Summary report for project

RATS - Robust Automatic Transcription of Speech

For year 2017

Submitted to Raytheon BBN Technologies

By Brno University of Technology

Lead author: Dr. Pavel Matějka

TECHNICAL ACCOMPLISHMENTS

Language identification

Throughout the RATS project, BUT developed many LID subsystems based on either acoustic features and ivectors or different variants of phonotactic systems. BUT also designed the calibration and fusion scheme which was used during all submissions of the Patrol team and also for NIST Language Recognition Evaluations. The great part of time was devoted to the engineering of the development datasets, which were used by the whole Patrol team.

Previously, we reported a success with using Neural Networks (NN) as a classifier on top of ivectors. Nowadays NNs can be found in every stage of a LID system. At the lowest level, DNNs are successfully used to extract bottleneck features (BN), in later stages, we can use them as a classifier or ultimately, one can build an End-to-End system, effectively combining the two cases into one big NN.

In the last two years of the project we investigated the bottleneck (BN) features which are extracted from a low-dimensional (30-80) layer of NN. These features convey information about phonetic content in a nonlinearly compressed form which can be directly used as (instead of) conventional acoustic features. Despite excellent results, these features exhibit strong coupling to the language used during the NN training. It can be shown in our analysis that this can be circumvented by means of multilingual training of such BN. The term multilingual means that the NN is trained on several languages simultaneously.

We have also participated in NIST Language Recognition Evaluation in 2015 where we built over 20 systems, but only 6 of them were selected for the final fusion and submission. The single best system was based on multilingual BNs and reached 50% relative improvement over the conventional acoustic features (MFCC+SDC).

Speaker identification

During the RATS project, BUT concentrated mainly on extending the state-of-the art modeling techniques for the SID (ivectors + PLDA) as well as on the development of acoustic features. Similarly to LID, BUT designed the calibration and fusion scheme which was used

during all submissions of the Patrol team for the RATS evaluations. Substantial part of the time was devoted to cooperation with BBN on the engineering and verification of the development datasets, which were used by the whole Patrol team.

NIST has run another Speaker Recognition Evaluation in 2016 which brought a completely new non-English dataset, short durations and mismatched channels; all of these generating a tough challenge in the domain adaptation. It revealed the weak side of current BN features that are tuned for English, brought back the issue of score normalization and in general significantly increased the difficulty which will undoubtedly inspire a lot of research. Adaptive score normalization was the crucial step for the best performance. Post analysis experiments revealed that the cohort for the score normalization should be pooled across several languages and channels from various sources, and an adaptive process should select the top X closest files (where $X=200$) to the final normalization cohort. Adaptive symmetric normalization (s-norm) performed the best.

Apart from speaker modeling, we were also working on neural networks for audio denoising. We trained the NN to directly map the noisy spectra to its clean representation. This approach allows for reconstructing the clean audio and therefore it can be applied as a pre-processing step for any other method. We have run extensive analysis of the training NN with different datasets, different levels of reverberation and noise levels and evaluated the results on the difficult speaker recognition conditions.

Lastly our focus is also on the End-to-End speaker recognition, where we decided to divide the system into several independent NNs that we can pre-train in order to obtain a good initialization. We built a NN that maps features to GMM statistics, another NN mapping these statistics to ivectors and finally a NN which maps statistics to scores. We were able to successfully train and connect these parts with a little loss in performance. Now, we are already in a position to jointly train all of the parts which together form an End-to-End speaker recognition system.

IMPLICATIONS FOR FUTURE RESEARCH

NIST Speaker Recognition Evaluation in 2016 revealed that current techniques are still not robust to data unseen during training and that domain adaption and score normalization are essential steps for a well performing system. There are also efforts in the community to work with the matched data without any labels – unsupervised adaptation/calibration.

NIST Language Recognition Evaluation 2015 revealed also that the current systems are not robust against channel. There were 2 dialects of French in the evaluation, both recorded with distinct and separate channels, but the channels were switched in the training and test data which caused complete failure of all systems.

End-to-End speaker recognition system are nowadays in focus of many research groups and results obtained on large (but not public) datasets are encouraging. This line of research in speaker ID is in its beginning; we expect to see more publications in the near future. As we already know from LID, this approach is data hungry and right now it is a challenging task to train an End-to-End SRE system from publicly available data.

