COMPENSATION OF EXTRINSIC VARIABILITY IN SPEAKER VERIFICATION SYSTEMS ON SIMULATED SKYPE AND HF CHANNEL DATA

Korbinian Riedhammer *Tobias Bocklet* *Elmar Nöth*

Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstraße 3, 91058 Erlangen, GERMANY
korbinian.riedhammer@informatik.uni-erlangen.de

ABSTRACT

In this work we focus on speaker verification on channels of varying quality, namely Skype and high frequency (HF) radio. In our setup, we assume to have telephone recordings of speakers for training, but recordings of different channels for testing with varying (lower) signal quality. Starting from a Gaussian mixture / support vector machine (GMM/SVM) baseline, we evaluate multi-condition training (MCT), an ideal channel classification approach (ICC), and nuisance attribute projection (NAP) to compensate for the loss of information due to the transmission. In an evaluation on Switchboard-2 data using Skype and HF channel simulators, we show that, for good signal quality, NAP improves the baseline system performance from 5% EER to 3.33% EER (for both Skype and HF). For strongly distorted data, MCT or, if adequate, ICC turn out to be the method of choice.

Index Terms— speaker verification, channel compensation

1. INTRODUCTION

The task of speaker verification describes the two-class problem of detecting speakers who pretend to be someone else, so-called impostors. In addition to the traditional scenario where speaker verification is applied to recordings from the telephone system or room microphones, other channels of communication draw more attention, e.g., Skype (`http://www.skype.com`) as a very popular (free) voice-over-IP service or HF radio for long range communication as for military, nautical or aviation purposes.

The state-of-the-art is to model a speaker by Gaussian mixture models (GMM) [1] that are estimated by features extracted from a spoken utterance, typically Mel frequency cepstrum coefficients (MFCCs). Our framework uses a universal background model (UBM) representing a set of background speakers. This UBM is then adapted to speaker specific models using maximum a posteriori (MAP) adaptation [2]. The mean values of these models represent each target speaker in a high-dimensional space. In a next step, for each training speaker, a support vector machine (SVM) is trained where the UBM is employed as imposter model. The classification task is to determine whether a test speaker is closer to the background speakers or to the target speaker [3].

One major problem in this scenario is session variability which contains both, extrinsic and intrinsic speaker variations [4]. This has been addressed by different techniques at different system levels. On feature level, feature mapping (FM) [5] can be used to reduce the effect of different channels. On model level, transformations like *nuisance attribute projection* (NAP) [6] or *joint factor analysis* (JFA) [7] can be applied. While FM and NAP do not handle the two kinds of variability differently, JFA tries to model extrinsic and intrinsic variations jointly.[1]

In this work we keep the intrinsic variations constant and focus on a "controlled" variation of the extrinsic factors, i.e., recording channel or codec differences. This is achieved by applying various channel simulators to a set of given telephone data. from the Switchboard-2 [8] corpus. We exemplarily simulate high-frequency (HF) recordings in various quality levels as defined by the CCIR [9] and Skype codec compression in various quality settings using the Skype API. For the latter, the variations are in packet loss and in bit rate. Note that, for Skype, we focus on the actual simulated audio data and not the encrypted stream as for example in [10].

In this work, we evaluate four types of systems, all based on the previously described GMM/SVM architecture.

1. As a baseline, we train a GMM/SVM system using the original telephone data, and test it on both, the original and simulated data. This system is confronted with a strong acosutic mismatch between training and test conditions.

2. An ideal channel classification system, i.e., we train an individual system for each channel setting using the simulated training data and test on the respective simulated test data. This results in one system trained specifically on each channel configuration. This system is designed to have the least mismatch in training and test.

3. A general GMM/SVM system trained in a multi-condition manner, all simulated data for each recording of the training set are employed to train multi-condition speaker models. The system is then tested on all test data.

4. A state-of-the-art intersession variability (ISV) compensation GMM/SVM system where a NAP transformation is estimated on various simulated quality settings of the training recordings. This results in a system with speaker models trained solely on the original (telephone) data but transformed into a "channel-free" space. The system is applied to all simulation conditions of the test data.

This article is structured as follows. After a brief introduction of the data and the channel simulation in Section 2, we describe the different speaker verification systems in Section 3. The results of the different systems are analyzed in Sec. 4. We conclude with a summary and an outlook in Sec. 5.

[1] As we use channel simulators to obtain several versions of the same recordings thus eliminating the speaker variability, we chose the computationally easier NAP for this work.

