Babelon

Monthly Report

Period of Performance: 5 March 2012 – 4 July 2013 (Base Period)

Organizations:	BBN Technologies (BBN) Brno University of Technology (BUT) Johns Hopkins University (JHU) Vocapia Research (LIMSI) Massachusetts Institute of Technologies (MIT) North-West University (NWU)	
Principal Investigators:	Dr. John Makhoul Tel: 617-873-3332 Fax: 617-873-2473 Email: makhoul@bbn.com	Dr. Stavros Tsakalidis Tel: 617-873-4976 Fax: 617-873-2473 Email: stavros@bbn.com
Reporting Period:	1 December – 31 December 2012	

1 Highlights/Significant Technical Achievements This Period

BBN Technologies (BBN)

- Bug fix in Tagalog keyword search module for the PI-KWS condition. After the bug fix the ATWV performance improved about 20% relative for both FullLP (0.455 vs. 0.382) and LimitedLP conditions (0.169 vs. 0.143).
- LM training from targeted web text collection. The ATWV for the Tagalog LimitedLP condition improved by 2 points.
- Updated system combination procedure.
- Fully integrated region-dependent transforms (RDT) into the training procedure.
- Using mapping to the "probability of correct" in score normalization.
- Using Powell's algorithm for tuning the normalization.

Brno University of Technology (BUT)

- Trained RDLT system for Tagalog and Pashto, reduction of parameters.
- Compare different NN configurations for Tagalog and Pashto.
- Compare feature normalizations and f0 extraction for Cantonese.
- Improving Kaldi KWS system.

Johns Hopkins University (JHU)

- Delivered bottleneck features for Tagalog data (FullLP) to BBN.
- Started neural network training procedure on Pashto data.

Vocapia Research (LIMSI)

- Work continues on Cantonese, Tagalog, Pashto and Turkish STT systems.
- Trained MLP with wLP raw features for Cantonese.
- Tested MLP features shared by BUT in our Cantonese and Tagalog STT systems.
- Test BBN VAD in our Tagalog STT system.
- KWS with graphemic STT system for Tagalog.

Massachusetts Institute of Technologies (MIT)

- We continue to develop DBN-based ASR for low-resource languages. We have also continued our research on completely unsupervised methods for spoken term detection.

North-West University (NWU)

- None.

2 Other Actions Worked on This Period

BBN Technologies (BBN)

- None

Brno University of Technology (BUT)

- None

Johns Hopkins University (JHU)

- Generated FDLP based feature representations for Pashto data.

Vocapia Research (LIMSI)

- Automatic pronunciation discovery using the segmental DTW algorithm was not very successful, probably due to the speaker variation and noise in the data. We have since trained a grapheme-based HMM system for Cantonese. By clustering the trained Gaussians, posteriorgrams can be produced for the data. We are currently investigating methods to learn pronunciations from the posteriorgrams.

Massachusetts Institute of Technologies (MIT)

- We have been exploring the use of unsupervised DTW re-scoring as a means of re-ranking spoken term hypotheses. We perform pair-wise DTW matching of all hypothesized spoken terms using an unsupervised posteriorgram-based representation. We then compute a lower-dimension projection on the resulting similarity matrix as a means of clustering highly similar keyword hypotheses. One potential application of this projection is as an additional feature for score normalization.

North-West University (NWU)

- (Ongoing) Multilingual pronunciation variant modeling: tool development and evaluation.
- Development of improved phonemic and graphemic lexicons for Pashto ASR systems
- Developing insight into requirements for algorithms for induction of *additional* pronunciations when phoneme accuracies are low (in contrast, existing algorithms are adequate for evaluation of *existing* pronunciations.

3 Concise Description of Each of the Experiments Being Done on the Language and its Outcome

BBN Technologies (BBN)

- LM training from targeted web text collection

  During the last reporting period, Microsoft Bing web search was performed to collect web text data for Tagalog keywords. During this reporting period, the collected web text was used to augment the LM. The web corpus consists of text extracted from web pages downloaded for the URLs produced by the Microsoft Bing search on the keywords. Table

1 below summarizes the n-gram statistics for the web text corpus and also the subset of n-grams that contain at least one keyword. It also compares the n-gram statistics for the baseline Language Model (LM) trained on the acoustic training text for the LimitedLP condition. The decoding lexicon for the LimitedLP language pack is approximately 6k words.

