May 6, 2010

1 Introduction

The ABC submission is a collaboration between:

- Agnitio Labs, South Africa
- Brno University of Technology, Czech Republic
- CRIM, Canada

We submit three different fusions of subsystems, as well as a mothballed version of the BUT 2008 JFA system. All four submissions are exercised only on the core condition.

As in 2008, our efforts were directed at handling different telephone and microphone channels. Additionally this year, we concentrated on English speech and on the special challenges posed by the new DCF weighting. All of our development decisions were made in order to optimize for *actual* DCF, with the new weighting. We made no special effort to compensate our systems for speech of low or high vocal effort.

2 Submitted Systems

Except for the mothballed system, our submissions are fusions of sub-systems. The fused scores are linear combinations of sub-system scores, followed by a saturating non-linearity. The combination weights, as well as two parameters controlling the non-linearity are numerically optimized to minimize a cross-entropy criterion, which is defined relative to the evaluation prior of 0.001. The fusion output is intended to act as a well-calibrated *log-likelihood-ratio*, which is thresholded at $-\operatorname{logit}(0.001) \approx 6.9$ to make hard decisions.

Some of the fusions include so-called quality measures. The quality measure for a trial is computed as a weighted bilinear combination of two quality vectors, one derived from the training segment, another from the test segment. These combination weights are optimized simultaneously with the other fusion parameters.

We used three different development trial lists assembled from English SRE 2008 data, on which we did development testing as well as optimization of the fusion parameters. These sets were:

tel-tel: Telephone speech in both sides of the trial.

int-tel: Interview microphone speech on one side and telephone speech on the other side of the trial.

int-int: (Courtesy of MIT) Interview microphone speech on both sides of the trial.

For 2010 trials, involving telephone conversations recorded on auxiliary microphone, we defaulted to the fusion trained on the corresponding int-tel or int-int list.

2.1 System 1: Primary

tel-tel CRIM, BUT JFA'10, SVM CMLLR-MLLR, Agnitio I-Vector, Prosodic JFA, BUT I-Vector Full Cov, PLDA I-Vector, BUT JFA'08

int-tel CRIM, BUT JFA'10, SVM CMLLR-MLLR, BUT I-Vector Full Cov, BUT I-Vector LVCSR, PLDA I-Vector, BUT JFA'08, BUT Quality Measures

int-int CRIM, BUT JFA'10, SVM CMLLR-MLLR, BUT I-Vector Full Cov, BUT I-Vector LVCSR, PLDA I-Vector, BUT JFA'08, BUT Quality Measures

2.2 System 2: Contrastive

This system is the same as the primary System 1, except that:

- for **tel-tel**, the Agnitio Quality Measures were included,
- for **int-tel** and **int-int**, the BUT Quality Measures were excluded.

2.3 System 3: Contrastive

This system is the same as the primary System 1, except that the CRIM subsystem, which generally performed best in development testing, was removed for all conditions.

2.4 System 4: Mothballed

This monolithic system is a linear calibration (without the non-linearity) of the BUT JFA'08 system on its own.

3 Development data

Our development data organization this year was designed specially to meet the challenge of the new DCF. The development data was split into two (mostly) non-overlapping parts:

Sub-system training: This contains pre-2008 SRE data, as well as some Switchboard and Fisher databases. This data is telephone and microphone speech in English and other languages.

Fusion training and development testing: This contains 2008 SRE and 2008 SRE follow-up data, all labelled as English.

The sub-system training subset was organized and employed much as we did in 2008, but the fusion training and development testing needed special attention because of the new DCF.

The new DCF weights false-alarms 1000 times more than misses. This ratio differs by *two orders of magnitude* from the old DCF weighting. This has two implications:

- The decision threshold becomes so high that very few, or even no false-alarms occur in the original SRE 2008 trial lists. This makes empirical measurement of the false-alarm rate very unreliable.
- Duplicate pins, erroneously assigned to the same speaker in the Mixer telephone collection, caused target trials to be mislabelled as non-targets. Systems that correctly score those targets above the threshold are then penalized with a relative weight of 1000 for each such trial.

