9 Conclusion

In this thesis, we addressed the problem of text-independent speaker verification from a machine learning point of view. The main purpose was to consider this problem as a two-class classification problem for each speaker. As suggested by the machine learning theory, the model used to solve this kind of problems should be discriminant, while the current state-of-the-art text-independent speaker verification models are based on Gaussian Mixture Models (GMMs) which are not apparently discriminant.

9.1 Contribution of the Thesis

We first described the state-of-the-art models as found in the speaker verification literature. Unfortunately, the performance measures used to compare models are often biased, including Equal Error Rate and Detection Error Trade-off (DET) curves. We have thus proposed new kinds of curves called Expected Performance Curves (EPCs) that allow to compare fairly systems for a range of decision thresholds. This work was published in:

CONTRIB S. Bengio, J. Mariéthoz, and M. Keller. The expected performance curve. In *International Conference on Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning*, 2005

and more specifically for speaker verification in:

CONTRIB S. Bengio and J. Mariéthoz. The expected performance curve: a new assessment measure for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, 2004

Moreover as no statistical test, such as Z-test, was applicable to the speaker verification problem, we adapted the Z-test in order to properly measure whether two systems were statistically significantly different in terms of Half Total Error Rate (HTER) and Detection Cost Function (DCF). This work was published in:

> CONTRIB S. Bengio and J. Mariéthoz. A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, pages 237–240, 2004

We have defined an experimental setup, including a protocol for the use of discriminant models. We performed experiments using three databases: Switchboard coming from the NIST campaign, the Banca database and the Polyvar database. The original benchmark Banca database and its protocol descriptions were published in:

> CONTRIB E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 625–638. Springer-Verlag, 2003

The Polyvar database and its protocol descriptions were published in:

> CONTRIB F. Bimbot, M. Blomberg, L. Boves, G. Chollet, C. Jaboulet, B. Jacob, J. Kharroubi, J. Koolwaaij, J. Lindberg, J. Mariéthoz, C. Mokbel, and H. Mokbel. An overview of the picasso project research activities in speaker verification for telephone applications. In *6th European Conference on Speech Communication and Technology — Eurospeech'99*, volume 5, pages 1963–1966, Budapest, Hungary, september 1999

In order to propose new approaches based on discriminant models, a general framework has been developed for speaker verification that includes several kinds of models: probabilistic models such as GMMs and non-probabilistic models such as Support Vector Machines (SVMs). This framework was originally presented in:

CONTRIB J. Mariéthoz and S. Bengio. A kernel trick for sequences applied to text-independent speaker verification systems. IDIAP-RR 77, IDIAP, 2005

This framework was then extended for the case of score normalization for both probabilistic and non-probabilistic based models. Score normalization is often used to compensate unmatched conditions between data used to train the model and test accesses. A generalized score normalization framework was proposed and enlights the hypothesis implicitly done when T- and Z- normalization are used. This new framework can be used to develop future score normalization procedures and is not limited to a Gaussian score distribution. This work was published in:

CONTRIB J. Mariéthoz and S. Bengio. A unified framework for score normalization techniques applied to text independent speaker verification. *IEEE Signal Processing Letters, Volume 12*, 12, 2005. IDIAP-RR 04-62

In order to better understand the state-of-the-art GMM based system, we analyzed it more deeply and mentioned several modifications suggested by the speaker verification community in order to reach the state-of-the-art performance. We showed that theses modifications make the GMM based model discriminant and is equivalent, using reasonable assumptions, to a mixture of linear classifiers. In order to interpret GMMs as mixtures of experts, we used an algorithm called "synchronous alignment", published in:

CONTRIB J. Mariéthoz, Dominique Genoud, Frédéric Bimbot, and Chafik Mokbel. Client / world model synchronous alignement for speaker verification. In *6th European Conference on Speech Communication and Technology — Eurospeech'99*, Budapest, Hungary, September 1999