Corpus		Unigrams	Bigrams	Trigrams
All	Acoustic training	6223	33172	57104
	Targeted Web text	6223	3,819,348	15,433,721
N-grams with keywords	Acoustic training	233	4243	5808
	Targeted Web text	344	714,844	2,235,550

Table 1: Ngram statistics for the LM text collections. The acoustic training corresponds to the Tagalog LimitedLP.

Next, we employed two different procedures to train LMs using the web-data:

1. Train a LM from the webdata and interpolate it with the main (baseline) LM that is trained from the acoustic transcripts. We used a weight of 0.7 for the baseline LM and 0.3 for the LM trained over the web-data.
2. In this experiment we train a LM using only N-grams that contain at least one keyword. Then, we interpolate the LM with the baseline LM with the same interpolation weights as in 1.

Table 2 below summarizes the STT and KWS performance for the Tagalog LimitedLP AA-KWS condition using the updated LMs. We observed MTWV gains for both procedures. We also compared the two LMs in terms of area under the curve (AUC). The results show a gain in AUC by training a LM only from N-grams that contain keywords. Next, we will investigate several weighting schemes for the n-grams that contain the keywords.

LM training	WER	ATWV	MTWV	AUC
acoustic	69.20	0.190	0.198	0.619
acoustic + web	68.96	0.201	0.213	0.623
acoustic + keyword n-grams from web	69.76	0.202	0.209	0.653

Table 2: STT and KWS performance for the Tagalog LimiteLP AA-KWS condition using targeted web-data

- Fully integrated region-dependent transforms (RDT) into the training procedure

  The region-dependent transform (RDT) is a feature extraction method for speech recognition that employs the Minimum Phoneme Error (MPE) criterion to optimize a set of feature transforms, each concentrating on a region of the acoustic space. The RDT procedure is used by the ASR training module to add or fuse various features (e.g. PLP, MLP) including external features provided by our partners (JHU, BUT).

  In this period, we fully integrated RDT into the training sequence. We also, added for robustness a stopping criterion into the estimation procedure that depends on a heldout set. The training iteration with best error rate on the heldout set is chosen.

- Updated system combination procedure

  In this period we updated the system combination procedure. The baseline system combination approach merges overlapping hits from multiple systems. The score of the merged hit was obtained via a liner-interpolation of the individual scores with pre-defined weights. The weights were optimized on a heldout set via a grid search.

  The output of the new combination algorithm is a common set of detections (hits), each with a vector of scores assigned to it, with scores from the multiple systems. These vectors are used as features in the machine learner of the score normalization. With this procedure we got a 3% absolute gain by combining the best Cantonese FullLP BBN output for the AA condition with the best BUT output (see Table 3).

System	**Unnormalized**		**Normalized**	
	ATWV	MTWV	ATWV	MTWV
BBN	0.463	0.475	0.557	0.564
BUT	0.514	0.524	0.567	0.570
System Combination	0.541	0.548	0.598	0.605

Table 3: System combination results for Cantonese FullLP AA-KWS condition.

- Using mapping to the "probability of correct" in score normalization

  The scores of the hits are sorted and binned so that a constant number of true hits (e.g., 10) appears in each bin. This means that bins with lower scores tend to be larger than bins with higher scores, as expected. This binning can then be used to learn a mapping from scores to the "probability of correct" (P(corr)), which is defined as the ratio of the number of true hits and the total number of hits in the bin. When using this mapping on the raw posteriors of the recognizer, there seems to be a small but consistent gain as reported below in Table 4.

Language	Probability of correct	Normalized	
		ATWV	ATWV
Cantonese	No	0.527	0.535
	Yes	0.530	0.536
Tagalog	No	0.448	0.459
	Yes	0.453	0.460
Pashto	No	0.408	0.412
	Yes	0.411	0.416

Table 4: Effect of using the "probability of correct" mapping in score normalization.