2. DATA

The original data was taken from the Switchboard-2 [8] corpus. The UBM was estimated on 1894 speakers. For the evaluation, 60 training and test speakers (disjoint from the background speakers) with two conversations each were selected to match a commercial evaluation setup of our industrial partner MEDAV GmbH (`http://www.medav.de`). One conversation was used for training, the other conversation was used for testing. For the actual evaluation, each of the 60 test utterances is paired with each of the training speakers resulting in a total of 3600 (60 target, 3540 non-target) trials.

2.1. HF Channel Simulation

The HF channel simulator used in this work was designed by the MEDAV GmbH and follows the respective CCIR recommendation [9]. It defines three quality levels "good", "moderate" and "poor". In addition, the resulting signal-to-noise ration can be varied between 0, 5, 10, 15, 20, 25 or 30 dB. Considering all different settings, the original 120 recordings result in 2520 simulated HF transmissions.

2.2. Skype Channel Simulation

The MEDAV GmbH provided us with Skype channel simulations using the Skype API. Using the quality settings 6, 8 and 12 kbps (kilo bit per second), a packet loss of 0%, 20%, 40%, 60% and 80% was simulated. These settings result in additional 1800 simulated Skype transmissions.

3. SPEAKER IDENTIFICATION SYSTEMS

From the 8 kHz, 16 bit telephone data, the MFCCs are extracted using a 25 ms window with a time-shift of 10 ms. The FFT coefficients are compressed using 25 equidistant Mel filters covering 300-3400 Hz. After a discrete cosine transform (DCT) the first 12 coefficients are selected. The very first coefficient is replaced by the short time energy. Furthermore, we extract deltas and delta-deltas, and perform an utterance-based mean and variance normalization. We used an energy-based voice activity detector to remove possible silence frames.

3.1. Baseline

The systems in this work use 1024 Gaussian mixtures with diagonal covariances to model the background and target speakers. The UBM is estimated from 1894 speakers using the well-known EM algorithm. From this UBM, the target speaker models are derived by applying MAP adaptation to the mean vectors with a relevance factor $r = 16$ [2]. For the background and each speaker, the mean vectors of the respective model form the supervector (SV). S linear kernel SVM using the `libSVM` toolkit [11] is trained for each speaker of the training set, labeling the background as -1 and target as $+1$. After scoring the impostor trials, we apply the T-norm. This general system design also applies to the following systems.

For the baseline system, only the original 60 training utterances are used to compute the speaker models. For the original telephone test set, it yields an EER of 5%.

3.2. Ideal Channel Classification

The motivation behind the ICC system(s) is that it is best to have matching training and testing conditions. Therefore, for each of the different channel configurations, an individual system is trained analog to the baseline system, but using simulated channel training data that matches the test conditions, instead of the original telephone data. Of course, the use of this system is somewhat theoretical as it would require a perfect channel condition classification prior to the actual system application.

3.3. Multi-Condition Training

The motivation for multi-condition training is to allow the statistical models to learn the actual speaker characteristics aside from varying channel side effects. The original 1894-speaker UBM is adapted to the training data (both original telephone and simulated) of the training data of the 60-speaker dataset in order to improve the modeling capabilities for the new channels. In a second step, the speaker models are adapted using all available training data. The resulting models are supposed to be less specific but more robust against channel variation.

3.4. Intersession Variability Compensation

Though the recent JFA [7] allows to jointly estimate session and channel variability, we chose to apply the computationally easier NAP [6] as we kept the intrinsic variabilty of the training and test speakers constant because of the simulated channel data.

In short, NAP tries to eliminate "unnecessary" subspaces in the training data $\{\boldsymbol{x}\}$ by finding a projection $P = (\mathbf{1} - vv^t)$ constrained by

$$v^* = \underset{v,||v||^2=1}{\operatorname{argmax}} \sum_{i,j} W_{ij} ||P\boldsymbol{x}_i - P\boldsymbol{x}_j||^2 \quad . \tag{1}$$

With a proper selection of the weights W_{ij}, P can be used to minimize the distance between two points that should be close to each other without reducing the dimensionality as for example with (kernel) PCA. Here, we set $W_{ij} = 0$, if the channels of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ match and $W_{ij} = 1$ otherwise.

The solution to this optimization problem can be found via solving an Eigenvalue problem. Furthermore, P can be constructed to have a "co-rank" corresponding to the number of Eigenvectors V used instead of the single Eigenvector v corresponding to the largest Eigenvalue. For this work, we use a co-rank of 4 Eigenvectors, as the evaluation setup covers only 60 speakers.

The NAP transformation is estimated for Skype and HF separately using both original and respective simulated data. This is motivated by the application scenario, where the type of channel is known, but the quality remains unknown or variable, and results in an individual system for the two channel groups.