Diarization is a crucial part of SRE but has received very little attention in the past. Current assumptions are to have one speaker in the audio which is not always true – for the real scenarios and also for the work with unsupervised/real/publicly available data, diarziation

front-end is needed and its poor performance can completely destroy the SRE system. We have explored diarization in our submission to SITW challenge with very positive results and further work will be done at SCALE/JHU summer workshop in 2017 with 5 BUT representatives. Besides diarization, the objectives of the workshop include also robustness, multilinguality, and adaptation to unseen data.

Long-term research opportunities

As the performance of SRE system for close-talk speech improves, it is time to suggest future research directions. Our ideas include the following:

Far-field speech and microphone arrays – the state of the art should be advanced in the following scenarios:

- „close-talk“ microphones in strongly noisy and reverberant environments - for example cell phone calls in a football stadium or music pub environment.
- „single distant microphone“ - a typical setup for covert listening, with microphones of varying quality and directivity - from cell phones to specialized directional microphones.
- „multiple distant microphones“ with fixed configuration, for example for conference rooms or large public spaces such as airport or railway station halls.
- „multiple distant microphones“ with fixed, but re-configurable positions typical for wall mounted wire-tapping devices or vehicle-mounting.
- “multiple distant microphones with dynamically changing configuration”, for example microphones worn by several agents tracking a target person in a noisy environment.

Tight cooperation will be required of signal capturing, signal processing (enhancement, beam-forming) and machine learning to come up with highly versatile solutions based on common mathematical background and ideally trained as a whole to maximize the target metric.

Link analysis (data-analysis technique used to evaluate relationships/connections between nodes) has long been used for both intelligence and investigation work. The situation can be compared to the early days of Internet search - Altavista, Excite and others had some results and market uptake, but the whole domain changed when Google started to exploit the relations between web-pages (TF/IDF metrics, PageRank, etc.). When using link analysis together with SRE, we expect similar breakthrough. We propose to combine them in the following way:

- *Massive use of conversational nature of speech data* - in case we know that A speaks often to B, then detecting A on one side of the call will automatically increase the prior probability of B even if the acoustic evidence is not reliable (due for example to illness, channel change or noise).
- *Use call content*. Standard i-vector based speaker identification ignores the content of the call, while a simple sentence “Peter speaking” heard on two different calls can completely change the game. It is not necessary to develop perfect ASR engines for all possible languages but commercially available ones can be deployed. For languages with missing ASR, language-independent techniques such as universal phoneme sets or automatically determined acoustic units (AUD) can be used.
- *Meta-information* is crucial for link analysis. Some of it is available (phone and IMEI numbers, geographical information, time-stamps) but the targets are aware such information is collected and have developed ways to falsify or obscure it. Significant amount of meta-information can however be automatically extracted from the speech

signal – for example, automatic detection of age, gender and accent of call participants. Another interesting meta-information is the *environment* – even if the speaker changes his cell phone number every day, he is not likely to change his favorite car, detecting that a call took place in given car can therefore help.

- By *time-relation analysis*, a classical problem of speaker recognition (speaker speaking very little in a call) can be turned into advantage, as this speaker can simply be identified by the fact that he is speaking little. Hierarchy and trust can be also partially inferred from this analysis.

Finally, SRE techniques can be used in **multi-modal distant person characterization**. Fusing several modalities is the only way to provide for reliable results when one or several modalities are missing or compromised (for example, a beard or sunglasses hamper face recognition, if the target is not speaking, SRE is unusable, etc). Techniques such as face and speaker recognition should be complemented by less traditional ones such movement and gesture analysis, infra-red sensing, walking pattern and general movement analysis etc. Modalities including remote recognition of heart-beat, breath, sweating, etc seem science-fiction nowadays, but with the progress in sensor technology, they might well be useable in near future. It is likely that the available machinery (i-vectors, NNs, calibration, fusion) used nowadays for SRE will find application also for these new modalities.

LIST OF PUBLICATIOS FUNDED UNDER RATS

- FÉR Radek, MATĚJKA Pavel, GRÉZL František, PLCHOT Oldřich and ČERNOCKÝ Jan. Multilingual Bottleneck Features for Language Recognition. In: *Proceedings of Interspeech 2015*. Dresden: International Speech Communication Association, 2015, pp. 389-393. ISBN 978-1-5108-1790-6. ISSN 1990-9772.