- Using Powell's algorithm for tuning the normalization

  We wrote a C++ version of the optimizer that has MTWV as the objective function. The starting point for the optimization is the weights obtained from the log-linear trainer. This seems to perform well on the tuning data, with gains of the order of 1-2 MTWV points in most cases, but seems to overfit. As can be seen from the figures in the Appendix (Figure 4 (MTWV split 1) and Figure 4 (MTWV split 2) the DET curve (red line) on the tuning data is better than the baseline (black line) only close to the operating point, but worse elsewhere. When we changed the optimizer to tune to AUC instead, the resulting DET curve became better (Figure 4 AUC split 1 and Figure 4 AUC split 2) but did not result in significant gains.

Brno University of Technology (BUT)

- ASR - Cantonese release B:
  - STK/HTK – Comparison of f0 and log f0 features using PLP/VTLN features, normalized using VAD, and speaker normalized f0, deltas/double deltas, training a simple ML system:
  - Heldout CER [%] - decoding using Appen segmentation:

    plp vtln VAD(CMN/CVN) 62.8%
    plp vtln VAD(CMN/CVN) + f0log_SpkrNorm 62.0%
    plp vtln VAD (CMN/CVN)+ f0log_SpkrNorm (then CMN/CVN) 62.7%
    plp vtln VAD (CMN/CVN)+ f0_SpkrNorm 58.6%
    plp vtln VAD (CMN/CVN)+ f0_SpkrNorm (then CMN/CVN) 59.0%

  - Plain f0 is giving significantly better results that log(f0)

- CMN/CVN on normalized f0 features (divided by speaker average f0) is hurting - so it makes sense to switch off CMN/CVN on PLP+NN+F0 feature streams.
- Current experiments shows that CMN/CVN is not beneficial anyway, effect CMN/CVN for RDT need to be investigated.

- ASR - Pashto release B:
  - Experiments with different NNs (trained ML system, scoring on devB, WER):

PLP VTLN HLDA	65.7%
UC bottleneck 6M params 4 layers	57.7%
UC bottleneck 2M params 4 layers	58.3%
UC bottleneck 2M params 3 layers	59.1%
UC bottleneck 2M params 4 layers (+f0, Red.Sil. , linear BN)	56.4%

  - 4 layer bottleneck NN structure gives 0.4% benefit in comparison to 3 layers with the same number of parameters.
  - Silence reduction + f0 + using linear bottle neck layer instead of sigmoid layer on both NN gives best results (second NN has linear layer anyway, and linear BN was standard configuration in previous experiments – see Tagalog).
  - RDLT training of PLP+NN+f0 features (WER on devB):

ML transform, reduce dim 72 dim to 69	57.4%
RDLT, reduce dim. 72 to 69, 1000 GMMs, context 7	54.4%
RDLT, reduce dim. 72 to 69, 125 GMMs, context 7	55.5%

  - 7*1000 transforms work significantly better that 7*125 - maybe it partly solves problems with dialect variations?
- ASR - Tagalog release B:
  - Experiments with different NNs (trained ML system, scoring on devB, WER):

PLP VTLN HLDA	67.0%
UC bottleneck 6M params 4 layers	58.9%
UC bottleneck 2M params 4 layers	59.5%
UC bottleneck 2M params 3 layers	60.8%
UC bottleneck 2M params 4 layers (reduced Sil)	59.3%
UC bottleneck 2M params 4 layers (linear bottleneck)	58.0%

  - Layer bottleneck NN structure gives 1.3% benefit in comparison to 3 layer NN with the same number of parameters.
  - Silence Reduction in NN training gives 0.2%.
  - Linear bottle neck layer instead of sigmoid layer on both NN. (second NN has linear layer anyway) gives 1.5% (used to be the standard configuration in previous experiments anyway).
  - RDLT training of PLP+NN+f0 features (WER on devB):

ML transform	58.6%
RDLT, reduce dim 72 to 69, 1000 GMMs, context 7	55.6%
RDLT, reduce dim 72 to 69, 125 GMMs, context 7	55.3%

  - 7*1000 transforms work slightly worse than the compact 7*125 system, as opposed to Pashto.