We addressed the first problem by creating extended trial lists, with many more non-target trials than in the original SRE trial lists. In order to decide how many non-targets we needed, we used two methods, which mostly agreed:

- Doddington's Rule of 30 [1], suggests there should be at least 30 false-alarms for *probably approximately correct* empirical false-alarm measurement. (The same applies to misses, but at the new DCF, the problem usually occurs with false alarms.) We tried to make our trial lists large enough so that number of false-alarms at the minimum DCF point of the system under evaluation exceeds 30.

- We calibrated each system under evaluation and then plotted normalized versions of actual DCF and minimum DCF as a function of the target prior[1]. That is, we parametrized the DCF as a function of the prior, which we swept from 0.001% to 50%. We observed that for low prior, as the number of false-alarms at min DCF dropped below 30, calibration generally broke down, with normalized actual DCF often above 1. For priors to the right of the 30 false-alarm boundary, calibration was generally good, with actual DCF close to min DCF.

The sizes of the final trial lists were as follows:

tel-tel: 1 228 targets, and 2 569 689 non-targets.

int-tel: 54 607 targets, and 1 807 366 non-targets.

int-int: 26 220 targets, and 1 136 693 non-targets.

Initially these lists were large enough for development testing of single systems, but later some fusions became so accurate that the number of false alarms at min DCF still fell below 30.

We addressed the second problem of labelling errors in the development data with a mixture of automatic and manual means. We tried to correct two kinds of errors:

- Some gender labelling errors, discovered via a mixture of automatic and manual means, were reported to us by Fabio Valente of Loquendo. Gender labelling errors affect all gender-dependent systems adversely.

- We used the Agnitio i-vector system and the speaker partitioning methods described in [2] to build a special duplicate pin detector. It uses *all* segments attributed to each pin, to test for each pair of pins, the hypothesis that the two pins may be of the same speaker. After listening tests confirmed that several of the top-scoring pin-pairs were indeed most probably the same speaker, we arbitrarily removed a number of the top scoring pin pairs from the non-target trials in our tel-tel list. This had a dramatic effect on the male DET-curves and less so on the female. We believe that the false non-targets thus removed were doing more damage to our development testing and fusion training than the few high-scoring true non-targets that we may have also removed in this way.

[1] Our MATLAB toolkit for making normalized DCF plots is available here: `http://focaltoolkit.googlepages.com/dcf-curves`.

4 Sub-systems

This section gives detailed descriptions of the various sub-systems that we used in the fusions.

4.1 BUT JFA'08

This is the configuration of the BUT JFA system as it was submitted to SRE 2008, with details described below.

4.1.1 Feature Extraction

Short time gaussianized MFCC 19 + energy augmented with their delta and double delta coefficients, making 60-dimensional feature vectors. The analysis window is 20 ms with a shift of 10 ms.

Short-time gaussianization uses a window of 300 frames (3 sec). For the first frame, only 150 frames on the right are used and the window grows till 300 while we move in time. When we approach the last frame, we use only 150 frames on the left side.

4.1.2 VAD

Speech/silence segmentation is performed by our Hungarian phoneme recognizer [3], where all phoneme classes are linked to the *speech* class. Segments labelled *speech* or *silence* are generated, but not merged yet to preserve smaller segments — a post-processing with two rules based on short time energy is applied first:

1. If the average energy in a *speech* segment is 30 dB less than the maximum energy of the utterance, the segment is labelled as silence.

2. If the energy in the other channel is greater than the maximum energy minus 3 dB in the processed channel, the segment is also labelled as silence.

After this post-processing, the resulting segments are merged together. Only speech segments are used. In case of 1-channel files, rule #2 is not applied. The *interview data* we processed as 1-channel and, we took ASR transcripts of the interviewer and removed his/her speech segments from our segmentation files based on time-stamps provided by NIST.

4.1.3 UBM

Two gender-dependent universal background models (UBMs) are trained on Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, and NIST SRE 2004 and 2005 telephone data. In total, there were 16307 recordings (574 hours) from 1307 female speakers and 13229 recordings (442 hours) from 1011 male speakers. We used 20 iterations of the EM algorithm and for each we do splitting up to 256 Gaussians and 25 iterations for 512 and up. No variance flooring was used.

4.1.4 JFA

The Factor analysis (FA) system closely follows the description of *Large Factor Analysis model* in Patrick Kenny's paper [4] with MFCC19 features. The two gender dependent UBMs are used to collect zero and first order statistic for training two gender dependent FA systems.