4. RESULTS

4.1. Skype

Fig. 1 shows the baseline system performance for the simulated Skype channel data. For the best quality (12 kbps, 0% loss), it yields an EER of 5% as for the original test data. It however degrades to 16.67% for the worst setting (6 kbps, 80% loss). Both kbps and loss seem to have a strong impact on the performance.

Figs. 2-4 show the four systems in comparison. Each figure focuses on a different skype quality (6,8, and 12 kbps) with varying package loss (0-80%) in each figure. Note that the NAP system outperforms all other systems as long as the loss remains below 40% (6 kbps) and 60% (other) yielding an EER up to 3.33%. For the strong lossy channels, MCT and ICC are the methods of choice. The

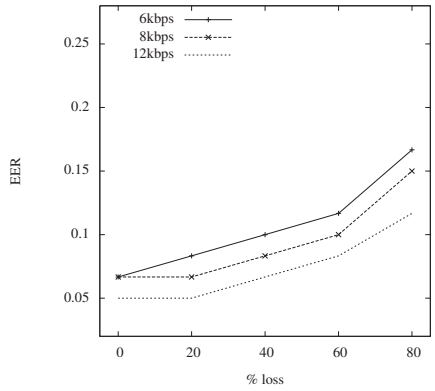

Fig. 1. Overview of the baseline system performance for the Skype channel data w.r.t to varying package loss.

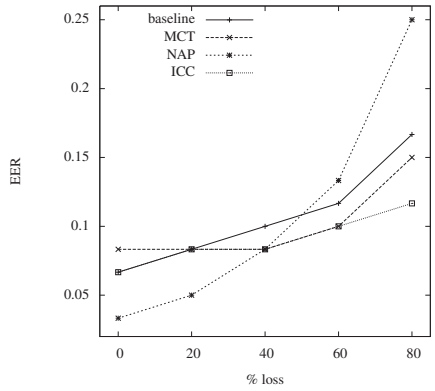

Fig. 2. System performance for Skype codec with various loss settings at 6 kbps.

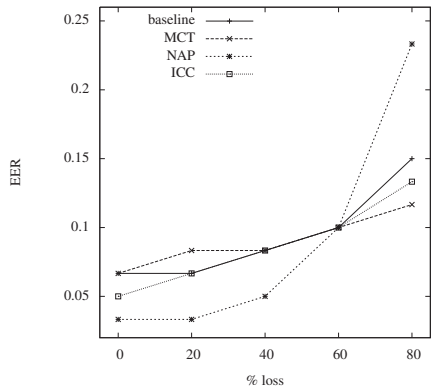

Fig. 3. System performance for Skype codec with various loss settings at 8 kbps.

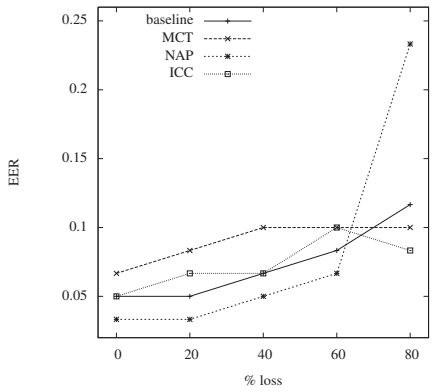

Fig. 4. System performance for Skype codec with various loss settings at 12 kbps.

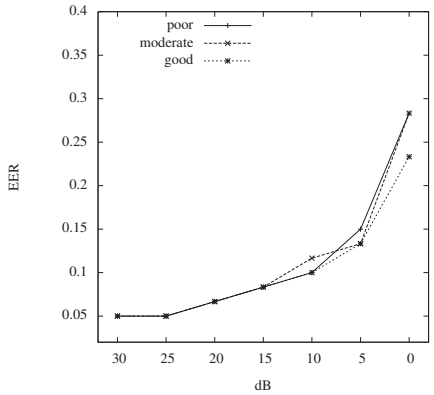

Fig. 5. Overview of the baseline system performance for the HF channel data.

rather weak performance of the MCT system for "good" channels can be explained by an overall robust performance as it is a compromise for all channel settings. Though the performance might look disappointing, it is still remarkable that the systems still work surprisingly good given the tremendous losses up to 80%. Interestingly, the baseline system sometimes outperforms the ICC which can be explained by overtraining to the channel effects.

4.2. HF

Fig. 5 shows the baseline system performance for the simulated HF channels, also starting at an EER of 5%. In contrast to the results on Skype data, the performance seems to be mainly dependent on the SNR rather than the actual CCIR quality level. It is remarkable that already the baseline system is somewhat able to handle 5 dB SNR data which is of course heavily distorted.