- ASR – Turkish release B: prepared systems for training Keyword Spotting on Cantonese using STK:
  - KWS on Cantonese using Kaldi: improved post-processing.
  - Term vocabulary is NOT in LVCSR - PI condition.
  - Currently best ASR system (PLP+NN+f0 RDT).
  - Based on indexing character lattices - paper Dogan Can.

	Unnorm.		After norm.	
	ATWV	MTWV	ATWV	MTWV
AVERAGE:	0.4518	0.4645	0.5159	0.5211 (smaller lattices)
AVERAGE:	0.4490	0.4586	0.5009	0.5082 (bigger lattices)

Johns Hopkins University (JHU)

- Processed conversational and scripted releases of Tagalog data. FDLP-S and FDLP-M features are generated on the data. Neural networks are trained and bottleneck features are extracted. These features are delivered to BBN, to integrate with Byblos based KWS system.
- Processed Pashto data and FDLP-S and FDLP-M features are derived. Neural networks are trained in discriminative fashion on this data.

Vocapia Research (LIMSI)

- **Cantonese STT**
  - Trained new MLP with wLP raw features (previous MLP used TRAP-DCT features).
  - Investigated using BUT features (MLP alone or MLP concatenated with PLP+f0).
  - Gender-independent and gender- dependent (MAP adaptation) without or with discriminative training (MMIE).
  - Tested use of BBN automatic segmentation (VAD) for training data.
  - Results CER on heldout data (BBN VAD not tested yet):

Type	# Features	GI	GD	GD MMIE
TD mlpplpf0	81	54.0	53.3	52.7
wLP mlpplpf0	81	53.5	52.3	-
BUT mlp+LIM plpf0	72	52.1	51.8	
BUT mlpplpf0	72	48.4	49.6	47.9

- Initial-Final system with BUT features:

Features	Phone	IF	ROVER
BUT mlpplpf0	44.5	44.6	43.9

- **Pashto STT:**
  - Phonemic and graphemic acoustic models built on release B data.
  - WER on release B dev: phonemic: 67.1%, graphemic 68.5%.
    - lower WER (about 1%) with multilingual MLP than with MLP trained on Arabic data.
    - plpf0 features better than plp alone (1%).
- **Tagalog STT:**
  - Investigating using BUT features and BBN VAD.
  - 1-pass phone-based system:

Partitioner	Features	val / tune / all
LIM	LIM	57.8 / 61.4 / 60.3
LIM	BUT	53.4 / 56.3 / 55.5
LIM	BUT	52.5 / 55.5 / 54.6 tuned
BBN	BUT	50.7/ 53.5 / 52.7 tuned

  - 1-pass graphemic system:

Partitioner	Features	val / tune / all
LIM	LIM	59.8 / 62.3 / 61.6
LIM	BUT	54.3 / 57.7 / 56.7
BBN	BUT	51.4 / 54.3 / 53.4

  - Rover of best phone system and graphemic systems: 51.9% on all dev.
  - Investigated 2-pass systems using BBN partitioner and BUT features.
  - WER on all dev:

- (1) phone → phone: 52.4%
- (2) phone → graphemic: 52.6%
- (3) graphemic → phone 52.1%
- Rover (2,3): 51.2 %

- **Tagalog KWS:**
  - KWS on the graphemic STT system with BUT features (WER=56.7%).
  - Internal RTTM file (the official RTTM not tested yet).
  - Pre-Indexed (PI) KWS system: MTWV = 0.3365 without score normalization.

Massachusetts Institute of Technologies (MIT)

- Evaluated unsupervised keyword search on Cantonese data using DBN-based posteriorgram representation. 606 keywords from the dry-run keyword list that existed in the training set were used for utterance-based keyword detection on the dev set. A MTWV of 0.206 was achieved for these keywords (a subset of the entire dry run keyword list).

North-West University (NWU)

- For induction of grapheme-based and phoneme-based lexicons in Pashto: "phone accuracies" (i.e. accuracies of phonemic or graphemic sub-word units) below 35% on development set (32.9% for graphemic system and 32.3% for phonemic system, corresponding to WERs of 59.1% and 62.1% for the graphemic and phonemic systems, respectively.)