First, for each FA system, 300 eigenvoices[2] are trained on the same data as the UBM, although only speakers with more than 8 recordings were considered here. For the estimated eigenvoices, MAP estimates of speaker factors are obtained and fixed for the following training of eigenchannels. A set of 100 eigenchannels is trained on NIST SRE 2004 and 2005 telephone data (5029 and 4187 recordings of 376 females and 294 males speaker respectively). Another set of 100 eigenchannels is trained on SRE 2005 auxiliary microphone data (1619 and 1322 recordings of 52 females and 45 males speaker respectively). Another set of 20 eigenchannels is trained on SRE08 interview development data (3 males and 3 females). All three sets are concatenated to form final U matrix. We used linear scoring to get the final score for the trial [5].

4.1.5 Normalization

Finally, scores are normalized using condition dependent zt-norm. For the telephone condition we used 200 females and 200 males z-norm and t-norm telephone segments, derived each from one speaker of NIST SRE 2004,05,06 data. For the interview and microphone we used together 400 utterances from which 200 from interview and 200 from microphone sets were randomly chosen from NIST 2008 interview data (speakers not present in dev set) and MIX05,06 microphone data respectively.

[2] We refer to *eigenvoices* and *eigenchannels* following the terminology defined in [4] although these sub-spaces are estimated using the EM-algorithm, not PCA.

Table 1: *Number of speakers per training list.*

set	#number of speakers
znorm.int.m	17
znorm.int.f	24
tnorm.int.m	17
tnorm.int.f	24
znorm.mic.m	41
znorm.mic.f	45
tnorm.mic.m	39
tnorm.mic.f	48
znorm.tel.m	200
znorm.tel.f	200
tnorm.tel.m	200
tnorm.tel.f	200

4.2 BUT JFA'10

This is a new configuration of the BUT JFA system, with details as described below.

4.2.1 Feature Extraction

Short time gaussianized MFCC 19 + energy augmented with their delta and double delta coefficients, making 60 dimensional feature vectors. The analysis window has 20 ms with shift of 10 ms.

Short-time gaussianization uses a window of 300 frames (3 sec). For the first frame, only 150 frames on the right are used and the window grows till 300 while we move in time. When we approach the last frame, we use only 150 frames on the left side.

4.2.2 VAD

Speech/silence segmentation is performed by our Hungarian phoneme recognizer [3], where all phoneme classes are linked to the *speech* class. Segments labelled *speech* or *silence* are generated, but not merged yet to preserve smaller segments — a post-processing with two rules based on short time energy is applied first:

1. If the average energy in a *speech* segment is 30 dB less than the maximum energy of the utterance, the segment is labelled as silence.

2. If the energy in the other channel is greater than the maximum energy minus 3 dB in the processed channel, the segment is also labelled as silence.

After this post-processing, the resulting segments are merged together. Only speech segments are used. In case of 1-channel files, rule #2 is not applied. The *interview data* we processed as 1-channel and, we took ASR transcripts of the interviewer and removed his/her speech segments from our segmentation files based on time-stamps provided by NIST.

4.2.3 UBM

Two gender-dependent universal background models (UBMs) are trained on Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, and NIST SRE 2004 and 2005 telephone data. In total, there were 16307 recordings (574 hours) from 1307 female speakers and 13229 recordings (442 hours) from 1011 male speakers. We used 20 iterations of the EM algorithm and for each we do splitting up to 256 Gaussians and 25 iterations for 512 and up. No variance flooring was used.

4.2.4 JFA

The Factor analysis (FA) system closely follows the description of *Large Factor Analysis model* in Patrick Kenny's paper [4] with MFCC19 features. The two gender dependent UBMs are used to collect zero and first order statistics for training two gender dependent FA systems.

First, a gender dependent system aimed at telephone speech was trained, using statistics from files from NIST 200[456] and Switchboard2 Phase[23] and Switchboard Cellular Part[12]. For the female sub-system, 21663 recordings were used, for the male-subsystem there were 16969 recordings. Both sub-systems had 300 eigenvoices and 100 eigenchannels. The eigenchannel and eigenvoice matrices were initialized randomly and then trained with 10 EM iterations of maximum likelihood and minimum divergence each. No D matrix was used.