Figs. 6-8 show the detailed system comparison for the three CCIR quality levels w.r.t. SNR. Similar to the observation for the Skype data, the NAP system outperforms the other systems between an SNR range of 30 to 15 db in all quality levels. For the range between 30 and 20 db the system achieves a constant EER of 3.33%. Also, the results suggest to use MCT if low SNR data is expected, suggesting that the MCT is indeed more robust than the baseline or

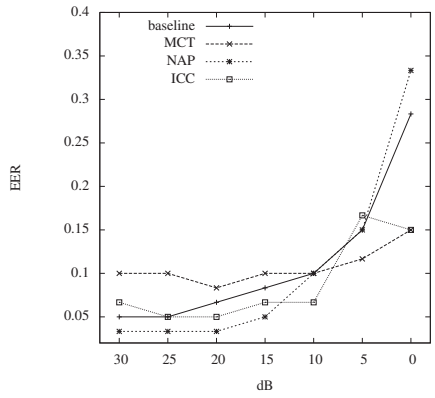

Fig. 6. System performance for HF simulator setting "poor" at various SNR.

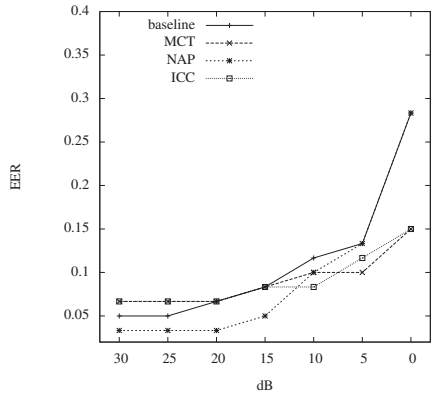

Fig. 7. System performance for HF simulator setting "moderate" at various SNR.

NAP system at the cost of performance on data with a high SNR.

5. SUMMARY

In this work we presented a detailed evaluation of channel compensation for simulated HF and Skype channel data, two increasingly important channels for speaker verification. The baseline performance could be greatly improved by applying NAP trained on various channel quality settings. The improvement did however not hold for low quality settings (high loss or low SNR). MCT or, if adequate, ICC systems show a rather average performance for high quality data but remain robust to strong distortions. As for future work, the presented experiments should be extended to a larger evaluation setup, to both confirm the results and to allow more complex intersession variability compensation techniques.

6. ACKNOWLEDGMENTS

The authors would like to thank the MEDAV GmbH for funding and supporting this work under the joint KAIMAN project (STMWVT grant IUK-0906-0002).

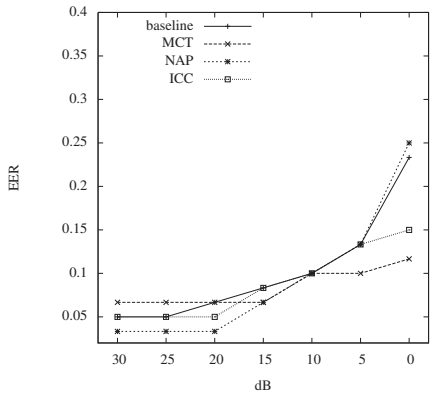

Fig. 8. System performance for HF simulator setting "good" at various SNR.

7. REFERENCES

[1] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transaction on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[3] W. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *Signal Processing Letters, IEEE*, vol. 13, pp. 308–311, 2006.

[4] M. Graciarena, T. Bocklet, E. Shriberg, A. Stolcke, and S. Kajarekar, "Feature-Based and Channel-Based Analyses of Intrinsic Variability in Speaker Verification," 2009, pp. 2015–2018.

[5] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, vol. II, pp. 53–56.

[6] A. Solomonoff, W. Campbell, and C. Quillen, "Nuisance attribute projection," *Speech Communication*, 2007.

[7] N. Dehak, P. Kenny, R. Dehak, O. Glembek, and P. Dumouchel L. Burget V. Hubeika F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proc. IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 4237–4240.

[8] D. Graff, D. Miller, and K .Walker, *Switchboard-2 Phase III Audio*, Linguistic Data Consortium, Philadelphia, 2002.

[9] ITU, "Recommendation 520-1. Use of High Frequency Ionospheric Channel Simulators," *Recommendations and Reports of the CCIR*, vol. III, pp. 57–58.

[10] M. Backes, G. Doychev, M. Dürmuth, and B. Köpf, "Speaker recognition in encrypted voice streams," in *Computer Security - ESORICS 2010*, vol. 6345, pp. 508–523. 2010.

[11] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines,* `http://www.csie.ntu.edu.tw/ cjlin/libsvm`, 2001.