4 Progress on Languages (report waypoints when possible)

BBN Technologies (BBN)

- Bug fix in Tagalog keyword search module

  We discovered a bug in the Tagalog keyword search module for the PI-KWS condition. After the bug fix the ATWV performance improved about 20% relative for both FullLP (0.455 vs. 0.382) and LimitedLP conditions (0.169 vs. 0.143).

- Comparison between the current score normalization system and the one from the dryrun

  Table 5 compares the current score normalization system to the one used in the dryrun (Sept 2012) for the FullLP AA-KWS condition.

Language	Score Normalization	Normalized	
		ATWV	ATWV
Cantonese	Dryrun (Sept 2012)	0.527	0.537
	Current (Dec 2012)	0.560	0.564
Tagalog	Dryrun (Sept 2012)	0.457	0.465
	Current (Dec 2012)	0.466	0.473
Pashto	Dryrun (Sept 2012)	0.399	0.408
	Current (Dec 2012)	0.420	0.425

Table 5: Comparison of score normalization systems.

Brno University of Technology (BUT)

- ASR recognizer for Tagalog
  - PLP+NN+f0, RDLT system: 55.3% WER (full dev set)
- ASR recognizer for Pashto
  - PLP+NN+f0 system: 54.4% WER (full dev set)

Johns Hopkins University (JHU)

- None

Vocapia Research (LIMSI)

- **Cantonese**
  - CER on held-out data: 43.9%
- **Turkish**
  - WER on release A data: 62.0% (phone-based system)
  - 65.0% (graphemic system)
- **Pashto**
  - WER on release B data: 67.1%
- **Tagalog**
  - WER on release B data: 51.2%

Massachusetts Institute of Technologies (MIT)

- We continue to develop acoustic models based on Deep Belief Networks (DBNs) and have been evaluating them on the Cantonese language. Our experiments on Cantonese

use the lexicon and transcripts provided by Appen from which we generate a trigram language model. The feature representation consisted of the 72 dimensional features shared by BUT. A 5-layer DBN containing 2048 hidden units/layer was trained on 52 hours of conversational speech from the Cantonese B data. Training consists of an unsupervised pre-training stage followed by a supervised back-propagation stage using frame-based state IDs generated from forced path alignments that were generated by a tied-state CD-HMM-GMM model. The resulting model achieves a 53.7% CER on the held-out set.

North-West University (NWU)

- Not measurable in this period.

5 Issues/Problems and Proposed Solutions (potential work plan risks and course corrections)

BBN Technologies (BBN)

- None.

Brno University of Technology (BUT)

- Good KWS results for Cantonese should be ported to all other Babel languages.

Johns Hopkins University (JHU)

- None.

Vocapia Research (LIMSI)

- None.

Massachusetts Institute of Technologies (MIT)

- None.

North-West University (NWU)

- None.

6 Babel-related Scholarly Activities including papers published and presentations given during this period (provide full reference for the material, and upload the item to SharePoint at least 2 weeks prior to publication or presentation; don't forget the acknowledgement)

BBN Technologies (BBN)

- None.

Brno University of Technology (BUT)

- None.

Johns Hopkins University (JHU)

- Samuel Thomas, Michael L. Seltzer, Kenneth Church and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition", submitted to ICASSP 2013.
- Pascal Clark, Sri Harish Mallidi, Aren Jansen, and Hynek Hermansky, "Frequency offset correction in speech without detecting pitch", submitted to ICASSP 2013.
- Vijayaditya Peddinti and Hynek Hermansky, "Filter-bank optimization for frequency domain linear prediction", submitted to ICASSP 2013.
- Tetsuji Ogawa, Feipeng Li, Hynek Hermansky, "Stream selection and integration in multi-stream asr using gmm-based performance monitoring", submitted to ICASSP 2013.
- Feipeng Li, Hynek Hermansky, "Effect of filter bandwidth and spectral sampling rate of analysis filterbank on automatic phoneme recognition", submitted to ICASSP 2013.
- Ehsan Variani, Feipeng Li, Hynek Hermansky, "Multi-stream recognition of noisy speech with performance monitoring", submitted to ICASSP 2013.
- Hynek Hermansky, Ehsan Variani and Vijayaditya Peddinti, "Mean temporal distance: predicting asr error from temporal properties of speech signal", submitted to ICASSP 2013.