The Maximum A Posteriori (MAP) adaptation algorithm is also an important modification in order to obtain good performance. MAP adaptation methods where compared to other standard approaches and this comparison was published in:

CONTRIB J. Mariéthoz and S. Bengio. A comparative study of adaptation methods for speaker verification. In *International Conference on Spoken Language Processing ICSLP*, pages 581–584, Denver, CO, USA, September 2002. IDIAP-RR 01-34

We first tried to develop discriminant models using information coming from the GMM based system by replacing the Bayes decision function of state-of-the-art GMM based systems, which can be seen as a linear function of two log likelihoods with a fixed slope equal to one, by learning a discriminant decision function with an SVM. Learning the decision function suggests that the discriminant models should be client dependent. This work was published in:

CONTRIB S. Bengio and J. Mariéthoz. Learning the decision function for speaker verification. In *IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP*, Salt Lake, City, USA, 2001. IDIAP-RR 00-40

Apart from log likelihoods, several other values could be inputted to an SVM. We can for instance enrich this representation with local log likelihood ratios (LLRs) for each Gaussian in order to increase the size of the input vector. After analyzing the results, we concluded that having only one discriminant model for all clients seems to be a limitation and it could be preferable to have a client dependent discriminant model that could be based on a whole sequence of feature vectors. The use of discriminant models as a decision function and using a large vector of LLRs was proposed in:

CONTRIB J. Mariéthoz and S. Bengio. An alternative to silence removal for text-independent speaker verification. IDIAP-RR 51, IDIAP, Martigny, Switzerland, 2003

These models suggest that the SVM is a good candidate for the speaker verification problem, especially with its ability to maximize the margin. Indeed, to train one model per speaker, we have very few client accesses (often one) and hundreds of impostor accesses. As we observed for SVM based systems, the problem is separable and maximizing the margin guarantees a reasonable solution over all possible solutions that give zero training error. Unfortunately, default SVMs can handle only fixed size vectors and we thus had to propose new kernels that can handle variable length sequences of vectors. We first developed a new Mean operator sequence kernel that computes the average of all sub-kernels over all pairs of frames. We showed that it generalizes the GLDS kernel proposed by Campbell (2002) with the advantage to better control the capacity of the SVM model, while making possible the use of infinite space kernels, such as Radial Basis Functions (RBFs).

We also proposed a new Max operator sequence kernel that searches for each frame of one sequence, the frame of the other sequence that best matches. It makes more sense and outperforms the standard approach. Unfortunately it does not satisfy the Mercer conditions but still converges very well for the studied databases. This work was published in:

> CONTRIB J. Mariéthoz and S. Bengio. A kernel trick for sequences applied to text-independent speaker verification systems. In *Second Workshop on Multimodal User Authentication, MMUA*, 2006. IDIAP-RR 05-77

A longer version of this paper has been submitted to the Patter Recognition journal.

We also proposed a method to smooth the Max operator based kernel. The good empirical results suggest that a more sophisticated method to enforce some temporal constraints can be a topic of future research.

Unfortunately, the Max operator method is computationally costly for long sequences. We thus proposed clustering techniques to make the algorithm tractable for long sequence based databases, such as Switchboard (NIST).

Finally, as speaker verification is a highly unbalanced two-class classification problem, it might be important to consider specific training criteria for such cases. As for most tested SVM kernels the problem is separable, the classical approach to compensate the unbalanced dataset are useless. We concluded that the solution found by the SVM is good even for highly unbalanced class examples.

A new SVM criterion that allows to deal with unbalanced class problems and interprets the output of an SVM as a probability has been published in:

> CONTRIB Y. Grandvalet, J. Mariéthoz, and S. Bengio. A probabilistic interpretation of svms with an application to unbalanced classification. In *Advances in Neural Information Processing Systems, NIPS 15*, 2005. IDIAP-RR 05-26

REF W.M. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. In *Proc IEEE International Conference on Audio Speech and Signal Processing*, pages 161–164, 2002.