Then, other eigenchannel matrices were trained (using the telephone eigenvoice matrix) on different training data: One with 50 eigenchannels on NIST 2008 SRE interview data (2514 female recordings, 1397 male recordings), and one with 100 eigenchannels on NIST 200[56] SRE microphone data (3702 female recordings, 2961 male recordings).

The three eigenchannel matrices (telephone, interview, and microphone) per gender were then combined to form two systems per gender: a telephone+interview+microphone system and a telephone+microphone system. Both used the telephone eigenvoice matrix.

The telephone+interview+microphone system was used to score the int-tel and int-int conditions, the telephone+microphone system was used to score the tel-tel conditions.

4.2.5 Normalization

Finally, scores are normalized using condition dependent zt-norm. For the telephone condition we used 200 females and 200 males z-norm and t-norm telephone segments, derived each from one speaker of NIST SRE 2004,05,06 data. For the interview and microphone we used together 400 utterances from which 200 from interview and 200 from microphone sets were randomly chosen from NIST 2008 interview data (speakers not present in dev set) and MIX05,06 microphone data respectively.

See table 1.

4.3 Agnitio I-Vector

The Agnitio I-Vector system used 60-dimensional MFCC features and a 2048-component UBM to extract 400-dimensional *i-vectors*[3] as proposed in [5, 6]. The i-vectors were modelled with the *two-covariance model* as proposed in [2].

Score normalization is a symmetrized version of ZT-norm called S-norm. The cohorts were synthetically generated by generating normally distributed i-vectors with the between-speaker covariance.

The UBM is gender-independent, but the i-vector extractors, two-covariance models and score normalizations are gender-dependent. This system was designed for telephone speech and was run only on tel-tel trials in 2010. For further details see [2].

4.4 BUT I-Vector Full Cov

This system uses 400-dimensional i-vectors extracted via 60-dimensional features and a 2048-component *full-covariance* UBM.

[3] The name *i-vector* is mnemonic for a vector of *intermediate* size (bigger than an acoustic feature vector and smaller than a supervector), which contains most of the relevant *information* about the speaker *identity*.

4.4.1 Feature Extraction

Short time gaussianized MFCC 19 + energy augmented with their delta and double delta coefficients, making 60 dimensional feature vectors. The analysis window has 20 ms with shift of 10 ms.

Short-time gaussianization uses a window of 300 frames (3 sec). For the first frame, only 150 frames on the right are used and the window grows till 300 while we move in time. When we approach the last frame, we use only 150 frames on the left side.

4.4.2 VAD

Speech/silence segmentation is performed by our Hungarian phoneme recognizer [3], where all phoneme classes are linked to the *speech* class. Segments labelled *speech* or *silence* are generated, but not merged yet to preserve smaller segments — a post-processing with two rules based on short time energy is applied first:

1. If the average energy in a *speech* segment is 30 dB less than the maximum energy of the utterance, the segment is labelled as silence.

2. If the energy in the other channel is greater than the maximum energy minus 3 dB in the processed channel, the segment is also labelled as silence.

After this post-processing, the resulting segments are merged together. Only speech segments are used. In case of 1-channel files, rule #2 is not applied. The *interview data* we processed as 1-channel and, we took ASR transcripts of the interviewer and removed his/her speech segments from our segmentation files based on time-stamps provided by NIST.

4.4.3 UBM

One gender-independent universal background model was represented as a full covariance, 2048-component GMM. It was trained on the NIST SRE 2004 and 2005 telephone data (376 female speakers in 171 hours of speech, 294 male speakers in 138 hours of speech). Variance flooring was applied in each iteration, where the threshold was computed as an average variance from each previous iteration, scaled by 0.1.

4.4.4 I-vector system

I-vector system aims at modelling overall variability of the training data and compressing the information to a low-dimensional vector. The technique is closely related to JFA in the sense that each training segment acts as a separate speaker. Speaker (and/or channel) modelling techniques are then applied on these low-dimensional vectors. This way, an i-vector system can be viewed as a front-end for further modelling.

To filter out the channel information from the i-vectors, LDA was used as in [5]. Each trial score is then computed as a cosine distance between two such vectors.