Vocapia Research (LIMSI)

- N/A.

Massachusetts Institute of Technologies (MIT)

- Presented a research update at the site visit on December 6.

North-West University (NWU)

- None.

7 Significant Anticipated Activities and Opportunities Next Period (next month)

BBN Technologies (BBN)

- Train Turkish system
- Continue research on LM training from targeted web-data
- LM inference for rare words
- Use Powell's algorithm for estimating the optimal weights used in system combination

Brno University of Technology (BUT)

- Train Turkish release B ASR, including VAD, NN, RDLT.
- Building of 10h systems (limited LP condition).
- Training NN networks with more parameters, silence reduction and f0 for Tagalog, Pashtu and also Turkish.
- Building of new ASR system on top of new NN based features.
- Building KWS on Pashto, Tagalog and Turkish.
- Prepare Kaldi based KWS in automatically adapted condition.
- Fusion of Kaldi and STK based KWS systems.
- Sub-word and hybrid word-subword KWS system for other languages.
- Experiments with multilingual neural nets.
- Trying different adaptation techniques, adaptation to dialects.

Johns Hopkins University (JHU)

- Deliver Pashto bottleneck features to BBN.
- Generate Tagalog and Pashto features for BabelLR and OtherLR conditions.

Vocapia Research (LIMSI)

- Work will continue on all transcription systems.
- Work will begin on building NNLM for Tagalog.
- Work will continue on updating the Cantonese KWS system with recent STT system outputs.
- Work will begin on building the KWS system for Pashto.

Massachusetts Institute of Technologies (MIT)

- We continue to develop ASR capability for Pashto, and Turkish. We plan to explore the pronunciation mixture model mechanism on these languages, in addition to continuing our research on DBN-based ASR for these languages.

North-West University (NWU)

- Improvements to grapheme-based systems: extend “lexicon” based on sub-word recognition results (Turkish & Pashto) and treat exceptional entries (loan words, spelled words etc.) separately.
- Improvements to phoneme-based systems: refined lexicons (Turkish & Pashto).
- Dialect analysis of comparative Pashto recognition accuracies: grapheme vs. phoneme.

8 Appendix

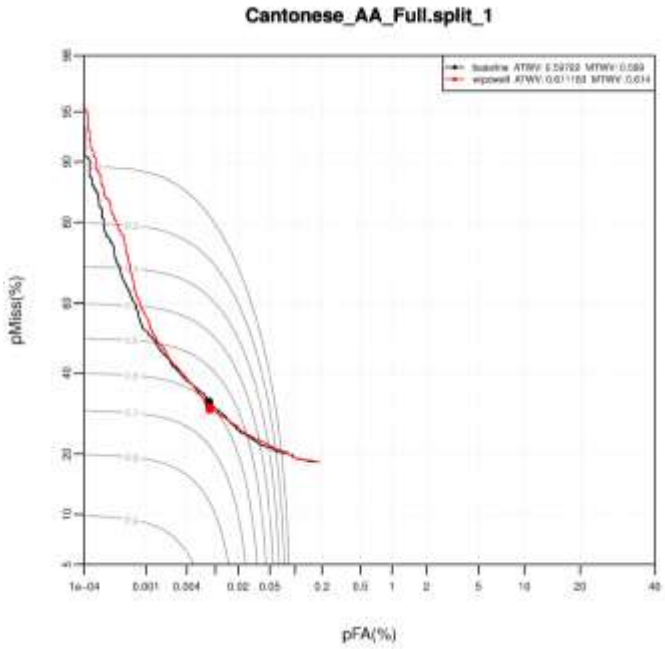

Figure 4: Optimizing for MTWV (split 1)