- BRUMMER Niko, SWART Albert du Preez, PRIETO Jesús J., GARCIA Perera Leibny Paola, MATĚJKA Pavel, PLCHOT Oldřich, DIEZ Sánchez Mireia, SILNOVA Anna, JIANG Xiaowei, NOVOTNÝ Ondřej, ROHDIN Johan A., GLEMBEK Ondřej, GRÉZL František, BURGET Lukáš, ONDEL Lucas, PEŠÁN Jan, ČERNOCKÝ Jan, KENNY Patrick, ALAM Jahangir, BHATTACHARYA Gautam and ZEINALI Hossein et al. *ABC NIST SRE 2016 SYSTEM DESCRIPTION*. San Diego: National Institute of Standards and Technology, 2016. http://www.fit.vutbr.cz/~matejkap/pubs.php?id=11372

- MATĚJKA Pavel, GLEMBEK Ondřej, NOVOTNÝ Ondřej, PLCHOT Oldřich, GRÉZL František, BURGET Lukáš and ČERNOCKÝ Jan. Analysis Of DNN Approaches To Speaker Identification. In: *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), 2016*. Shanghai: IEEE Signal Processing Society, 2016, pp. 5100-5104. ISBN 978-1-4799-9988-0.

- NOVOTNÝ Ondřej, MATĚJKA Pavel, GLEMBEK Ondřej, PLCHOT Oldřich, GRÉZL František, BURGET Lukáš and ČERNOCKÝ Jan. Analysis of the DNN-Based SRE Systems in Multi-language Conditions. In: *Proceedings of SLT 2016*. San Diego: IEEE Signal Processing Society, 2016, pp. 199-204. ISBN 978-1-5090-4903-5.

- NOVOTNÝ Ondřej, MATĚJKA Pavel, PLCHOT Oldřich, GLEMBEK Ondřej, BURGET Lukáš and ČERNOCKÝ Jan. Analysis of Speaker Recognition Systems in Realistic Scenarios of the SITW 2016 Challenge. In: *Proceedings of Interspeech 2016*. San Francisco:

International Speech Communication Association, 2016, pp. 828-832. ISBN 978-1-5108-3313-5.

- PLCHOT Oldřich, BURGET Lukáš, ARONOWITZ Hagai and MATĚJKA Pavel. Audio Enhancing With DNN Autoencoder For Speaker Recognition. In: *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), 2016*. Shanghai: IEEE Signal Processing Society, 2016, pp. 5090-5094. ISBN 978-1-4799-9988-0.

- PLCHOT Oldřich, MATĚJKA Pavel, FÉR Radek, GLEMBEK Ondřej, NOVOTNÝ Ondřej, PEŠÁN Jan, VESELÝ Karel, ONDEL Lucas, KARAFIÁT Martin, GRÉZL František, KESIRAJU Santosh, BURGET Lukáš, BRUMMER Niko, SWART Albert du Preez, CUMANI Sandro, MALLIDI Sri Harish and LI Ruizhi. BAT System Description for NIST LRE 2015. In: *Proceedings of Odyssey 2016, The Speaker and Language Recognition Workshop*. Bilbao: International Speech Communication Association, 2016, pp. 166-173. ISSN 2312-2846.

- CUMANI Sandro, PLCHOT Oldřich and FÉR Radek. Exploiting i-vector posterior covariances for short-duration language recognition. In: *Proceedings of Interspeech 2015*. Dresden: International Speech Communication Association, 2015, pp. 1002-1006. ISBN 978-1-5108-1790-6. ISSN 1990-9772.

- LI Ruizhi, MALLIDI Sri Harish, PLCHOT Oldřich, BURGET Lukáš and DEHAK Najim. Exploiting Hidden-Layer Responses of Deep Neural Networks for Language Recognition. In: *Proceedings of Interspeech 2016*. San Francisco: International Speech Communication Association, 2016, pp. 2365-2369.

Submitted and not yet published papers:

- Ondrej Novotny, Oldrich Plchot, Pavel Matejka, Lukas Burget, “On the use of DNN Autoencoder for Robust Speaker Recognition”, submitted to Interspeech 2017.

- Pavel Matejka, Ondrej Novotny, Oldrich Plchot, Lukas Burget, Mireia Diez Sanchez, Jan Cernocky, “Analysis of Score Normalization in Multilingual Speaker Recognition”, submitted to Interspeech 2017.

Oldrich Plchot et al., “Analysis and Description of ABC Submission to NIST SRE 2016”, submitted to Interspeech 2017.