The i-vectors extractor was trained on the following telephone data: NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2 giving 8396 female speaker in 1463 hours of speech, and 6168 male speakers in 1098 hours of speech (both after VAD).

The tel-tel LDA matrix was trained on the same data as the i-vector extractor, except the Fisher data was excluded, resulting in 1684 female speakers in 715 hours of speech and 1270 male speakers in 537 hours of speech. The int-tel and int-int LDA matrices were trained on the same data as tel-tel, augmented with all possible microphone and interview data, resulting in 1830 female speakers in 832 hours of speech and 1387 speakers in 621 hours of speech.

4.4.5 Normalization

Simplified symmetrical normalization—*s-norm*—is applied to the scores: For each trial, the enrolled model and the test segment (both represented by an i-vector) are scored against some s-norm cohort (400 speakers, gender dependent) and their score distributions are estimated. Two scores are then computed for the trial: one normalized by the enrolled model score distribution, and one by the test segment score distribution. The final score is computed as an average of the two normalized scores.

The normalization cohorts were chosen to satisfy the corresponding condition (e.g., interview models were normalized by telephone segments in the int-tel condition).

4.5 BUT I-Vector LVCSR

Apart from the UBM, this system is identical to the full-covariance i-vector system as described in 4.4. It uses i-vectors extracted via 60-dimensional features and a 2048-component *diagonal-covariance* UBM derived from an LVCSR system.

4.5.1 UBM

One gender-independent universal background model was represented as a diagonal covariance, 2048-component GMM. It was based on Gaussian components extracted from our LVCSR system trained on fisher (2000 hours) + 300 hours from switchboard and callhome. All Gaussians were pooled together, and repetitive clustering was applied to satisfy maximum likelihood increase in each step.

4.6 PLDA I-Vector

The system is based on the same 400-dimensional i-vectors as in the case of the BUT I-Vector Full Cov system. However, this time the i-vectors are modelled using a PLDA [7] model.

4.6.1 PLDA Training

The algorithms for PLDA training and scoring were implemented by Agnitio. We note that in contrast to [7], the EM-algorithm for training the PLDA model parameters makes use of an additional *minimum-divergence* update [8, 9], which helps the algorithm to converge more quickly and to a generally better solution than plain EM.

tel-tel condition PLDA model with 90 eigenvoices and 400 eigenchannels (full rank) is trained using mixer 04,05,06 and switchboard telephone data. No snorm is applied.

int-tel and int-int condition PLDA model with 90 eigenvoices and 400 eigenchannels (full rank) is trained using pooled switchboard, mixer 04,05,06 telephone and microphone data and heldout 2008 interview data. After the model is trained, V and D matrices are fixed and U matrix is retrained once using telephone data, once using 05,06 microphone data (plus corresponding telephone recordings) and once using heldout 2008 interview data. For scoring trials, the original matrix U (trained on everything) and the three telephone, microphone and interview specific U matrices are stacked into one matrix of 1600 eigenchannels. No snorm is applied.

4.7 Prosodic JFA

4.7.1 Features

The prosodic feature generation closely follows the description in [10]. The features incorporate duration, pitch and energy measurements. Pitch and energy values are estimated every 10 ms and energy is further normalized by its maximum. The temporal trajectory of pitch and energy is modeled by Discrete Cosine Transformation (DCT), over a fixed frame long temporal window of 300 ms, with a 50 ms frame shift. The first 6 coefficients of both, transformed pitch and energy trajectories, form a fixed length feature vector. Only voiced frames (where pitch is detected) are used, all other frames are cut out prior to DCT transformation. Further, duration information is appended as one discrete coefficient, that is the number of voiced frames within the 30 frame interval.

4.7.2 Model

The Factor analysis (FA) system closely follows the description of *Large Factor Analysis model* in Patrick Kenny's paper [4]. First, gender dependent UBMs with 512 components each, are trained with variance flooring on Switchboard and SRE04-06 data. After PCA initialization of $\mathbf{V}$ and $\mathbf{U}$, 100 eigenvoices are iteratively re-estimated on Switchboard and SRE04-06 data. For the *tel-tel* condition, we further estimate 40 eigenchannels on Switchboard and SRE04-06 data. For *int-tel* and *int-int* conditions, 10 telephone channels are estimated on SRE04-05 data, concatenated with 10 interview channels, trained on SRE08 interview data and additional 10 microphone channels, estimated on SRE05-06 microphone recordings. During verification, we use fast linear scoring [11]. All scores are further zt-normalized, using task dependent z- and t-norm lists.