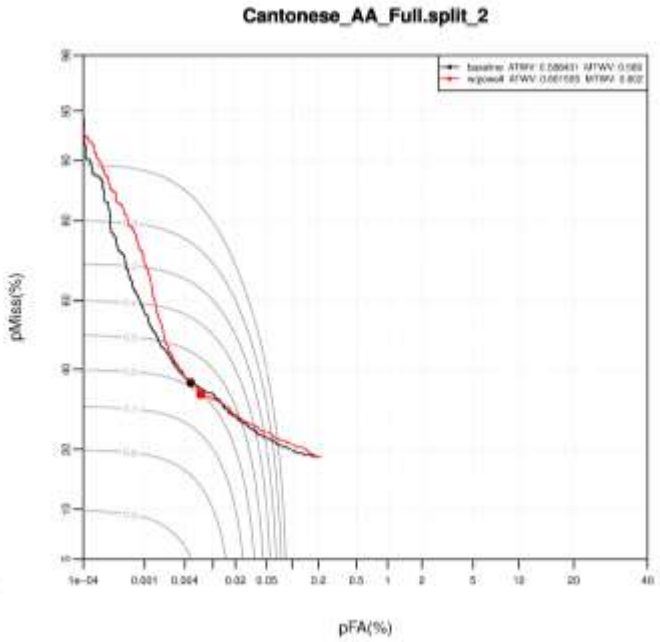

Figure 4: Optimizing for MTWV (split 2)

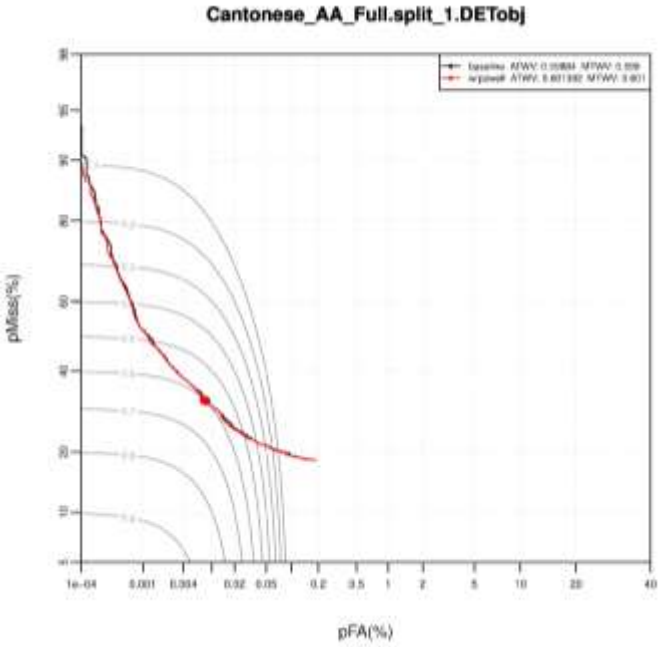

Figure 4: Optimizing for AUC (split 1)

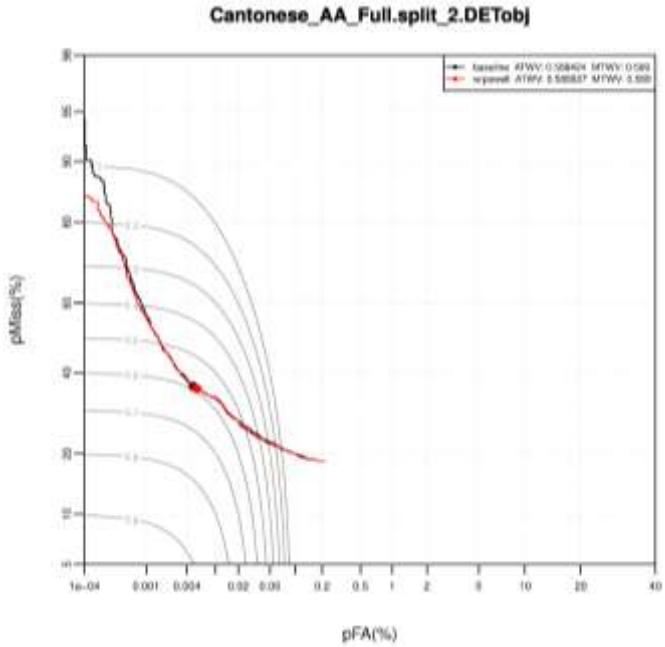

Figure 4: Optimizing for AUC (split 2)