We finally proposed a new research direction based on new distance measures. Such a measure should allow a training negative example to cover other unseen impostors. Our new approach is based on the vicinity function proposed by Vapnik (2000). The main idea is to assume a Gaussian noise on the negative examples. Even if this method is not principled, it gives good empirical results and suggests several extensions of our research for this problem. A Gaussian noise can also be added in order to capture the acquisition channel variability.

Overall, in this thesis, we used the machine learning theory to develop a good methodology and a good framework for the speaker verification problem. We proposed several new discriminant models that improve the HTER performance of the state-of-the-art systems, but more importantly that increase the understanding of these models.

This opens several new research perspectives. For example, the score normalization framework allows the use of new score normalization procedures based on non-Gaussian score distribution estimation. The smoothing Max operator kernel suggests to consider some temporal constraints. It seems also very promising to develop a new similarity measure that includes some noise on the data distribution, either to allow a negative example to cover more unseen impostors but also to account for acquisition channel variation. An other general problem is that in "real" life we have no idea of what is a true impostor, which kind of strategy he/she can develop to break the system. Moreover, we cannot base our intuition on a human criterion: we showed in Mariéthoz and Bengio (2005) that current verification systems are robust to professional imitators while humans are not, while at the opposite automatic systems are less robust to noise than humans. In terms of applications, there is an evident need for mobile phone applications using some form of person identification. Thus speaker verification systems should be more and more robust to various recording conditions. Even if already existing solutions are robust for reasonable levels of noise, better robustness is still needed for high levels of noise. A potential solution could be the use of pre-processing methods such as selecting an audio source using a microphone array. An other interesting approach could be to use different biometric modalities such as speech, face, lips, etc. Existing approaches often simply fuse the scores obtained by each modality, but more principled approach to jointly consider all modalities during training are still needed.

REF V. N. Vapnik. *The nature of statistical learning theory.* Springer, second edition, 2000.

REF J. Mariéthoz and S. Bengio. Can a professional imitator fool a GMM-based speaker verification system? IDIAP-RR 61, IDIAP, 2005.

9.2 Other contributions

All the algorithms developed in this thesis are based on a machine learning library called “Torch”. This library is widely used by the machine learning community and is available at http://www.torch.ch. The author is one of the main contributor of this software.

During the course of this thesis, several other, yet related, scientific contributions were accepted for publications but not described here. They are:

CONTRIB S. Marcel, J. Mariéthoz, Y. Rodriguez, and F. Cardinaux. Bi-modal face and speech authentication: a biologin demonstration system. In *Workshop on Multimodal User Authentication (MMUA)*, 2006. IDIAP-RR 06-18

CONTRIB Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24(8):882–893, 2006

CONTRIB M. Liwicki, A. Schlapbach, H. Bunke, S. Bengio, J. Mariéthoz, and J. Richiardi. Writer identification for smart meeting room systems. In *Seventh IAPR Workshop on Document Analysis Systems, DAS*, 2006

CONTRIB J. Mariéthoz and S. Bengio. Can a professional imitator fool a GMM-based speaker verification system? IDIAP-RR 61, IDIAP, 2005

CONTRIB J. Mariéthoz and S. Bengio. A new speech recognition baseline system for numbers 95 version 1.3 based on torch. IDIAP-RR 16, IDIAP, 2004

CONTRIB Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Estimating the quality of face localization for face verification. In *IEEE International Conference on Image Processing, ICIP*, 2004

CONTRIB C. Sanderson, S. Bengio, H. Bourlard, J. Mariéthoz, R. Collobert, M.F. BenZeghiba, F. Cardinaux, and S. Marcel. Speech & face based biometric authentication at idiap. In *International Conference on Multimedia and Expo, ICME*, 2003

CONTRIB S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz. Confidence measures for multimodal identity verification. *Information Fusion*, 3(4):267–276, 2002