4.8 SVM CMLLR-MLLR

4.8.1 Feature Extraction

LVCSR system was trained on 2000 hours of Fisher data + 300 hours of Switchboard and Callhome. The used features were PLP12_0_D_A_T (in HTK notation), with VTLN applied, and HLDA with dimensionality reduction (52 to 39). Speaker adaptive training was done using fMPE + MPE models with crossword triphones, WER 24% on NIST eval01 task. 2-class CMLLR was used to model speech and silence, and 3-class MLLR was used to model 2 data clusters and silence. From the resulting 5 matrices, only the triplet of the speech-related matrices was taken.

For MLLR and CMLLR estimation, we ran 4 iterations with intermediate data-to-model realignment.

We did not run our own ASR recognition, rather we used the provided NIST ASR transcripts.

While phoneme alignment is estimated using VTLN features, the MLLR and CMLLR transformation matrices are estimated using non-VTLN features.

4.8.2 SVM Training

We constructed one supervector for each utterance by vectorizing one CMLLR and two MLLR matrices. Then we applied rank normalization trained on the NIST SRE 2004 and 2005 telephone data.

After the normalization, a Nuisance Attribute Projection (NAP) was performed. We computed three projection matrices: first is trained on NIST SRE 2004 and 2005 telephone data, second on NIST SRE 2005 and 2006 microphone data and third is trained on NIST SRE 2008 interview data. For **tel-tel** system 20 dimensions from first and 10 dimensions from second matrix are used. For **int-tel** and **int-int** system, 20 dimensions from first, 10 dimensions from second and 10 dimensions from third matrix are used.

Background data sets (data for the negative class) are constructed separately for **tel-tel** system and for **int-tel** and **int-int** systems.

For the **tel-tel** system, first all NIST SRE 2004 and 2005 telephone are used. This set is then reduced to one third by selecting the segments which appeared most frequently as support vectors. To this reduced telephone set, the microphone data from **int-tel** and **int-int** reduced background set are added.

For the **int-tel** and **int-int** systems, first all NIST SRE 2005 and 2006 microphone data plus NIST SRE 2008 interview data were used. Then this set was reduced to one third in the same way as for the **tel-tel** system. To this reduced microphone and interview data, the whole reduced telephone set was added.

Finally SVM training was applied using Matlab interface for libsvm. A precomputed linear kernels were provided to the libsvm during the training and testing.

4.9 CRIM

This is another i-vector system, which uses a modified PLDA [7] model with heavy-tailed distributions instead of normal distributions. For details see [12] and the CRIM SRE 2010 System Description.

4.10 Agnitio Quality Measures

Quality measures are ancillary statistics computed from the input data, which on their own contain little or no speaker discriminating information, but which can aid calibration of discriminative scores. In particular, if the quality measures indicate that the input data is of poor quality, e.g. because of low SNR, then the output log-likelihood-ratio can be modulated by the quality measures to be close to zero. Agnitio computed two types of quality measure:

- A **segment-based** quality measure is a scalar value computed for each test segment and each train segment.
- A **trial-based** quality measure is a scalar value computed for every detection trial.

The Agnitio quality measures were computed only for telephone data.

4.10.1 Segment-based Quality Measures

1. UBM state visit statistic: $\frac{1}{2048}\sum_{i=1}^{2048}\frac{n_i}{n_i+1}$, where n_i is the expected number of visits to UBM state i in the speech segment.

2. The average gender-dependent i-vector extractor log-likelihood, where the gender agrees with the NIST gender label.

3. Gender mismatch: For segments labelled *male* by NIST, the average male i-vector extractor log-likelihood, minus the average female i-vector extractor log-likelihood. Conversely, for segments labelled *female*, the gender mismatch statistic is the female log-likelihood minus the male log-likelihood.

4. The average UBM frame log-likelihood. The UBM is gender-independent.

5. The average posterior entropy for the UBM state, given the acoustic frame feature vector.

6. The logarithm of the number of MFCC frames in the segment that passed the VAD.

4.10.2 Trial-based Quality measures

Both measures below were computed from the i-vector two-covariance model [2], under the assumption that the speaker in the train and test segment are the same.

1. Channel mismatch: The squared Mahalanobis distance between the channel variables for the two segments, where the distance is parametrized by the within-speaker covariance, $\mathbf{C}_W$. The formula is $\boldsymbol{\delta}'\mathbf{C}_W^{-1}\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is the difference between the two i-vectors of the trial.

2. Speaker rarity: The *expected* squared Mahalanobis distance between the speaker-identity variable and the origin, where the distance is parametrized by the between-speaker covariance. The expectation is over the posterior for the speaker identity variable.

4.11 BUT Quality Measures

The BUT quality measures are all segment-based:

1. Gender mismatch, similar to the Agnitio ones, but computed from MMI-trained GMMs as described below.

2. logarithm of number of frames

3. number of frames

4. SNR

5. Speech GMM log-likelihood

6. Speech+Silence GMM log-likelihood

More details are given below.

4.11.1 Gender identification likelihoods

Two likelihoods of two GMM models with 32 Gaussians. The first one was trained only on speech frames using BUT segmentation from female speakers. The second one was trained only on speech frames using BUT segmentation from male speakers. The models are trained using MMI criterion, and features are MFCC_0_D_A_Z. The training data for both GMMs are from NIST 2004 data.

4.11.2 GMM likelihoods

Two likelihoods of two GMM models with 256 Gaussians. The first one was trained on only speech frames using BUT segmentation. The second one was trained on whole files (speech and silence). The training data for both GMMs:

- 3000 randomly chosen files from MIXER 04,05,06 telephone data
- 2000 randomly chosen files from MIXER 05,06 telephone:mic data
- 1000 randomly chosen files from MIXER 08 interview data (heldout speakers defined by MIT).

The likelihoods are computed only for speech frames and normalized by the number of frames.

4.11.3 SNR estimation

This SNR estimator was implemented according the article [13].

5 Fusion and calibration

The fused and calibrated log-likelihood-ratio output for a trial with train segment i and test segment j is:

$$\ell_{ij} = f_{\epsilon,\delta}\big(\alpha + \sum_{k=1}^{N} \beta_k s_k(i,j) + \sum_{k=1}^{M} \gamma_k r_k(i,j) + \mathbf{q}(i)'\mathbf{W}\mathbf{q}(j)\big) \tag{1}$$

where $s_k(i,j)$ is the score of system k for the trial; $r_k(i,j)$ is the kth trial-based quality measure; $\mathbf{q}(i)$ and $\mathbf{q}(j)$ are vectors of segment-based quality measures, where the vector is augmented by appending 1. The fusion parameters are: offset α; linear combination weights β_k and γ_k; the bilinear combination matrix $\mathbf{W}$, constrained to be symmetric; and the saturation parameters δ and ϵ. The saturating non-linearity is defined as:

$$f_{\epsilon,\delta}(x) = \log \frac{\exp(x) + \exp(\delta)}{1 + \exp(\delta)} - \log \frac{\exp(x) + \exp(-\epsilon)}{1 + \exp(-\epsilon)} \tag{2}$$

The saturation parameters may be interpreted as the log-odds for mislabelling in the fusion training data, respectively a target as a non-target, or a non-target as a target. If $\delta, \epsilon \ll 0$, then f is an increasing sigmoid passing through the origin, with lower saturation approximately at δ and upper saturation approximately at $-\epsilon$. As the parameters approach 0 from below, the saturation levels approach 0 respectively from above and below. In our experiments, the optimizer typically set $\delta, \epsilon \approx -9$.

5.1 Training

The fusion parameters were trained separately on each of the three lists of supervised detection trials (tel-tel, int-tel and int-int), by minimizing w.r.t. the fusion parameters the following cross-entropy objective function:

$$
\begin{aligned}
\mathcal{C}_{\text{xe}}(P) = \frac{P}{T} \sum_{(i,j)\in\mathcal{T}} \log\bigl(1 + \exp(-\ell_{ij} - \operatorname{logit} P)\bigr) \\
+ \frac{1-P}{N} \sum_{(i,j)\in\mathcal{N}} \log\bigl(1 + \exp(\ell_{ij} + \operatorname{logit} P)\bigr)
\end{aligned}
\tag{3}
$$

where $P = 0.001$ to agree with the new DCF definition; $\mathcal{T}$ is the set of target trials of size T; and $\mathcal{N}$ the set of non-target trials of size N.

The optimization was done with the *Trust-Region-Newton-Conjugate-Gradient* [14, 15] optimization algorithm, which uses first and second order partial derivatives. The derivatives were calculated via algorithmic differentiation, which essentially computes derivatives of complicated functions by applying the chain rule to combine the derivatives of simple building blocks [16, 17].

Although we used (3) as training objective, we evaluated our development results by actual and minimum DCF. We found that in most cases, when the saturating non-linearity was used, we got a well-calibrated actual DCF close to minimum DCF. If we omitted the non-linearity and used a plain affine fusion transform, the actual DCF was often much higher. We conjecture that the non-linearity compensates both for labelling errors and data scarcity at the challenging new DCF.

5.2 Scores and decisions

The fusion output scores are intended to act as well-calibrated *log-likelihood-ratios*. These scores are compared to the Bayes threshold of $-\operatorname{logit}(0.001) \approx 6.9$ to make hard decisions.

Although our scores can be evaluated as log-likelihood-ratios via C_{llr} as defined in the evaluation plan, we optimized our scores to minimize the closely related $\mathcal{C}_{\text{xe}}(0.001)$. Note that $C_{llr} = \mathcal{C}_{\text{xe}}(0.5)$.

6 Team members

Agnitio Niko Brümmer, Luis Buera and Edward de Villiers.

BUT Doris Baum, Lukáš Burget, Ondřej Glembek, Martin Karafiát, Marcel Kockmann, Pavel Matějka and Oldřich Plchot.

CRIM Patrick Kenny, Pierre Ouellet and Mohammed Senoussaoui.

References

[1] George R. Doddington, "Speaker recognition evaluation methodology: a review and perspective," in *Proceedings of RLA2C*, Avignon, France, Apr. 1998, pp. 60–66. 3

[2] Niko Brümmer and Edward de Villiers, "The speaker partitioning problem," in *Proceedings of Speaker Odyssey 2010*, Brno, Czech Republic, June 2010. 4, 9, 16

[3] Pavel Matějka, Lukáš Burget, Petr Schwarz, and Jan "Honza" Černocký, "Brno university of technology system for nist 2005 language recognition evaluation," in *Proceedings of the IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006, pp. 57–64. 5, 7, 10

[4] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, July 2008. 6, 8, 13

[5] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech 2009*, Brighton, UK, Sept. 2009, pp. 1559–1562. 6, 9, 11

[6] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," submitted to IEEE Transactions on Audio, Speech and Language Processing, 2010. 9

[7] Simon J. D. Prince and James H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007. 12, 15

[8] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007. 12

[9] Niko Brümmer, "The em algorithm and minimum divergence," Agnitio Labs Technical Report. Online: `http://niko.brummer.googlepages.com/EMandMINDIV.pdf`, Oct. 2009. 12

[10] Marcel Kockmann, Lukáš Burget, and Jan "Honza" Černocký, "Investigations into prosodic syllable contour features for speaker recognition," in *Proceedings of ICASSP*, Dallas, United States, Mar. 2010, pp. 1–4. 13

[11] Ondřej Glembek, Lukáš Burget, Najim Dehak, Niko Brümmer, and Patrick Kenny, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," in *Proceedings of ICASSP*, Taipei, Taiwan, Apr. 2009. 13

[12] Patrick Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proceedings of Speaker Odyssey 2010*, Brno, Czech Republic, June 2010. 15

[13] Chanwoo Kim and Richard M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proceedings of Interspeech 2008*, Brisbane, Australia, Sept. 2008. 17

[14] Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, Springer, 2006. 18

[15] Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi, "Trust region newton method for large-scale logistic regression," *Journal of Machine Learning Research*, Sept. 2008. 18

[16] Niko Brümmer, "Notes on computation of first and second order partial derivatives for optimization," Agnitio Labs Technical Report. Online: `http://niko.brummer.googlepages.com/2diff_chain_rules.pdf`, Apr. 2009. 18

[17] Jan R. Magnus and Heinz Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, 1999. 18